Martin Aigner - Günter M. Ziegler
Traduction : Nicolas Puech

Raisonnements divins

Quelques démonstrations mathématiques particulièrement élégantes

Troisième édition

Springer

Raisonnements divins

Springer

Paris
Berlin
Heidelberg
New York
Hong Kong
Londres
Milan
Tokyo

Martin Aigner
Günter M. Ziegler

Raisonnements divins

Quelques démonstrations mathématiques particulièrement élégantes

Traduit de l'anglais par Nicolas Puech

Illustrations de Karl H. Hofmann

Troisième édition

Springer

Martin Aigner

Freie Universität Berlin
Institut für Mathematik II
Arnimallee 3
14195 Berlin, Germany
aigner@math.fu-berlin.de

Günter M. Ziegler

Freie Universität Berlin
Institut für Mathematik
Arnimallee 2
14195 Berlin, Germany
ziegler@math.fu-berlin.de

Traduction de la quatrième édition anglaise de :
Proofs from THE BOOK by Martin Aigner, Günter M. Ziegler
ISBN : 978-3-642-00855-9

ISBN 978-2-8178-0399-9 Springer Paris Berlin Heidelberg New York

Springer-Verlag France est membre du groupe Springer Science + Business Media
Traduction de la première édition par Nicolas Puech et Jean-Marie Morvan
Traduction des deuxième, troisième et quatrième éditions par Nicolas Puech
Mise en page des première, deuxième et troisième éditions françaises par Nicolas Puech

Couverture : Jean-François Montmarché

Préface

Paul Erdős aimait parler du Grand Livre, dans lequel Dieu inscrit les preuves parfaites des théorèmes mathématiques, suivant ainsi la remarque de Hardy : il n'y a pas de lieu permanent pour les mathématiques laides. Erdős disait aussi que l'on n'a pas besoin de croire en Dieu, mais qu'en tant que mathématicien, on doit croire au Grand Livre. Il y a quelques années, nous lui avions suggéré d'écrire une première (et très modeste) ébauche du Grand Livre. L'idée l'enthousiasma et, comme à son habitude, il se mit immédiatement au travail, remplissant des pages et des pages de suggestions. Notre livre devait paraître en mars 1998 comme cadeau pour le 85^{e} anniversaire d'Erdős. Malheureusement, la mort de Paul durant l'été 1996 n'a pas permis qu'il en soit l'un des co-auteurs. Ce livre est dédié à sa mémoire.

Paul Erdős

Nous n'avons pas de définition ni de caractérisation de ce qui constitue une preuve du Grand Livre : nous ne faisons qu'offrir les exemples que nous avons sélectionnés, en espérant que nos lecteurs partageront notre enthousiasme pour des idées brillantes, des points de vue ingénieux et de magnifiques observations. Nous espérons aussi que nos lecteurs ne nous tiendront pas rigueur des imperfections de notre exposé. La sélection est essentiellement due à Paul Erdős lui-même. Il a suggéré un grand nombre de sujets, plusieurs preuves lui reviennent directement, d'autres ont été amorcées par la bonne question ou la bonne conjecture qu'il a su proposer. On peut donc considérer que, dans une grande mesure, cet ouvrage reflète le point de vue de Paul Erdős à propos de ce qui devrait être considéré comme une preuve du Grand Livre.

« Le Grand Livre ».

Nous avons limité notre sélection de sujets de sorte que le livre tout entier soit accessible aux lecteurs dont les connaissances mathématiques se restreignent à un premier cycle universitaire. Un peu d'algèbre linéaire, quelques notions fondamentales d'analyse et de théorie des nombres, et une bonne maîtrise des concepts élémentaire de mathématiques discrètes devraient suffire pour comprendre ce livre dans son intégralité et en tirer profit.

Nous sommes extrêmement reconnaissants envers certaines personnes qui nous ont aidés et soutenus dans ce projet ; on compte notamment parmi elles les étudiants d'un séminaire pendant lequel nous avons discuté d'une version préliminaire de l'ouvrage. Merci à Benno Artmann, Stephan Brandt, Stefan Felsner, Eli Goodman, Torsten Heldmann et Hans Mielke. Nous remercions également Margrit Barrett, Christian Bressler, Ewgenij Gawrilow, Michael Joswig, Elke Pose et Jörg Rambau pour leur aide technique dans la composition de ce livre. Nous devons beaucoup à Tom Trotter qui a lu le manuscrit de la première à la dernière page, à Karl H. Hofmann pour ses magnifiques illustrations, et surtout au grand Paul Erdős.

Berlin, Mars 1998 *Martin Aigner · Günter M. Ziegler*

Préface de la deuxième édition

La première édition de ce livre a été très bien accueillie. Nous avons reçu un nombre inhabituel de lettres contenant des commentaires et des corrections, quelques raccourcis, d'intéressantes suggestions de preuves différentes et de nouveaux sujets à traiter. Bien que nous essayions de répertorier des preuves *parfaites*, notre exposé ne l'est pas.

La seconde édition nous donne l'occasion de présenter une nouvelle version de notre livre : il contient trois chapitres supplémentaires, des modifications substantielles, de nouvelles preuves, ainsi que des améliorations mineures suggérées par nos lecteurs. Le chapitre consacré au « problème des treize sphères » a été supprimé. La démonstration reposait sur des détails que nous ne pouvions pas rédiger de manière succincte et élégante.

Merci à tous les lecteurs qui nous ont écrit et aidés ; parmi eux, Stephan Brandt, Christian Elsholtz, Jürgen Elstroth, Daniel Grieser, Roger Heath-Brown, Lee L. Keener, Christian Lebœuf, Hanfried Lenz, Nicolas Puech, John Scholes, Bernulf Weißbach et *beaucoup* d'autres. Merci encore pour leur aide et leur soutien à Ruth Allewelt et Karl-Friedrich Koch de Springer Heidelberg, à Christoph Eyrich et Torsten Heldmann de Berlin, et à Karl H. Hofmann pour les nouvelles superbes illustrations.

Berlin, Septembre 2000 *Martin Aigner · Günter M. Ziegler*

Préface de la troisième édition

Lors de la préparation de la première édition, nous n'aurions jamais osé rêver d'un tel succès : le livre a été traduit dans plusieurs langues et nous avons reçu de nombreux courriers de lecteurs enthousiastes avec des suggestions d'améliorations et des propositions pour développer de nouveaux thèmes si nombreuses qu'elles pourraient nous occuper pendant plusieurs années.

Ainsi, cette troisième édition comporte deux nouveaux chapitres (sur les partitions d'entiers d'Euler et sur les manières de mélanger un jeu de cartes), le regroupement de trois démonstrations pour le calcul de la série d'Euler au sein d'un chapitre à part entière. Il présente aussi de nombreuses améliorations comme le procédé de Calkin-Wilf-Newman pour « énumérer les rationnels ».

Nous remercions tous ceux qui ont soutenu ce projet pendant les cinq dernières années et dont les indications ont permis les améliorations que l'on trouve dans cette nouvelle édition. Plus particulièrement, merci à David Bevan, Anders Björner, Dietrich Braess, John Cosgrave, Hubert Kalf, Günter Pickert, Alistair Sinclair et Herb Wilf.

Berlin, Juin 2003 *Martin Aigner · Günter M. Ziegler*

Préface de la quatrième édition

Lorsque nous avons entrepris ce projet il y a près de quinze ans, nous n'avions pas imaginé l'accueil enthousiaste qu'il a reçu et qui s'est traduit par de nombreux courriers chaleureux, des commentaires intéressants, de nouvelles éditions et des traductions en treize langues différentes. Il n'est pas exagéré de dire que ce livre est devenu une partie de nos vies.

Outre de nombreuses améliorations, en partie suggérées par nos lecteurs, cette quatrième édition comporte cinq nouveaux chapitres. Deux thèmes classiques sont abordés : la loi de réciprocité quadratique et le théorème fondamental de l'algèbre ; deux chapitres traitent des pavages ; enfin un chapitre s'intéresse au nombre chromatique des graphes de Kneser.

Nous remercions tous ceux qui nous ont aidés et encouragés tout au long de ces années. Pour la deuxième édition, nous sommes reconnaissants à Stephan Brandt, Christian Elsholtz, Jürgen Elstrodt, Daniel Grieser, Roger Heath-Brown, Lee L. Keener, Christian Lebœuf, Hanfried Lenz, Nicolas Puech, John Scholes, Bernulf Weißbach, et de nombreux autres encore. La troisième édition bénéficia des suggestions de David Bevan, Anders Björner, Dietrich Braess, John Cosgrave, Hubert Kalf, Günter Pickert, Alistair Sinclair, and Herb Wilf. Pour la présente édition, nous sommes particulièrement reconnaissants pour les contributions de France Dacar, Oliver Deiser, Anton Dochtermann, Michael Harbeck, Stefan Hougardy, Hendrik W. Lenstra, Günter Rote, Moritz Schmitt, and Carsten Schultz. Nous remercions aussi Ruth Allewelt de Springer à Heidelberg ainsi que Christoph Eyrich, Torsten Heldmann et Elke Pose de Berlin pour leur aide et leur soutien tout au long de ce projet. Enfin, cet ouvrage n'aurait certainement pas la même allure sans les conseils prodigués par Karl-Friedrich Koch pour la mise en page et sans les superbes illustrations réalisées pour chaque édition par Karl H. Hofmann.

Berlin, Juillet 2009 *Martin Aigner · Günter M. Ziegler*

Préface de la troisième édition française

Compte tenu du succès rencontré par l'ouvrage anglais, et afin de le mettre à la portée de tous, nous avons souhaité qu'il soit traduit en français. Nous sommes heureux qu'une nouvelle édition rende accessible au lectorat français les améliorations de la plus récente édition anglaise.

Cette édition française correspond à la quatrième édition anglaise (et corrige même certaines coquilles ou erreurs subsistant dans celle-ci). Ces modifications font de cet ouvrage la version la plus à jour de ce texte à la date de publication.

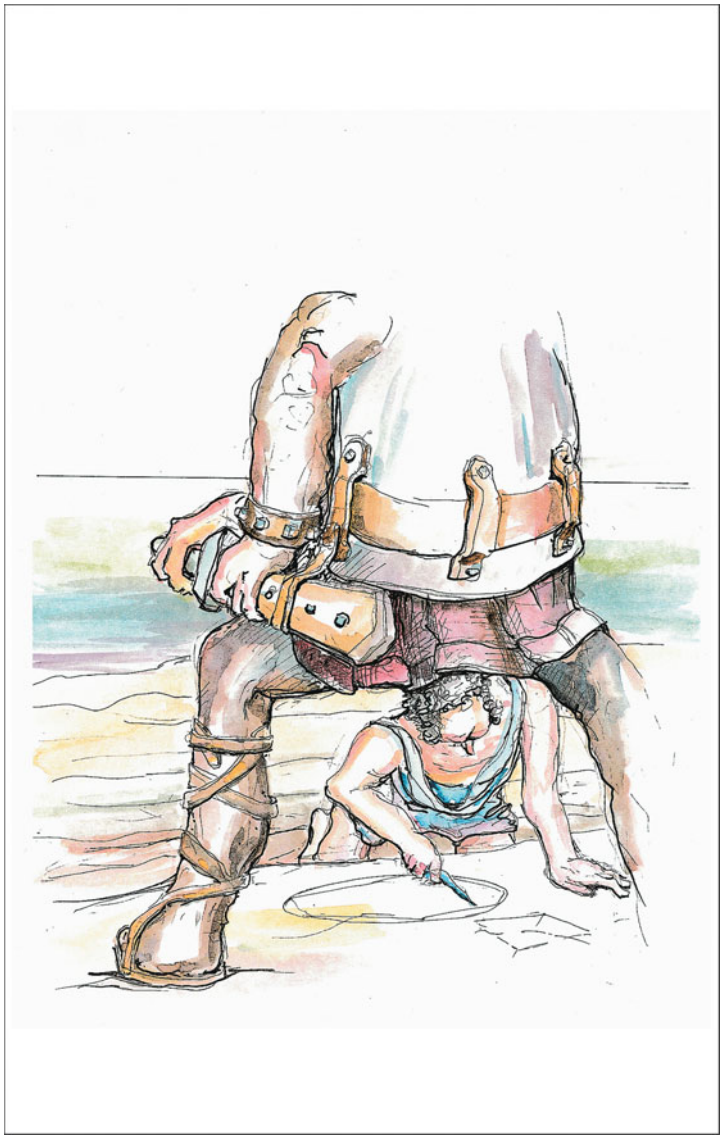
Nous sommes très reconnaissants aux traducteurs [1] pour le travail qu'ils ont effectué. Nous souhaitons remercier particulièrement Nicolas Puech pour ses suggestions, pour la rigueur qu'il a apportée à la traduction de certains passages et pour le considérable travail de mise en page accompli afin que puisse être réalisé l'ouvrage français dans sa forme actuelle.

Berlin, Juin 2012 *Martin Aigner · Günter M. Ziegler*

1. Nicolas Puech remercie Laurent Decreusefond, Oivier Hudry, Mohamed Koubàa, Christian Lebœuf, Antoine Lobstein, Philippe Martins, Alain Maruani, Bruno Petazzoni, Hugues Randriambololona et Günter Ziegler pour leur aide dans la mise au point des éditions successives de cet ouvrage.

Sommaire

Théorie des nombres



« Irrationalité et π ».

Six preuves de l'infinité de l'ensemble des nombres premiers

Chapitre 1

Il est bien naturel de commencer ces notes avec la preuve probablement la plus ancienne du Grand Livre, habituellement attribuée à Euclide (*Éléments* IX, 20). Elle montre que la suite des nombres premiers est infinie.

■ **La preuve d'Euclide.** Étant donné un ensemble fini $\{p_1, \ldots, p_r\}$ de nombres premiers, considérons le nombre $n = p_1 p_2 \cdots p_r + 1$. Ce nombre n a un diviseur premier p. Cependant p n'est pas l'un des p_i sinon p serait un diviseur de n, du produit $p_1 p_2 \cdots p_r$, et donc aussi de la différence $n - p_1 p_2 \ldots p_r = 1$, ce qui est impossible. Ainsi, un ensemble fini $\{p_1, \ldots, p_r\}$ ne peut constituer la collection de *tous* les nombres premiers. □

Avant de poursuivre, fixons quelques notations. $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ est l'ensemble des entiers naturels, $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$ l'ensemble des entiers naturels non nuls, $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ l'ensemble des entiers relatifs et $\mathbb{P} = \{2, 3, 5, 7, \ldots\}$ l'ensemble des nombres premiers.

Dans ce qui suit, nous allons exposer quelques autres démonstrations (choisies parmi bien d'autres) de ce résultat ; nous espérons que le lecteur les appréciera autant que nous. Bien qu'elles utilisent des points de vue différents, elles utilisent toutes les résultats suivants : les entiers naturels croissent au-delà de toute borne et tout entier naturel $n \geqslant 2$ admet un diviseur premier. Ensemble, ces deux faits contraignent $\mathbb{P}$ à être infini. Les trois preuves suivantes viennent du folklore, la cinquième a été proposée par Harry Fürstenberg et la dernière est attribuée à Paul Erdős.

Les deuxième et troisième preuves utilisent des suites particulières d'entiers bien connues.

■ **Deuxième preuve.** Examinons tout d'abord les *nombres de Fermat* $F_n = 2^{2^n} + 1$ où $n = 0, 1, 2, \ldots$ Nous allons montrer que deux nombres de Fermat (distincts) sont premiers entre eux ; en conséquence, il doit y avoir un nombre infini de nombres premiers [1]. À cet effet, vérifions la formule de récurrence :

$$\prod_{k=0}^{n-1} F_k \;=\; F_n - 2 \qquad (n \geqslant 1)$$

à partir de laquelle on déduit immédiatement notre assertion. En effet, si m est, par exemple, un diviseur de F_k et F_n $(k < n)$ alors m divise 2 et, par conséquent, $m = 1$ ou $m = 2$. Cependant, $m = 2$ est impossible puisque tous les nombres de Fermat sont impairs.

$$\begin{aligned} F_0 &= 3 \\ F_1 &= 5 \\ F_2 &= 17 \\ F_3 &= 257 \\ F_4 &= 65537 \\ F_5 &= 641 \times 6700417 \end{aligned}$$

Les premiers nombres de Fermat.

1. N.d.T. : puisque chaque nombre de Fermat comporte dans sa décomposition en facteurs premiers au moins un facteur qu'on ne retrouve pas dans la décomposition des autres.

Le théorème de Lagrange

Si G est un groupe (multiplicatif) fini et U un sous-groupe de G, alors $|U|$ divise $|G|$.

■ **Preuve.** Considérons la relation binaire :

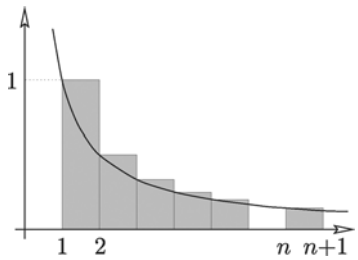
$$a \sim b : \Longleftrightarrow\ ba^{-1} \in U$$

On déduit des axiomes de groupe que $\sim$ est une relation d'équivalence. La classe d'équivalence qui contient l'élément a est exactement la classe :

$$Ua = \{xa : x \in U\}$$

Puisque l'on a clairement $|Ua| = |U|$, on en déduit que G se décompose en classes d'équivalence ayant toutes le même cardinal $|U|$ et que, par conséquent, $|U|$ divise $|G|$. □

Dans le cas particulier où U est un sous-groupe cyclique $\{a, a^2, \ldots, a^m\}$, on trouve que m (le plus petit entier positif tel que $a^m = 1$, appelé *l'ordre* de a) divise le cardinal $|G|$ du groupe.



Escalier au dessus de la fonction $t \mapsto f(t) = \frac{1}{t}$.

Pour montrer la formule, nous faisons un raisonnement par récurrence sur n. Pour $n = 1$, nous avons $F_0 = 3$ et $F_1 - 2 = 3$. Nous constatons ensuite que :

$$\prod_{k=0}^{n} F_k = \Big(\prod_{k=0}^{n-1} F_k\Big) F_n = (F_n - 2)F_n =$$
$$= (2^{2^n} - 1)(2^{2^n} + 1) = 2^{2^{n+1}} - 1 = F_{n+1} - 2 \qquad \square$$

■ **Troisième preuve.** Supposons que $\mathbb{P}$ soit fini et que p soit le plus grand nombre premier. Considérons le *nombre de Mersenne* $2^p - 1$ et montrons que tout facteur premier q de $2^p - 1$ est plus grand que p, ce qui implique la conclusion désirée. Soit q un nombre premier qui divise $2^p - 1$. Nous avons donc $2^p \equiv 1 \pmod q$. Puisque p est un nombre premier, cela signifie que l'élément 2 est d'ordre p dans le groupe multiplicatif $\mathbb{Z}_q \backslash \{0\}$ du corps $\mathbb{Z}_q$. Ce groupe a $q - 1$ éléments. Grâce au théorème de Lagrange (voir encadré), nous savons que l'ordre de chaque élément divise le cardinal du groupe, c'est-à-dire que $p \mid q - 1$, et par conséquent $p < q$. □

Penchons-nous maintenant sur une preuve qui utilise l'analyse.

■ **Quatrième preuve.** Soit $\pi(x) := \text{Card}\{p \leqslant x : p \in \mathbb{P}\}$ le cardinal de l'ensemble des nombres premiers qui sont inférieurs ou égaux au nombre réel x. Énumérons les nombres premiers $\mathbb{P} = \{p_1, p_2, p_3, \ldots\}$ dans l'ordre croissant. Considérons le logarithme naturel $\ln x$, défini par $\ln x = \int_1^x \frac{1}{t} dt$. Comparons maintenant l'aire qui se trouve sous le graphe de $f(t) = \frac{1}{t}$ avec une fonction en escalier qui se trouve au dessus (voir aussi l'appendice en page 11 à propos de cette méthode). Si $n \leqslant x < n + 1$ nous avons :

$$\begin{aligned} \ln x &\leqslant 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n-1} + \frac{1}{n} \\ &\leqslant \sum \frac{1}{m} \quad \text{où la somme s'étend à tous les } m \in \mathbb{N}^* \text{ qui n'ont} \\ &\qquad\qquad \text{que des diviseurs premiers } p \leqslant x. \end{aligned}$$

Puisque chaque m s'écrit de manière *unique* comme un produit de la forme $\prod_{p \leqslant x} p^{k_p}$, nous voyons que la dernière somme est égale à :

$$\prod_{\substack{p \in \mathbb{P} \\ p \leqslant x}} \Big(\sum_{k \geqslant 0} \frac{1}{p^k}\Big)$$

La somme qui se trouve à l'intérieur est une série géométrique de raison $\frac{1}{p}$ et par conséquent :

$$\ln x \leqslant \prod_{\substack{p \in \mathbb{P} \\ p \leqslant x}} \frac{1}{1 - \frac{1}{p}} = \prod_{\substack{p \in \mathbb{P} \\ p \leqslant x}} \frac{p}{p-1} = \prod_{k=1}^{\pi(x)} \frac{p_k}{p_k - 1}$$

Il est maintenant clair que $p_k \geqslant k+1$, donc :

$$\frac{p_k}{p_k - 1} \;=\; 1 + \frac{1}{p_k - 1} \;\leqslant\; 1 + \frac{1}{k} \;=\; \frac{k+1}{k}$$

et par suite :

$$\ln x \;\leqslant\; \prod_{k=1}^{\pi(x)} \frac{k+1}{k} \;=\; \pi(x) + 1$$

Tout le monde sait que $\ln x$ n'est pas borné. On en conclut que $\pi(x)$ est également non borné et qu'il existe une infinité de nombres premiers. □

■ **Cinquième Preuve.** Après l'analyse, la topologie ! Considérons la curieuse topologie définie de la manière suivante sur l'ensemble $\mathbb{Z}$ des entiers. Pour $a, b \in \mathbb{Z}$, $b > 0$, posons :

$$N_{a,b} = \{a + nb : n \in \mathbb{Z}\}$$

Chaque ensemble $N_{a,b}$ est une progression arithmétique infinie des deux côtés. Nous disons qu'un ensemble $O \subseteq \mathbb{Z}$ est *ouvert* si O est vide, ou si pour chaque $a \in O$ il existe $b > 0$ tel que $N_{a,b} \subseteq O$. Il est clair qu'une réunion d'ensembles ouverts est encore un ouvert. Si O_1, O_2 sont ouverts et que $a \in O_1 \cap O_2$ vérifie $N_{a,b_1} \subseteq O_1$ et $N_{a,b_2} \subseteq O_2$, alors $a \in N_{a,b_1 b_2} \subseteq O_1 \cap O_2$. Il en résulte que toute intersection finie d'ensembles ouverts est encore ouverte. Cette famille d'ouverts induit donc une véritable topologie sur $\mathbb{Z}$.

Notons les deux résultats suivants :
(A) Tout ensemble ouvert non vide est infini.
(B) Tout ensemble $N_{a,b}$ est fermé.
Le premier résultat est une conséquence de la définition. Pour le deuxième, on observe que :

$$N_{a,b} \;=\; \mathbb{Z} \setminus \bigcup_{i=1}^{b-1} N_{a+i,b}$$

ce qui prouve que $N_{a,b}$ est le complémentaire d'un ensemble ouvert et qu'il est donc fermé.

Jusqu'à présent, les nombres premiers n'interviennent pas mais ils arrivent ici. Puisque tout nombre $n \neq 1, -1$ a un diviseur premier p — et qu'il est donc contenu dans $N_{0,p}$ — nous pouvons affirmer que :

$$\mathbb{Z} \setminus \{1, -1\} \;=\; \bigcup_{p \in \mathbb{P}} N_{0,p}$$

Toutefois, si $\mathbb{P}$ était fini, $\bigcup_{p\in\mathbb{P}} N_{0,p}$ serait une réunion finie d'ensembles fermés (d'après le résultat (B)) et serait donc fermé. Par conséquent, $\{1, -1\}$ serait un ensemble ouvert, ce qui contredit le résultat (A). □

« Jeter des galets, indéfiniment ».

■ **Sixième Preuve.** Cette dernière preuve représente un pas en avant considérable car elle démontre non seulement qu'il y a une infinité de nombres

premiers mais aussi que la série $\sum_{p\in\mathbb{P}} \frac{1}{p}$ diverge. La première preuve de ce résultat important a été énoncée par Euler (elle est intéressante en elle-même) mais la preuve donnée ici, inventée par Erdős, est d'une irrésistible beauté.

Soit $p_1, p_2, p_3, \ldots$ la suite des nombres premiers écrite dans l'ordre croissant. Supposons que $\sum_{p\in\mathbb{P}} \frac{1}{p}$ converge. Il doit donc y avoir un entier naturel k tel que $\sum_{i\geqslant k+1} \frac{1}{p_i} < \frac{1}{2}$. Appelons $p_1, \ldots, p_k$ les *petits* nombres premiers et $p_{k+1}, p_{k+2}, \ldots$ les *grands* nombres premiers. Si N est un entier naturel arbitraire, nous avons donc :

$$\sum_{i\geqslant k+1} \frac{N}{p_i} \;<\; \frac{N}{2} \tag{1}$$

Soit N_b le nombre d'entiers positifs $n \leqslant N$ divisibles par un grand nombre premier au moins et N_s le nombre d'entiers positifs $n \leqslant N$ qui n'ont que des petits diviseurs premiers. Nous allons montrer que pour un N convenable :

$$N_b + N_s \;<\; N$$

ce qui aboutira à la contradiction souhaitée, puisque par définition $N_b + N_s$ devrait être égal à N.

Pour estimer N_b, notons que $\lfloor \frac{N}{p_i} \rfloor$ dénombre les entiers positifs $n \leqslant N$ qui sont des multiples de p_i. Par conséquent, d'après (1) on obtient :

$$N_b \;\leqslant\; \sum_{i\geqslant k+1} \left\lfloor \frac{N}{p_i} \right\rfloor \;<\; \frac{N}{2} \tag{2}$$

Examinons maintenant N_s. Écrivons chaque $n \leqslant N$ n'ayant que des petits diviseurs premiers sous la forme $n = a_n b_n^2$, où a_n est le facteur sans carré. Chaque a_n est ainsi un produit de petits nombres premiers *différents*. On en déduit qu'il y a exactement 2^k facteurs sans carré différents. Par ailleurs, comme $b_n \leqslant \sqrt{n} \leqslant \sqrt{N}$, il y a au plus $\sqrt{N}$ facteurs différents qui sont des carrés . Ainsi :

$$N_s \;\leqslant\; 2^k \sqrt{N}$$

Puisque l'on a établi l'inégalité (2) pour *tout* N, il reste à déterminer un nombre N tel que $2^k\sqrt{N} \leqslant \frac{N}{2}$, c'est-à-dire $2^{k+1} \leqslant \sqrt{N}$. $N = 2^{2k+2}$ convient. □

Bibliographie

[1] B. ARTMANN : *Euclid — The Creation of Mathematics,* Springer-Verlag, New York 1999.

[2] P. ERDŐS : *Über die Reihe* $\sum \frac{1}{p}$, Mathematica, Zutphen B **7** (1938), 1-2.

[3] L. EULER : *Introductio in Analysin Infinitorum,* Tomus Primus, Lausanne 1748 ; Opera Omnia, Ser. 1, Vol. 8.

[4] H. FÜRSTENBERG : *On the infinitude of primes,* Amer. Math. Monthly **62** (1955), 353.

Le postulat de Bertrand

Chapitre 2

Nous avons vu que la suite des nombres premiers $2, 3, 5, 7, \ldots$ est infinie. Pour nous convaincre que la distance séparant deux nombres premiers consécutifs n'est pas bornée, considérons le produit $N := 2.3.5.\cdots.p$ de tous les nombres premiers inférieurs à $k+2$. Remarquons qu'aucun des k nombres

$$N+2, N+3, N+4, \ldots, N+k, N+(k+1)$$

n'est premier, puisque pour $2 \leqslant i \leqslant k+1$ nous savons que i a un facteur premier plus petit que $k+2$, que ce facteur divise également N, et donc aussi $N+i$. Par ce biais, nous trouvons, par exemple pour $k = 10$ qu'aucun des dix nombres :

$$2312, 2313, 2314, \ldots, 2321$$

n'est premier.

La distance séparant deux nombres premiers consécutifs peut également être majorée. Il existe une borne célèbre qui se définit comme suit : « l'écart entre un nombre premier et le nombre premier suivant ne peut pas être supérieur au nombre duquel on est parti ». Ce principe est connu sous le nom du postulat de Bertrand, puisqu'il a été supposé et vérifié empiriquement pour $n < 3\,000\,000$ par Joseph Bertrand. Il a été démontré pour la première fois pour tout n par Pafnuty Chebyshev en 1850. Une preuve plus simple a été fournie par le génie indien Ramanujan. Notre Preuve du Grand Livre est due à Paul Erdős : elle est extraite du premier article publié par Erdős, paru en 1932 alors qu'il avait 19 ans.

Le postulat de Bertrand
Pour tout $n \geqslant 1$, il existe un nombre premier p tel que $n < p \leqslant 2n$.

■ **Preuve.** Nous allons estimer l'ordre de grandeur du coefficient binomial $\binom{2n}{n}$ de façon suffisamment précise pour nous assurer que, s'il n'admettait aucun facteur premier dans l'ensemble $\{n+1, \ldots, 2n\}$, il serait « trop petit ». Notre argumentation se décompose en cinq étapes.

(1) Nous montrons d'abord le postulat de Bertrand pour $n < 4000$. Pour cela, on n'a pas besoin de vérifier 4000 cas. Il suffit (le procédé est connu sous le nom d'« astuce de Landau ») de vérifier que :

$$2, 3, 5, 7, 13, 23, 43, 83, 163, 317, 631, 1259, 2503, 4001$$

Joseph Bertrand

Beweis eines Satzes von Tschebyschef.

Von P. Erdős in Budapest.

Für den zuerst von Tschebyschef bewiesenen Satz, laut dessen es zwischen einer natürlichen Zahl und ihrer zweifachen stets wenigstens eine Primzahl gibt, liegen in der Literatur mehrere Beweise vor. Als einfachsten kann man ohne Zweifel den Beweis von Ramanujan[1]) bezeichnen. In seinem Werk *Vorlesungen über Zahlentheorie* (Leipzig, 1927), Band I, S. 66—68 gibt Herr Landau einen besonders einfachen Beweis für einen Satz über die Anzahl der Primzahlen unter einer gegebenen Grenze, aus welchem unmittelbar folgt, daß für ein geeignetes q zwischen einer natürlichen Zahl und ihrer q-fachen stets eine Primzahl liegt. Für die augenblicklichen Zwecken des Herrn Landau kommt es nicht auf die numerische Bestimmung der im Beweis auftretenden Konstanten an; man überzeugt sich aber durch eine numerische Verfolgung des Beweises leicht, daß q jedenfalls größer als 2 ausfällt.

In den folgenden Zeilen werde ich zeigen, daß man durch eine Verschärfung der dem Landauschen Beweis zugrunde liegenden Ideen zu einem Beweis des oben erwähnten Tschebyschefschen Satzes gelangen kann, der — wie mir scheint — an Einfachheit nicht hinter dem Ramanujanschen Beweis steht. Griechische Buchstaben sollen im Folgenden durchwegs positive, lateinische Buchstaben natürliche Zahlen bezeichnen, die Bezeichnung p ist für Primzahlen vorbehalten.

1. Der Binomialkoeffizient

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2}$$

[1]) Siehe Ramanujan, A Proof of Bertrand's Postulate. *Journal of the Indian Mathematical Society*, 11 (1919), S. 181—182. — *Collected Papers of Srinivasa Ramanujan* (Cambridge, 1927), S. 208—276.

L'article publié par Erdős en 1932.

est une suite de nombres premiers tels que chacun d'entre eux est plus petit que le double du précédent. Par conséquent, chaque intervalle $\{y : n < y \leqslant 2n\}$ tel que $n \leqslant 4000$, contient un de ces 14 nombres premiers.

(2) Nous montrons ensuite que :

$$\prod_{p \leqslant x} p \ \leqslant \ 4^{x-1} \qquad \text{pour tout réel } x \geqslant 2 \tag{1}$$

où notre notation — ici et dans ce qui suit — signifie que le produit s'étend à tous les nombres *premiers* $p \leqslant x$. Pour cela, nous utilisons une démonstration par récurrence sur le cardinal de ces nombres premiers. Elle n'est pas tirée de l'article original d'Erdős, mais elle est encore due à Erdős (voir en marge). C'est réellement une Preuve du Grand Livre. Nous notons tout d'abord que si q est le plus grand nombre premier tel que $q \leqslant x$, alors :

$$\prod_{p \leqslant x} p \ = \ \prod_{p \leqslant q} p \qquad \text{et} \qquad 4^{q-1} \ \leqslant \ 4^{x-1}$$

Il suffit donc de vérifier (1) lorsque $x = q$ est un nombre premier. Pour $q = 2$ on obtient $2 \leqslant 4$ et l'on continue ensuite avec les nombres premiers impairs $q = 2m+1$. Pour ces nombres, on sépare le produit en deux parties et calculons :

$$\prod_{p \leqslant 2m+1} p \ = \ \prod_{p \leqslant m+1} p \ . \ \prod_{m+1 < p \leqslant 2m+1} p \ \leqslant \ 4^m \binom{2m+1}{m} \ \leqslant \ 4^m 2^{2m} \ = \ 4^{2m}$$

Tous les éléments de cette ligne de calcul sont faciles à vérifier. En effet :

$$\prod_{p \leqslant m+1} p \ \leqslant \ 4^m$$

se montre par récurrence.

L'inégalité :

$$\prod_{m+1 < p \leqslant 2m+1} p \ \leqslant \ \binom{2m+1}{m}$$

est une conséquence du fait que $\binom{2m+1}{m} = \frac{(2m+1)!}{m!(m+1)!}$ est un entier et que les nombres premiers que nous considérons sont tous des facteurs du numérateur $(2m+1)!$ mais pas du dénominateur $m!(m+1)!$. Enfin, on a :

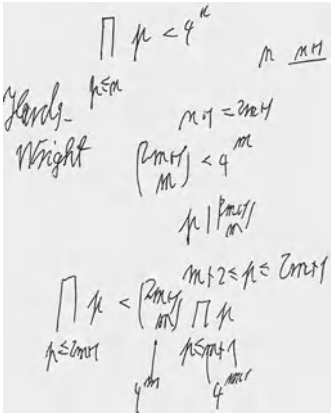
$$\binom{2m+1}{m} \ \leqslant \ 2^{2m}$$

puisque :

$$\binom{2m+1}{m} \quad \text{et} \quad \binom{2m+1}{m+1}$$

sont deux termes (égaux !) qui apparaissent dans la somme :

$$\sum_{k=0}^{2m+1} \binom{2m+1}{k} = 2^{2m+1}$$



Notes manuscrites d'Erdős.

Le théorème de Legendre

Le nombre $n!$ contient le facteur premier p exactement

$$\sum_{k \geqslant 1} \left\lfloor \frac{n}{p^k} \right\rfloor$$

fois.

■ **Preuve.** Il y a exactement $\lfloor \frac{n}{p} \rfloor$ termes parmi les facteurs de $n! = 1.2.3. \cdots .n$ qui sont divisibles par p et qui fournissent $\lfloor \frac{n}{p} \rfloor$ p-facteurs. Il y a $\lfloor \frac{n}{p^2} \rfloor$ termes parmi les facteurs de $n!$ qui sont divisibles par p^2 et qui fournissent les $\lfloor \frac{n}{p^2} \rfloor$ facteurs premiers p suivants de $n!$, *etc.* □

(3) Nous déduisons du théorème de Legendre (voir encadré) que $\binom{2n}{n} = \frac{(2n)!}{n!n!}$ contient le facteur premier p exactement :

$$\sum_{k\geqslant 1}\left(\left\lfloor\frac{2n}{p^k}\right\rfloor - 2\left\lfloor\frac{n}{p^k}\right\rfloor\right)$$

fois. Ici, chaque terme de la somme est au plus égal à 1, puisqu'il vérifie :

$$\left\lfloor\frac{2n}{p^k}\right\rfloor - 2\left\lfloor\frac{n}{p^k}\right\rfloor \;<\; \frac{2n}{p^k} - 2\left(\frac{n}{p^k} - 1\right) \;=\; 2$$

et qu'il est entier. En outre, les termes de la somme disparaissent dès que $p^k > 2n$.
Ainsi, $\binom{2n}{n}$ contient p exactement :

$$\sum_{k\geqslant 1}\left(\left\lfloor\frac{2n}{p^k}\right\rfloor - 2\left\lfloor\frac{n}{p^k}\right\rfloor\right) \;\leqslant\; \max\{r : p^r \leqslant 2n\}$$

fois. Par conséquent, la plus grande puissance de p qui divise $\binom{2n}{n}$ n'est pas supérieure à $2n$. En particulier, les nombres premiers $p > \sqrt{2n}$ apparaissent au plus une fois dans la factorisation de $\binom{2n}{n}$.
En outre — et d'après Erdős, c'est le point clé de sa preuve — les nombres premiers p qui vérifient $\frac{2}{3}n < p \leqslant n$ ne divisent pas $\binom{2n}{n}$. En effet, $3p > 2n$ implique (pour $n \geqslant 3$, et par conséquent $p \geqslant 3$) que p et $2p$ sont les seuls multiples de p qui apparaissent comme facteurs du numérateur de $\frac{(2n)!}{n!n!}$, alors qu'il y a deux p-facteurs au dénominateur.

Des exemples comme :
$\binom{26}{13} = 2^3.5^2.7.17.19.23$
$\binom{28}{14} = 2^3.3^3.5^2.17.19.23$
$\binom{30}{15} = 2^4.3^2.5.17.19.23.29$
illustrent le fait que de « tout petits » facteurs premiers $p < \sqrt{2n}$ peuvent apparaître avec des puissances élevées dans $\binom{2n}{n}$, de « petits » nombres premiers tels que $\sqrt{2n} < p \leqslant \frac{2}{3}n$ apparaissent au plus une fois, alors que des facteurs tels que $\frac{2}{3}n < p \leqslant n$ n'apparaissent pas du tout.

(4) Nous sommes désormais prêts pour estimer $\binom{2n}{n}$. Si $n \geqslant 3$, en utilisant une estimation de la page 12 pour la borne inférieure, nous obtenons :

$$\frac{4^n}{2n} \;\leqslant\; \binom{2n}{n} \;\leqslant\; \prod_{p\leqslant\sqrt{2n}} 2n \;\cdot \prod_{\sqrt{2n}<p\leqslant\frac{2}{3}n} p \;\cdot \prod_{n<p\leqslant 2n} p$$

Ainsi, puisqu'il n'y a pas plus de $\sqrt{2n}$ nombres premiers $p \leqslant \sqrt{2n}$,

$$4^n \;\leqslant\; (2n)^{1+\sqrt{2n}} \cdot \prod_{\sqrt{2n}<p\leqslant\frac{2}{3}n} p \;\cdot \prod_{n<p\leqslant 2n} p \qquad \text{pour} \quad n \geqslant 3 \qquad (2)$$

(5) Supposons maintenant qu'il n'y ait pas de nombre premier p tel que $n < p \leqslant 2n$. Le deuxième produit dans (2) vaut donc 1. En substituant (1) dans (2), nous obtenons :

$$4^n \;\leqslant\; (2n)^{1+\sqrt{2n}}\, 4^{\frac{2}{3}n}$$

ou :

$$4^{\frac{1}{3}n} \;\leqslant\; (2n)^{1+\sqrt{2n}} \qquad (3)$$

ce qui est faux dès que n est suffisamment grand ! En fait, en utilisant la relation $a+1 < 2^a$ (que l'on peut montrer par récurrence pour tout $a \geqslant 2$), nous obtenons :

$$2n = \left(\sqrt[6]{2n}\right)^6 < \left(\lfloor \sqrt[6]{2n} \rfloor + 1\right)^6 < 2^{6\left\lfloor \sqrt[6]{2n} \right\rfloor} \leqslant 2^{6\sqrt[6]{2n}} \tag{4}$$

Ainsi, si $n \geqslant 50$ (et par conséquent $18 < 2\sqrt{2n}$), on déduit de (3) et de (4) que :

$$2^{2n} \leqslant (2n)^{3\left(1+\sqrt{2n}\right)} < 2^{\sqrt[6]{2n}\left(18+18\sqrt{2n}\right)} < 2^{20\sqrt[6]{2n}\sqrt{2n}} = 2^{20(2n)^{2/3}}$$

Cela implique $(2n)^{1/3} < 20$ et donc $n < 4000$. □

On peut déduire bien plus de ce type d'estimations : avec les mêmes méthodes , on peut déduire de (2) que :

$$\prod_{n<p\leqslant 2n} p \geqslant 2^{\frac{1}{30}n} \quad \text{si} \quad n \geqslant 4000$$

et qu'il y a donc au moins :

$$\log_{2n}\left(2^{\frac{1}{30}n}\right) = \frac{1}{30}\,\frac{n}{\log_2 n + 1}$$

nombres premiers dans l'intervalle compris entre n et $2n$.

Cette estimation n'est pas si mauvaise : le « vrai » nombre d'entiers premiers contenus dans cet intervalle est *grosso modo* $n/\ln n$. C'est une conséquence du « théorème des nombres premiers », qui dit que la limite :

$$\lim_{n\to\infty} \frac{\text{Card}\{p \leqslant n : p \text{ est premier}\}}{n/\ln n}$$

existe et est égale à 1. Ce résultat célèbre a été prouvé pour la première fois par Hadamard et de la Vallée-Poussin en 1896 ; Selberg et Erdős ont mis au point une preuve élémentaire (ne faisant intervenir aucun outil d'analyse complexe, mais la démonstration demeure encore longue et compliquée) en 1948.

En ce qui concerne le théorème des nombres premiers lui-même, on peut encore espérer des progrès : par exemple, une preuve de l'hypothèse de Riemann (voir page 56), un des problèmes ouverts majeurs en mathématiques, fournirait également une amélioration substantielle des estimations du théorème des nombres premiers. En ce qui concerne le postulat de Bertrand, on peut aussi attendre des améliorations spectaculaires. En fait, le problème suivant n'est toujours pas résolu [3, p. 19] :

Y a-t-il toujours un nombre premier entre n^2 *et* $(n+1)^2$ *?*

Appendice - Quelques estimations

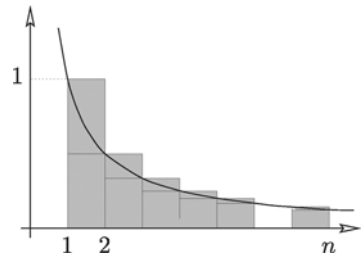
Comparaison à une intégrale

Il existe une méthode très simple, mais efficace, pour estimer certaines sommes au moyen d'intégrales (comme celle déjà rencontrée à la page 4). Pour estimer les *nombres harmoniques* :

$$H_n \;=\; \sum_{k=1}^{n} \frac{1}{k}$$

nous dessinons la figure qui se trouve dans la marge et déduisons de celle-ci :

$$H_n - 1 \;=\; \sum_{k=2}^{n} \frac{1}{k} \;<\; \int_1^n \frac{1}{t}\,dt \;=\; \ln n$$



en comparant l'aire qui se trouve sous le graphe $f(t) = \frac{1}{t}$ $(1 \leqslant t \leqslant n)$ à celle des rectangles ombrés, puis :

$$H_n - \frac{1}{n} \;=\; \sum_{k=1}^{n-1} \frac{1}{k} \;>\; \int_1^n \frac{1}{t}\,dt \;=\; \ln n$$

en comparant à l'aire des grands rectangles (en incluant les parties légèrement ombrées). Ces deux résultats impliquent que :

$$\ln n + \frac{1}{n} \;<\; H_n \;<\; \ln n \;+\; 1\,.$$

En particulier, $\lim\limits_{n\to\infty} H_n \to \infty$, et l'ordre de croissance de H_n est donné par $\lim\limits_{n\to\infty} \frac{H_n}{\ln n} = 1$. En fait, on connaît de bien meilleures estimations (voir [2]), comme :

$$H_n \;=\; \ln n + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4} + O\left(\frac{1}{n^6}\right)$$

Ici $O\left(\frac{1}{n^6}\right)$ désigne une fonction $f(n)$ telle que l'on ait $f(n) \leqslant c\frac{1}{n^6}$ où c est une constante.

où $\gamma \approx 0.5772$ est la *constante d'Euler*.

Estimation des factorielles – Formule de Stirling

La même méthode appliquée à :

$$\ln(n!) \;=\; \ln 2 + \ln 3 + \ldots + \ln n \;=\; \sum_{k=2}^{n} \ln k$$

conduit à :

$$\ln((n-1)!) \;<\; \int_1^n \ln t\,dt \;<\; \ln(n!)$$

où l'intégrale se calcule facilement :

$$\int_1^n \ln t\,dt \;=\; \Big[t\ln t - t\Big]_1^n \;=\; n\ln n - n + 1$$

Ainsi, nous obtenons un minorant de $n!$:

$$n! \;>\; e^{n\ln n - n + 1} \;=\; e\Big(\frac{n}{e}\Big)^n$$

et aussi un majorant :

$$n! \;=\; n\,(n-1)! \;<\; ne^{n\ln n - n + 1} \;=\; en\Big(\frac{n}{e}\Big)^n$$

Ici $f(n) \sim g(n)$ signifie que $\lim\limits_{n\to\infty} \dfrac{f(n)}{g(n)} = 1$.

On aurait besoin d'une étude plus précise pour obtenir l'équivalent de $n!$, donné par la *formule de Stirling* :

$$n! \;\sim\; \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n$$

Là encore, il existe des résultats plus précis comme :

$$n! \;=\; \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n \left(1 + \frac{1}{12n} + \frac{1}{288n^2} - \frac{139}{5140n^3} + O\left(\frac{1}{n^4}\right)\right)$$

Estimation des coefficients binomiaux

En se fondant sur la définition des coefficients binomiaux selon laquelle $\binom{n}{k}$ est le nombre de k-parties d'un n-ensemble, on sait que la suite $\binom{n}{0}$, $\binom{n}{1}$, …, $\binom{n}{n}$ de coefficients binomiaux

- a pour somme $\sum\limits_{k=0}^{n} \binom{n}{k} = 2^n$ et
- est symétrique : $\binom{n}{k} = \binom{n}{n-k}$.

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
 & 1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 \\
1 & & 7 & & 21 & & 35 & & 35 & & 21 & & 7 & & 1
\end{array}$$

Le triangle de Pascal.

À partir de l'équation fonctionnelle $\binom{n}{k} = \frac{n-k+1}{k}\binom{n}{k-1}$, il est facile de voir que, pour tout n, les coefficients binomiaux $\binom{n}{k}$ forment une suite qui est symétrique : elle croît vers son milieu, de sorte que les coefficients binomiaux du milieu sont les plus grands de la suite :

$$1 = \tbinom{n}{0} < \tbinom{n}{1} < \ldots < \tbinom{n}{\lfloor n/2\rfloor} = \tbinom{n}{\lceil n/2\rceil} > \ldots > \tbinom{n}{n-1} > \tbinom{n}{n} = 1$$

La notation $\lfloor x\rfloor$ (respectivement $\lceil x\rceil$) désigne le nombre x arrondi à l'entier inférieur (respectivement supérieur) le plus proche.

À partir des développements asymptotiques de $n!$ mentionnés précédemment, on peut obtenir des estimations très précises des coefficients binomiaux. Cependant, dans la suite, nous n'aurons besoin que d'estimations assez modestes comme $\binom{n}{k} \leqslant 2^n$ pour tout k, alors que pour $n \geqslant 2$, on a :

$$\binom{n}{\lfloor n/2\rfloor} \;\geqslant\; \frac{2^n}{n}$$

l'égalité ayant lieu seulement si $n = 2$. En particulier pour $n \geqslant 1$, on a :

$$\binom{2n}{n} \geqslant \frac{4^n}{2n}$$

En effet, $\binom{n}{\lfloor n/2 \rfloor}$, coefficient binomial médian, est le plus grand élément de la suite $\binom{n}{0}, \binom{n}{n}, \binom{n}{1}, \binom{n}{2}, \ldots, \binom{n}{n-1}$, dont la somme est 2^n et dont la moyenne est donc $\frac{2^n}{n}$.
Par ailleurs, on peut majorer de manière élémentaire les coefficients binomiaux en écrivant :

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \leqslant \frac{n^k}{k!} \leqslant \frac{n^k}{2^{k-1}}$$

ce qui est une estimation raisonnable pour les « petits » coefficients binomiaux aux extrémités de la suite, lorsque n est grand (par rapport à k).

Bibliographie

[1] P. ERDŐS : *Beweis eines Satzes von Tschebyschef,* Acta Sci. Math. (Szeged) **5** (1930-32), 194-198.

[2] R. L. GRAHAM, D. E. KNUTH & O. PATASHNIK : *Concrete Mathematics. A Foundation for Computer Science,* Addison-Wesley, Reading MA 1989.

[3] G. H. HARDY & E. M. WRIGHT : *An Introduction to the Theory of Numbers,* fifth edition, Oxford University Press 1979.

Les coefficients binomiaux ne sont (presque) jamais des puissances

Chapitre 3

Il existe un épilogue au postulat de Bertrand qui conduit à un beau résultat sur les coefficients binomiaux. En 1892, Sylvester a amélioré le postulat de Bertrand de la manière suivante :

> *Si $n \geqslant 2k$, alors l'un au moins des nombres $n, n-1, \ldots, n-k+1$ a un diviseur premier p plus grand que k.*

Notons que, pour $n = 2k$, nous obtenons précisément le postulat de Bertrand. En 1934, Erdős a établi une Preuve du Grand Livre courte et élémentaire du résultat de Sylvester, en suivant les grandes lignes de sa démonstration du postulat de Bertrand. Voici une manière équivalente d'énoncer le théorème de Sylvester :

> *Le coefficient binomial*
>
> $$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \qquad (n \geqslant 2k)$$
>
> *a toujours un facteur premier $p > k$.*

Gardant cette observation à l'esprit, nous abordons un nouveau bijou d'Erdős. Quand $\binom{n}{k}$ est-il égal à une puissance m^ℓ ? Il est facile de voir que l'équation $\binom{n}{2} = m^2$ admet une infinité de solutions. En effet, si $\binom{n}{2}$ est un carré, alors $\binom{(2n-1)^2}{2}$ en est un également. Pour s'en convaincre, il suffit de poser $n(n-1) = 2m^2$. Cela implique que :

$$(2n-1)^2((2n-1)^2-1) = (2n-1)^2 4n(n-1) = 2(2m(2n-1))^2$$

et par conséquent :

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

En commençant avec $\binom{9}{2} = 6^2$, nous obtenons une infinité de solutions ; la suivante est $\binom{289}{2} = 204^2$. Cependant, cela ne donne pas toutes les solutions. Par exemple, $\binom{50}{2} = 35^2$ est le début d'une autre série, ainsi que $\binom{1682}{2} = 1189^2$. Si $k = 3$, on sait que $\binom{n}{3} = m^2$ a pour unique solution $n = 50$, $m = 140$. Nous sommes au bout du chemin. Pour $k \geqslant 4$ et pour tout $\ell \geqslant 2$, il n'existe aucune solution. C'est ce qu'Erdős a prouvé par un argument ingénieux.

$\binom{50}{3} = 140^2$ est la seule solution pour $k = 3$, $\ell = 2$.

Théorème. *L'équation* $\binom{n}{k} = m^\ell$ *n'a pas de solution entière pour* $\ell \geqslant 2$ *et* $4 \leqslant k \leqslant n-4$.

■ **Preuve.** Remarquons d'abord que l'on peut supposer $n \geqslant 2k$ puisque $\binom{n}{k} = \binom{n}{n-k}$. Supposons que le théorème soit faux et que $\binom{n}{k} = m^\ell$. La preuve, par l'absurde, se construit autour des quatre points suivants.

(1) Selon le théorème de Sylvester, il existe un facteur premier p de $\binom{n}{k}$ plus grand que k. Par conséquent, p^ℓ divise $n(n-1)\cdots(n-k+1)$. Il est clair qu'un seul des facteurs $n-i$ peut être multiple de p (puisque $p > k$). On en déduit que $p^\ell \mid n-i$ et donc que :

$$n \;\geqslant\; p^\ell \;>\; k^\ell \;\geqslant\; k^2 .$$

(2) Considérons un facteur quelconque $n-j$ du numérateur et notons-le sous la forme $n-j = a_j m_j^\ell$, où a_j n'est divisible par aucune puissance ℓ-ième non triviale. Nous remarquons grâce à **(1)** que a_j n'a que des diviseurs premiers inférieurs ou égaux à k. Nous voulons montrer ensuite que $a_i \neq a_j$ pour $i \neq j$. Supposons, au contraire, que $a_i = a_j$ pour un certain $i < j$. Alors $m_i \geqslant m_j + 1$ et :

$$\begin{aligned} k \;&>\; (n-i)-(n-j) \;=\; a_j(m_i^\ell - m_j^\ell) \;\geqslant\; a_j((m_j+1)^\ell - m_j^\ell) \\ &>\; a_j \ell m_j^{\ell-1} \;\geqslant\; \ell(a_j m_j^\ell)^{1/2} \;\geqslant\; \ell(n-k+1)^{1/2} \\ &\geqslant\; \ell\left(\tfrac{n}{2}+1\right)^{1/2} \;>\; n^{1/2} \end{aligned}$$

ce qui contredit l'inégalité $n > k^2$ établie ci-dessus.

(3) Nous prouvons ensuite que les a_i sont les entiers $1, 2, \ldots, k$ écrits dans un certain ordre (ce qui constitue pour Erdős, le point crucial de la preuve). Puisque nous savons déjà qu'ils sont tous distincts, il suffit de prouver que :

$$a_0 a_1 \cdots a_{k-1} \quad \text{divise} \quad k!$$

En substituant $n-j = a_j m_j^\ell$ dans l'équation $\binom{n}{k} = m^\ell$, on trouve :

$$a_0 a_1 \cdots a_{k-1}(m_0 m_1 \cdots m_{k-1})^\ell \;=\; k!\, m^\ell$$

En simplifiant par les facteurs communs à $m_0 \cdots m_{k-1}$ et m, on obtient :

$$a_0 a_1 \cdots a_{k-1} u^\ell \;=\; k!\, v^\ell$$

avec $\operatorname{pgcd}(u, v) = 1$. Il reste à montrer que $v = 1$. Si tel n'était pas le cas, v contiendrait un diviseur premier p. Puisque $\operatorname{pgcd}(u, v) = 1$, p doit être un diviseur premier de $a_0 a_1 \cdots a_{k-1}$ et, par conséquent, il est inférieur ou égal à k. D'après le théorème de Legendre (voir page 8) nous savons que $k!$ contient p à la puissance $\sum_{i \geqslant 1} \lfloor \frac{k}{p^i} \rfloor$. Estimons maintenant l'exposant de p dans $n(n-1)\cdots(n-k+1)$. Soit i un entier positif et soient $b_1 < b_2 < \ldots < b_s$ les multiples de p^i qui se trouvent parmi $n, n-1, \ldots, n-k+1$.

Alors $b_s = b_1 + (s-1)p^i$ et par conséquent :

$$(s-1)p^i \;=\; b_s - b_1 \;\leqslant\; n-(n-k+1) \;=\; k-1$$

ce qui implique :

$$s \;\leqslant\; \Big\lfloor \frac{k-1}{p^i} \Big\rfloor + 1 \;\leqslant\; \Big\lfloor \frac{k}{p^i} \Big\rfloor + 1$$

Ainsi, pour chaque i, le nombre de multiples de p^i qui se trouvent parmi $n, \ldots, n-k+1$, et par conséquent parmi les a_j, est borné par $\lfloor \frac{k}{p^i} \rfloor + 1$. Cela implique que l'exposant de p dans $a_0 a_1 \cdots a_{k-1}$ est au plus :

$$\sum_{i=1}^{\ell-1} \Big(\Big\lfloor \frac{k}{p^i} \Big\rfloor + 1 \Big)$$

grâce au raisonnement que nous avons utilisé pour le théorème de Legendre dans le chapitre 2. La seule différence est que, cette fois, la somme s'arrête à $i = \ell - 1$, puisque les a_j ne contiennent pas de puissance ℓ-ième.
En considérant ces deux résultats simultanément, nous concluons que l'exposant de p dans v^ℓ est au plus :

$$\sum_{i=1}^{\ell-1} \Big(\Big\lfloor \frac{k}{p^i} \Big\rfloor + 1 \Big) - \sum_{i \geqslant 1} \Big\lfloor \frac{k}{p^i} \Big\rfloor \;\leqslant\; \ell - 1$$

et nous obtenons la contradiction souhaitée, puisque v^ℓ est une puissance ℓ-ième.
Cela suffit déjà à régler le cas $\ell = 2$. En effet, puisque $k \geqslant 4$, l'un des a_i doit être égal à 4, mais les a_i ne contiennent pas de carré. Supposons donc maintenant que $\ell \geqslant 3$.

On constate que notre analyse est en accord avec $\binom{50}{3} = 140^2$, puisque :
$50 = 2 \cdot 5^2$
$49 = 1 \cdot 7^2$
$48 = 3 \cdot 4^2$
et $5 \cdot 7 \cdot 4 = 140$.

(4) Puisque $k \geqslant 4$, nous devons avoir $a_{i_1} = 1$, $a_{i_2} = 2$, $a_{i_3} = 4$ pour certains i_1, i_2, i_3, c'est-à-dire :

$$n - i_1 = m_1^\ell, \;\; n - i_2 = 2m_2^\ell, \;\; n - i_3 = 4m_3^\ell$$

Nous affirmons que $(n-i_2)^2 \neq (n-i_1)(n-i_3)$. Sinon, on pose $b = n - i_2$ et $n - i_1 = b - x$, $n - i_3 = b + y$, où $0 < |x|, |y| < k$. Par suite :

$$b^2 = (b-x)(b+y) \quad \text{ou} \quad (y-x)b = xy$$

où $x = y$ est clairement impossible. Nous avons maintenant grâce à la partie **(1)** :

$$|xy| \;=\; b|y-x| \;\geqslant\; b \;>\; n-k \;>\; (k-1)^2 \;\geqslant\; |xy|$$

ce qui est absurde.

Nous avons donc $m_2^2 \neq m_1 m_3$. Supposons que $m_2^2 > m_1 m_3$ (l'autre cas étant analogue) et écrivons nos dernières chaînes d'inégalités. Nous obtenons :

$$\begin{aligned} 2(k-1)n &> n^2 - (n-k+1)^2 > (n-i_2)^2 - (n-i_1)(n-i_3) \\ &= 4[m_2^{2\ell} - (m_1 m_3)^\ell] \geqslant 4[(m_1 m_3 + 1)^\ell - (m_1 m_3)^\ell] \\ &\geqslant 4\ell m_1^{\ell-1} m_3^{\ell-1} \end{aligned}$$

Puisque $\ell \geqslant 3$ et $n > k^\ell \geqslant k^3 > 6k$, cela implique :

$$\begin{aligned} 2(k-1)nm_1m_3 &> 4\ell m_1^\ell m_3^\ell = \ell(n-i_1)(n-i_3) \\ &> \ell(n-k+1)^2 > 3(n-\frac{n}{6})^2 > 2n^2 \end{aligned}$$

Puisque $m_i \leqslant n^{1/\ell} \leqslant n^{1/3}$, nous obtenons finalement :

$$kn^{2/3} \geqslant km_1m_3 > (k-1)m_1m_3 > n$$

ou $k^3 > n$. Avec cette contradiction, la preuve est terminée. □

Bibliographie

[1] P. ERDŐS : *A theorem of Sylvester and Schur,* J. London Math. Soc. **9** (1934), 282-288.

[2] P. ERDŐS : *On a diophantine equation,* J. London Math. Soc. **26** (1951), 176-178.

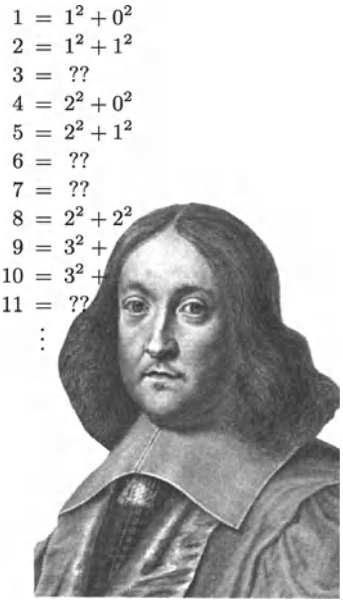
[3] J. J. SYLVESTER : *On arithmetical series,* Messenger of Math. **21** (1892), 1-19, 87-120 ; Collected Mathematical Papers Vol. 4, 1912, 687-731.

Représentation des nombres comme somme de deux carrés

Chapitre 4

Quels sont les nombres qui peuvent s'écrire comme somme de deux carrés ?

Cette question est aussi vieille que la théorie des nombres et sa solution est un classique dans ce domaine. La partie difficile de la solution consiste à vérifier que chaque nombre premier de la forme $4m+1$ est une somme de deux carrés. G. H. Hardy écrit que ce *théorème des deux carrés* de Fermat « est à juste titre considéré comme l'un des plus fins de l'arithmétique ». Néanmoins, la démonstration développée ci-dessous est relativement récente.



Pierre de Fermat

Commençons par quelques échauffements. En premier lieu, il convient de distinguer trois classes de nombres : le nombre premier $p = 2$, les nombres premiers de la forme $p = 4m + 1$ et les nombres premiers de la forme $p = 4m + 3$. Chaque nombre premier appartient exactement à l'une de ces trois classes. Sur ce point, nous pouvons noter (en utilisant une méthode « à la Euclide ») qu'il existe une infinité de nombres premiers de la forme $4m + 3$. En fait, s'il y en avait seulement un nombre fini, nous pourrions considérer p_k le plus grand nombre premier de ce type. En posant :

$$N_k \;:=\; 2^2.3.5.\cdots.p_k - 1$$

(où $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, ... désigne la suite de tous les nombres premiers), on trouverait alors que N_k est congru à $3 \pmod 4$. Il devrait donc avoir un facteur premier de la forme $4m + 3$ et ce facteur premier serait plus grand que p_k ce qui contredit la définition de p_k. À la fin de ce chapitre, nous verrons également qu'il existe une infinité de nombres premiers du type $p = 4m + 1$.

Notre premier lemme caractérise les nombres premiers pour lesquels -1 est un carré dans le corps $\mathbb{Z}_p$ (dont les propriétés sont résumées dans l'encadré de la page suivante). Il va aussi nous fournir un moyen rapide de constater qu'il existe un nombre infini de nombres premiers de la forme $4m + 1$.

Lemme 1. *Pour tout nombre premier $p = 4m+1$, l'équation $s^2 \equiv -1 \pmod p$ admet deux solutions $s \in \{1, 2, \ldots, p-1\}$, pour $p = 2$, l'équation admet une solution, et pour les nombres premiers de la forme $p = 4m+3$, l'équation n'admet pas de solution.*

■ **Preuve.** Pour $p = 2$, on prend $s = 1$. Pour p impair, nous construisons la relation d'équivalence sur $\{1, 2, \ldots, p-1\}$ qui est définie en associant

chaque élément à son symétrique pour l'addition et à son symétrique pour la multiplication dans $\mathbb{Z}_p$. Ainsi, les classes d'équivalence « générales » contiendront quatre éléments :

$$\{x, -x, \overline{x}, -\overline{x}\}$$

puisqu'un ensemble à 4 éléments de ce type contient les deux inverses de chacun de ses éléments. Cependant, il existe des classes d'équivalence plus petites lorsque ces quatre nombres ne sont pas distincts :

- $x \equiv -x$ est impossible si p est impair.
- $x \equiv \overline{x}$ est équivalent à $x^2 \equiv 1$. Cette équation a deux solutions, respectivement $x = 1$ et $x = p - 1$, conduisant à une classe d'équivalence $\{1, p-1\}$ de cardinal 2.
- $x \equiv -\overline{x}$ est équivalent à $x^2 \equiv -1$. Cette équation peut soit ne pas avoir de solution, soit avoir deux solutions distinctes : $x_0, p - x_0$ auquel cas la classe d'équivalence est $\{x_0, p - x_0\}$.

L'ensemble $\{1, 2, \ldots, p-1\}$ contient $p-1$ éléments. Nous l'avons partitionné en ensembles à 4 éléments (que nous appellerons quadruplets) et en une ou deux paires. Pour $p - 1 = 4m + 2$, nous constatons qu'il n'y a que la paire $\{1, p-1\}$, le reste est composé de quadruplets et donc $s^2 \equiv -1 \pmod{p}$ n'a pas de solution. Si $p - 1 = 4m$, il doit y avoir une deuxième paire, ce sont les deux solutions de $s^2 \equiv -1$ que nous cherchions. □

Si $p = 11$, la partition est : $\{1, 10\}$, $\{2, 9, 6, 5\}, \{3, 8, 4, 7\}$; si $p = 13$, c'est : $\{1, 12\}, \{2, 11, 7, 6\}$, $\{3, 10, 9, 4\}, \{5, 8\}$, la paire $\{5, 8\}$ fournissant les deux solutions de $s^2 \equiv -1 \bmod 13$.

Le lemme 1 signifie que tout nombre premier impair qui divise un nombre de la forme $M^2 + 1$ doit être de la forme $4m + 1$. Cela implique qu'il y a une infinité de nombre premiers de cette forme. En effet, si tel n'était pas le cas, désignons par q_k le plus grand nombre premier de la forme $4m + 1$ et considérons le nombre $2.3.5. \cdots .q_k)^2 + 1$. Le même raisonnement que précédemment conduit à une contradiction.

Corps premiers

Si p est un nombre premier, l'ensemble $\mathbb{Z}_p = \{0, 1, \ldots, p-1\}$ muni de l'addition et la multiplication « modulo p » est un corps fini. Nous aurons besoin des propriétés simples suivantes :

- Si $x \in \mathbb{Z}_p$, $x \neq 0$, le symétrique pour l'addition (habituellement noté $-x$) est $p - x \in \{1, 2, \ldots, p-1\}$. Si $p > 2$, x et $-x$ sont des éléments distincts de $\mathbb{Z}_p$.
- Tout $x \in \mathbb{Z}_p \backslash \{0\}$ a un inverse unique pour la multiplication noté $\overline{x} \in \mathbb{Z}_p \backslash \{0\}$ qui vérifie donc $x\overline{x} \equiv 1 \pmod{p}$.
  La définition des nombres premiers implique en effet que l'application $\mathbb{Z}_p \backslash \{0\} \to \mathbb{Z}_p \backslash \{0\}$, $z \mapsto xz$ est injective. $\mathbb{Z}_p \backslash \{0\}$ étant fini, elle est aussi surjective ; donc pour chaque $x \neq 0$ il existe un unique $\overline{x} \neq 0$ tel que $x\overline{x} \equiv 1 \pmod{p}$.
- Les carrés $0^2, 1^2, 2^2, \ldots, h^2$ définissent des éléments distincts de $\mathbb{Z}_p$, pour $h = \lfloor \frac{p}{2} \rfloor$.
  En effet, $x^2 \equiv y^2$, ou $(x + y)(x - y) \equiv 0$, implique que $x \equiv y$ ou que $x \equiv -y$. Les $1 + \lfloor \frac{p}{2} \rfloor$ éléments $0^2, 1^2, \ldots, h^2$ sont appelés des *carrés* dans $\mathbb{Z}_p$.

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

·	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

Addition et multiplication dans $\mathbb{Z}_5$

À ce stade, notons que pour *tous* les nombres premiers, l'équation $x^2+y^2 \equiv -1 \pmod p$ a des solutions. En effet, il y a $\lfloor \frac{p}{2} \rfloor + 1$ carrés distincts de la forme x^2 dans $\mathbb{Z}_p$ et $\lfloor \frac{p}{2} \rfloor + 1$ nombres distincts de la forme $-(1+y^2)$. Ces deux ensembles de nombres sont trop grands pour être disjoints, puisque $\mathbb{Z}_p$ n'a que p éléments. Il doit donc exister x et y tels que $x^2 \equiv -(1+y^2) \pmod p$.

Lemme 2. *Aucun nombre $n = 4m+3$ n'est somme de deux carrés.*

■ **Preuve.** Le carré de tout nombre pair est $(2k)^2 = 4k^2 \equiv 0 \pmod 4$, tandis que les carrés des nombres impairs sont de la forme $(2k+1)^2 = 4(k^2+k)+1 \equiv 1 \pmod 4$. Ainsi, toute somme de deux carrés est congrue à $0, 1$ ou $2 \pmod 4$. □

Il est évident que les nombres premiers $p = 4m+3$ sont « mauvais ». On poursuit donc en espérant obtenir de « bonnes » propriétés pour les nombres premiers de la forme $p = 4m+1$. Voici le point clé pour parvenir au théorème principal.

Proposition. *Tout nombre premier de la forme $p = 4m+1$ est somme de deux carrés, c'est-à-dire qu'il peut s'écrire sous la forme $p = x^2+y^2$ avec $x, y \in \mathbb{N}$.*

Nous allons présenter ici deux preuves de ce résultat, aussi élégantes et surprenantes l'une que l'autre. La première preuve combine une application saisissante du « principe des tiroirs » (déjà utilisé avant le lemme 2 ; voir aussi le chapitre 25 pour en savoir davantage sur ce principe) et des congruences. L'idée est due au théoricien des nombres norvégien Axel Thue.

■ **Preuve.** On considère les couples d'entiers (x', y') qui vérifient la propriété $0 \leqslant x', y' \leqslant \sqrt{p}$, c'est-à-dire tels que $x', y' \in \{0, 1, \ldots, \lfloor \sqrt{p} \rfloor\}$. Il y a $(\lfloor \sqrt{p} \rfloor + 1)^2$ couples de ce type. En utilisant l'estimation $\lfloor x \rfloor + 1 > x$ si $x = \sqrt{p}$, on voit qu'il y a plus de p couples d'entiers de ce type. Ainsi, pour tout $s \in \mathbb{Z}$, il est impossible que toutes les valeurs $x' - sy'$ produites par les paires (x', y') soient distinctes modulo p. Donc pour chaque s, il existe deux couples distincts :

$$(x', y'),\ (x'', y'') \ \in \ \{0, 1, \ldots, \lfloor \sqrt{p} \rfloor\}^2$$

tels que :

$$x' - sy' \ \equiv \ x'' - sy'' \pmod p$$

On constate alors que $x' - x'' \equiv s(y' - y'') \pmod p$. Ainsi, en posant :

$$x := |x' - x''| \quad \text{et} \quad y := |y' - y''|$$

on obtient :

$$(x, y) \in \{0, 1, \ldots, \lfloor \sqrt{p} \rfloor\}^2 \quad \text{avec} \quad x \equiv \pm sy \pmod p$$

Pour $p = 13$, $\lfloor \sqrt{p} \rfloor = 3$; on est donc amené à considérer $x', y' \in \{0, 1, 2, 3\}$. Pour $s = 5$, la somme $x' - sy' \pmod{13}$ prend les valeurs suivantes :

x' \ y'	0	1	2	3
0	0	8	3	11
1	1	9	4	12
2	2	10	5	0
3	3	11	6	1

Nous savons aussi que x et y ne peuvent être nuls tous les deux, parce que les paires (x', y') et (x'', y'') sont distinctes.

Soit maintenant s une solution de $s^2 \equiv -1 \pmod{p}$ (cette solution existe d'après le lemme 1). Alors $x^2 \equiv s^2 y^2 \equiv -y^2 \pmod{p}$; nous avons donc mis en évidence un couple $(x, y) \in \mathbb{Z}^2$ tel que $0 < x^2 + y^2 < 2p$ et $x^2 + y^2 \equiv 0 \pmod{p}$. Mais p est le seul nombre qui se trouve entre 0 et $2p$ et qui est divisible par p. Ainsi, $x^2 + y^2 = p$ ce qui termine la démonstration. □

Notre deuxième démonstration de la proposition — il est clair que c'est aussi une preuve du Grand Livre — a été découverte par Roger Heath-Brown en 1971. Elle est parue en 1984. (Une version condensée en « une phrase » a été donnée par Don Zagier). Elle est tellement élémentaire que nous n'avons même pas besoin d'utiliser le Lemme 1.

L'argument de Heath-Brown est structuré autour de trois involutions linéaires : la première complètement évidente, une deuxième cachée et une troisième triviale qui constitue une sorte de bouquet final. La deuxième involution, inespérée, repose sur une structure cachée sur l'ensemble des solutions entières de l'équation $4xy + z^2 = p$.

■ **Preuve.** Étudions l'ensemble :

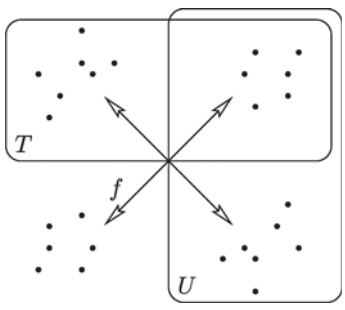
$$S \;:=\; \{(x, y, z) \in \mathbb{Z}^3 : 4xy + z^2 = p, \quad x > 0, \quad y > 0\}$$

Cet ensemble est fini. En effet, $x \geqslant 1$ et $y \geqslant 1$ impliquent que $y \leqslant \frac{p}{4}$ et $x \leqslant \frac{p}{4}$. Il n'y a donc qu'un nombre fini de valeurs possibles pour x et y. En outre, étant donnés x et y, il y a au plus deux valeurs possibles pour z.

1. La première involution linéaire est définie par :

$$f : S \longrightarrow S, \quad (x, y, z) \longmapsto (y, x, -z)$$

elle échange x et y, et change le signe de z. Elle envoie S sur lui-même ; c'est une *involution* : appliquée à deux reprises, elle donne l'identité. Remarquons que f n'a pas de point fixe, puisque $z = 0$ implique $p = 4xy$, ce qui est impossible. D'autre part, f envoie les éléments de



$$T \;:=\; \{(x, y, z) \in S : z > 0\}$$

sur les éléments de $S \backslash T$ qui satisfont $z < 0$. Enfin, f inverse les signes de $x - y$ et de z et elle envoie donc les éléments de

$$U \;:=\; \{(x, y, z) \in S : (x - y) + z > 0\}$$

sur les solutions de $S \backslash U$. Pour nous en assurer, nous devons nous convaincre qu'il n'y a pas de solutions telles que $(x - y) + z = 0$, ce qui est le cas puisque cela impliquerait : $p = 4xy + z^2 = 4xy + (x - y)^2 = (x + y)^2$.

Que peut-on conclure de l'étude de f ? L'observation principale est la suivante : puisque f envoie les ensembles T et U sur leurs complémentaires, elle échange aussi les éléments de $T \backslash U$ avec ceux de $U \backslash T$. C'est-à-dire

qu'il y a le même nombre de solutions dans U qui ne sont pas dans T que de solutions dans T qui ne sont pas dans U. Donc T *et* U *ont le même cardinal.*

2. La deuxième involution que nous étudions est une application définie sur l'ensemble U :

$$g : U \longrightarrow U, \quad (x, y, z) \longmapsto (x - y + z, y, 2y - z)$$

Vérifions d'abord que cette application est bien définie : si $(x, y, z) \in U$, alors $x - y + z > 0, y > 0$ et $4(x - y + z)y + (2y - z)^2 = 4xy + z^2$, donc $g(x, y, z) \in S$. Puisque $(x - y + z) - y + (2y - z) = x > 0$, on en conclut que $g(x, y, z) \in U$.

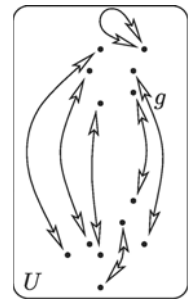
Par ailleurs, g est bien une involution puisque $g(x, y, z) = (x-y+z, y, 2y-z)$ est envoyé par g sur $((x - y + z) - y + (2y - z), y, 2y - (2y - z)) = (x, y, z)$.

Enfin g a exactement un point fixe :

$$(x, y, z) \;=\; g(x, y, z) \;=\; (x - y + z, y, 2y - z)$$

ne se produit que si $y = z$. Il s'ensuit que $p = 4xy + y^2 = (4x + y)y$, ce qui n'est le cas que si $y = 1 = z$ et $x = \frac{p-1}{4}$.

Comme g est une involution sur U comportant exactement un point fixe, *le cardinal de* U *est impair.*

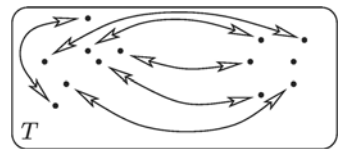


3. La troisième involution (triviale) que nous devons étudier est l'application définie sur T qui échange x et y :

$$h : T \longrightarrow T, \quad (x, y, z) \longmapsto (y, x, z)$$

Il est clair que cette application est bien définie et que c'est une involution. Combinons maintenant ce que nous savons des deux autres involutions : le cardinal de T est égal au cardinal de U, qui est impair. h étant une involution sur un ensemble fini de cardinal impair, elle *a un point fixe*. Il existe donc un point $(x, y, z) \in T$ tel que $x = y$, c'est-à-dire une solution de :

$$p = 4x^2 + z^2 = (2x)^2 + z^2 \qquad \square$$



Sur un ensemble fini de cardinal impair, une involution admet au moins un point fixe.

Notons que cette démonstration conduit à un résultat plus précis : le nombre de représentations de p de la forme $p = x^2 + (2y)^2$ est *impair* pour tous les nombres premiers de la forme $p = 4m + 1$ (la représentation est en réalité unique, voir [4]). Notons aussi que les deux preuves ne sont pas efficaces : essayez de trouver x et y pour un nombre premier à dix chiffres ! Des méthodes efficaces pour trouver de telles représentations en sommes de deux carrés sont décrites dans [1] et [8].

Le théorème suivant répond complètement à la question posée au début du chapitre.

Théorème. *Un entier naturel n peut être représenté comme une somme de deux carrés si et seulement si chaque facteur premier de la forme $p = 4m + 3$ apparaît avec un exposant pair dans la décomposition en facteurs premiers de n.*

■ **Preuve.** On dit qu'un nombre n est *représentable* s'il est la somme de deux carrés, c'est-à-dire s'il existe $x, y \in \mathbb{N}$ tels que $n = x^2 + y^2$. Le théorème est une conséquence des cinq résultats suivants.

(1) Les nombres $1 = 1^2 + 0^2$ et $2 = 1^2 + 1^2$ sont représentables ; tout nombre premier de la forme $p = 4m + 1$ est représentable.

(2) le produit de deux nombres représentables quelconques $n_1 = x_1^2 + y_1^2$ et $n_2 = x_2^2 + y_2^2$ est représentable puisque $n_1 n_2 = (x_1 x_2 + y_1 y_2)^2 + (x_1 y_2 - x_2 y_1)^2$.

(3) Si n est représentable, $n = x^2 + y^2$, nz^2 est aussi représentable puisque $nz^2 = (xz)^2 + (yz)^2$.

Les résultats (1), (2) et (3) permettent d'établir l'implication directe du théorème.

(4) Si $p = 4m + 3$ est un nombre premier qui divise un nombre représentable $n = x^2 + y^2$, alors p divise x et y, et donc p^2 divise n. En effet, si $x \not\equiv 0 \,(\mathrm{mod}\, p)$, on pourrait trouver $\overline{x}$ tel que $x\overline{x} \equiv 1 \,(\mathrm{mod}\, p)$ et, en multipliant l'équation $x^2 + y^2 \equiv 0$ par $\overline{x}^2$, on obtiendrait :

$$1 + y^2\overline{x}^2 = 1 + (\overline{x}y)^2 \equiv 0 \,(\mathrm{mod}\ )p$$

ce qui est impossible pour $p = 4m + 3$ d'après le lemme 1.

(5) Si n est représentable et si $p = 4m + 3$ divise n, alors p^2 divise n et n/p^2 est représentable. Cela se déduit de (4) et ainsi la démonstration est terminée. □

Pour terminer, voici deux remarques :

- Si a et b sont deux nombres naturels premiers entre eux, alors il existe une infinité de nombres premiers de la forme $am + b$ ($m \in \mathbb{N}$). C'est un théorème célèbre (et difficile) de Dirichlet. Plus précisément, on peut montrer que, pour de grandes valeurs de x, le nombre d'entiers premiers $p \leqslant x$ de la forme $p = am + b$ est approché de manière assez précise par $\frac{1}{\varphi(a)} \frac{x}{\ln x}$ où $\varphi(a)$ désigne le nombre d'entiers b, $1 \leqslant b < a$, qui sont premiers avec a (ce résultat constitue un raffinement substantiel du théorème sur les nombres premiers évoqué en page 10).
- Cela signifie que, pour a fixé, en faisant varier b les nombres premiers apparaissent à la même fréquence. Néanmoins, pour $a = 4$, une observation attentive montre qu'il existe « bien plus » de nombres premiers de la forme $4m + 3$ que de la forme $4m + 1$. Cet effet est connu sous le nom de « biais de Chebyshev » (voir Granville et Martin [2], Riesel [5] ainsi que Rubinstein et Sarnak [6]).

Bibliographie

[1] F. W. CLARKE, W. N. EVERITT, L. L. LITTLEJOHN & S. J. R. VORSTER : *H. J. S. Smith and the Fermat Two Squares Theorem,* Amer. Math. Monthly **106** (1999), 652-665.

[2] A. GRANVILLE & G. MARTIN : *Prime number races,* American Math. Monthly **113** (2006), 1-33.

[3] D. R. HEATH-BROWN : *Fermat's two squares theorem,* Invariant (1984), 2-5.

[4] I. NIVEN & H. S. ZUCKERMAN : *An Introduction to the Theory of Numbers,* Fifth edition, Wiley, New York 1972.

[5] H. RIESEL : *Prime Numbers and Computer Methods for Factorization,* Second edition, Progress in Mathematics **126**, Birkhäuser, Boston MA 1994.

[6] M. RUBINSTEIN & P. SARNAK : *Chebyshev's bias,* Experimental Mathematics **3** (1994), 173-197.

[7] A. THUE : *Et par antydninger til en taltheoretisk metode,* Kra. Vidensk. Selsk. Forh. **7** (1902), 57-75.

[8] S. WAGON : *Editor's corner : The Euclidean algorithm strikes again,* Amer. Math. Monthly **97** (1990), 125-129.

[9] D. ZAGIER : *A one-sentence proof that every prime* $p \equiv 1 \pmod 4$ *is a sum of two squares,* Amer. Math. Monthly **97** (1990), 144.

La loi de réciprocité quadratique — Chapitre 5

Quel est le théorème qui admet le plus grand nombre de démonstrations différentes ? Le théorème de Pythagore ou le théorème fondamental de l'algèbre sont certainement de bons candidats mais le champion est sans aucun doute la loi de réciprocité quadratique en théorie des nombres. Dans une monographie admirable, Franz Lemmermeyer établit à la date de l'an 2000 une liste ne comportant pas moins de 196 démonstrations différentes de ce résultat. Bien sûr, certaines d'entre elles ne sont que de légères variantes mais la palette des idées mises en œuvre et la liste des contributeurs sont impressionnantes.

Euler, Legendre et Gauss ont beaucoup travaillé sur les résidus quadratiques et ont étudié en particulier les relations possibles entre q résidu quadratique modulo p et p résidu quadratique modulo q, lorsque p et q sont des nombres premiers impairs. Euler et Legendre découvrirent ainsi le théorème remarquable qui suit mais ne le démontrèrent que pour des cas particuliers. Le 8 avril 1796, Gauss rédigea la première preuve complète de ce résultat dans son journal. Un peu plus tard, Ferdinand Gotthold Eisenstein proposa cinq nouvelles preuves. La liste de ceux qui ont établi des démonstrations originales de la loi de réciprocité quadratique ressemble à un *Who's Who* des mathématiques !

Carl Friedrich Gauss

Face à tant de démonstrations différentes, la question de savoir laquelle doit figurer dans le Grand Livre est extrêmement délicate. Est-ce la plus courte ? Ou la plus inattendue ? Faut-il privilégier la démonstration qui se prête le mieux à des généralisations pour aboutir à d'autres lois de réciprocité plus sophistiquées ? Nous avons choisi deux démonstrations (qui reposent sur les troisième et sixième preuves de Gauss) ; la première peut être considérée comme la plus simple et la plus satisfaisante tandis que l'autre constitue le point de départ d'une série de résultats fondamentaux dans des structures plus générales.

Comme dans le précédent chapitre, nous travaillons « modulo p », où p est un nombre premier impair. $\mathbb{Z}_p$ désigne l'anneau des entiers modulo p — aussi appelés *restes* dans la suite — qui est dans ce cas un corps. On note généralement (mais pas toujours) $0, 1, \ldots, p-1$ les éléments de ce corps. On considère $a \not\equiv 0 \,(\mathrm{mod}\, p)$, c'est-à-dire $p \nmid a$. On dit que a est un *résidu quadratique* modulo p s'il existe b tel que $a \equiv b^2 \,(\mathrm{mod}\, p)$. Dans le cas contraire, on dit que a est un non-résidu quadratique. Les résidus quadratiques modulo p sont donc $1^2, 2^2, \ldots, (\frac{p-1}{2})^2$ si bien qu'il y a $\frac{p-1}{2}$ résidus quadratiques et $\frac{p-1}{2}$ non-résidus quadratiques modulo p. En effet, si $i^2 \equiv j^2 (\mathrm{mod}\, p)$ avec $1 \leqslant i, j \leqslant \frac{p-1}{2}$, alors p divise $i^2 - j^2 = (i-j)(i+j)$.

Pour $p = 13$, les résidus quadratiques sont $1^2 \equiv 1$, $2^2 \equiv 4$, $3^2 \equiv 9$, $4^2 \equiv 3$, $5^2 \equiv 12$ et $6^2 \equiv 10$; les non-résidus sont $2, 5, 6, 7, 8, 11$.

Comme $2 \leqslant i + j \leqslant p - 1$, p divise $i - j$, c'est-à-dire $i \equiv j \pmod{p}$.

À ce stade, il est commode d'introduire le *symbole de Legendre*. Pour $a \not\equiv 0 \pmod{p}$, on pose :

$$\left(\frac{a}{p}\right) := \begin{cases} 1 & \text{si } a \text{ est un résidu quadratique modulo } p, \\ -1 & \text{si } a \text{ est un non-résidu quadratique modulo } p. \end{cases}$$

L'histoire commence avec le « petit théorème de Fermat ». Pour $a \not\equiv 0 \pmod{p}$,

$$a^{p-1} \equiv 1 \pmod{p}. \tag{1}$$

En fait, comme $\mathbb{Z}_p^* = \mathbb{Z}_p \setminus \{0\}$ est un groupe pour la multiplication, on observe que $\{1a, 2a, 3a, \ldots, (p-1)a\} = \mathbb{Z}_p^*$ si bien que :

$$(1a)(2a)\cdots((p-1)a) \equiv 1 \cdot 2 \cdots (p-1) \pmod{p},$$

et en divisant par $(p-1)!$ on obtient $a^{p-1} \equiv 1 \pmod{p}$.
En d'autres termes, le *polynôme* $x^{p-1} - 1 \in \mathbb{Z}_p[x]$ a pour racines tous les éléments non nuls de $\mathbb{Z}_p$. On constate ensuite que :

$$x^{p-1} - 1 = (x^{\frac{p-1}{2}} - 1)(x^{\frac{p-1}{2}} + 1).$$

Supposons $a \equiv b^2 \pmod{p}$ est un résidu quadratique. Alors, selon le petit théorème de Fermat, $a^{\frac{p-1}{2}} \equiv b^{p-1} \equiv 1 \pmod{p}$. Ainsi, les résidus quadratiques sont exactement les racines du premier facteur, $x^{\frac{p-1}{2}} - 1$, et les $\frac{p-1}{2}$ non-résidus sont donc les racines du second facteur $x^{\frac{p-1}{2}} + 1$. En mettant ce résultat en correspondance avec la définition du symbole de Legendre, on obtient le résultat important qui suit.

Par exemple, pour $p = 17$ et $a = 3$, on obtient $3^8 = (3^4)^2 = 81^2 \equiv (-4)^2 \equiv -1 \pmod{17}$, alors que pour $a = 2$ on trouve $2^8 = (2^4)^2 \equiv (-1)^2 \equiv 1 \pmod{17}$. Ainsi, 2 est un résidu quadratique tandis que 3 est un non-résidu.

Critère d'Euler. *Pour $a \not\equiv 0 \pmod{p}$,*

$$\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \pmod{p}.$$

Le critère d'Euler fournit immédiatement la *règle du produit* suivante :

$$\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right), \tag{2}$$

puisque le résultat se lit directement dans le membre de droite du critère d'Euler. Cette règle est très utile lorsqu'on veut calculer pratiquement des valeurs du symbole de Legendre. Puisque tout entier est le produit de ± 1 et de nombre premiers, il suffit de connaître $\left(\frac{-1}{p}\right)$, $\left(\frac{2}{p}\right)$ et $\left(\frac{q}{p}\right)$ pour les entiers premiers impairs q.
Selon le critère d'Euler, $\left(\frac{-1}{p}\right) = 1$ si $p \equiv 1 \pmod{4}$ et $\left(\frac{-1}{p}\right) = -1$ si $p \equiv 3 \pmod{4}$, résultats déjà observés dans le chapitre précédent. La valeur de $\left(\frac{2}{p}\right)$ résulte du lemme de Gauss que nous allons établir un peu plus loin : $\left(\frac{2}{p}\right) = 1$ si $p \equiv \pm 1 \pmod{8}$ tandis que $\left(\frac{2}{p}\right) = -1$ si $p \equiv \pm 3 \pmod{8}$.

Gauss procéda à de nombreux calculs sur les résidus quadratiques et, en particulier, essaya de voir quelle relation il pouvait y avoir entre le fait que q soit un résidu quadratique modulo p et que p soit un résidu quadratique modulo q, lorsque p et q sont des nombres premiers impairs. Il proposa finalement la conjecture suivante qu'il parvint à démontrer.

Loi de réciprocité quadratique. *Soient p et q deux nombres premiers impairs distincts. Alors :*

$$\left(\frac{q}{p}\right)\left(\frac{p}{q}\right) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}.$$

Si $p \equiv 1 \pmod 4$ ou $q \equiv 1 \pmod 4$, alors $\frac{p-1}{2}$ (respectivement $\frac{q-1}{2}$) est pair et donc $(-1)^{\frac{p-1}{2}\frac{q-1}{2}} = 1$. Ainsi, $\left(\frac{q}{p}\right) = \left(\frac{p}{q}\right)$. Lorsque $p \equiv q \equiv 3 \pmod 4$, on observe que $\left(\frac{p}{q}\right) = -\left(\frac{q}{p}\right)$. Pour des nombres premiers impairs, on obtient donc $\left(\frac{p}{q}\right) = \left(\frac{q}{p}\right)$ sauf si p et q sont *simultanément* congrus à $3 \pmod 4$.

Exemple : $\left(\frac{3}{17}\right) = \left(\frac{17}{3}\right) = \left(\frac{2}{3}\right) = -1$, ainsi 3 est un non-résidu modulo 17.

Première démonstration. La clé de notre première preuve (qui est au centre de la troisième preuve proposée par Gauss) consiste en une formule de dénombrement qui fut rapidement connue sous le nom de *Lemme de Gauss*.

Lemme de Gauss. *On suppose $a \not\equiv 0 \pmod p$. On considère les nombres $1a, 2a, \ldots, \frac{p-1}{2}a$ et on réduit chacun d'eux modulo p à son représentant le plus petit possible en valeur absolue, c'est-à-dire au nombre $ka \equiv r_k \pmod p$ avec $-\frac{p-1}{2} \leq r_k \leqslant \frac{p-1}{2}$ pour tout k. Alors :*

$$\left(\frac{a}{p}\right) = (-1)^s, \quad \text{où } s = Card\{k : r_k < 0\}.$$

■ **Preuve.** Désignons par $u_1, \ldots, u_s$ les représentants inférieurs à 0 et par $v_1, \ldots, v_{\frac{p-1}{2}-s}$ ceux supérieurs à 0. On remarque que les nombres $-u_1, \ldots, -u_s$ sont tous compris entre 1 et $\frac{p-1}{2}$ et qu'ils sont tous distincts des v_j (voir note en marge). Il s'ensuit que $\{-u_1, \ldots, -u_s, v_1, \ldots, v_{\frac{p-1}{2}-s}\} = \{1, 2, \ldots, \frac{p-1}{2}\}$. On en déduit :

$$\prod_i (-u_i) \prod_j v_j = \tfrac{p-1}{2}!,$$

ce qui implique :

$$(-1)^s \prod_i u_i \prod_j v_j \equiv \tfrac{p-1}{2}! \pmod p.$$

Si $-u_i = v_j$, alors $u_i + v_j \equiv 0 \pmod p$. Par ailleurs, u_i et v_j sont respectivement de la forme $u_i \equiv ka$ et $v_j \equiv \ell a \pmod p$, ce qui implique $p \mid (k+\ell)a$. Comme p et a sont premiers entre eux, p divise $k + \ell$, ce qui est impossible car $k + \ell \leqslant p - 1$.

Souvenons-nous à présent de la manière dont ont été définis les u_i et les v_j. Ce sont les représentants de $1a, \cdots, \frac{p-1}{2}a$. Ainsi :

$$\tfrac{p-1}{2}! \equiv (-1)^s \prod_i u_i \prod_j v_j \equiv (-1)^s \tfrac{p-1}{2}! a^{\frac{p-1}{2}} \pmod{p}.$$

En simplifiant par $\frac{p-1}{2}!$ et en utilisant le critère d'Euler on trouve :

$$(\frac{a}{p}) \equiv a^{\frac{p-1}{2}} \equiv (-1)^s \pmod{p},$$

et donc $(\frac{a}{p}) = (-1)^s$ car p est impair. □

À l'aide de ce résultat, on peut facilement calculer $(\frac{2}{p})$. Comme $1.2, 2.2, \ldots, \frac{p-1}{2}.2$ sont tous compris entre 1 et $p-1$, on obtient :

$s = \text{Card}\{i : \frac{p-1}{2} < 2i \leq p-1\} = \frac{p-1}{2} - \text{Card}\{i : 2i \leqslant \frac{p-1}{2}\} = \lceil \frac{p-1}{4} \rceil$.

On vérifie que s est pair exactement lorsque $p = 8k \pm 1$.

Le lemme de Gauss est le point central de nombreuses démonstrations de la loi de réciprocité quadratique. La plus élégante pourrait bien être celle suggérée par Ferdinand Gotthold Eisenstein, qui avait appris la théorie des nombres en lisant le célèbre *Disquisitiones Arithmeticae* de Gauss et qui proposa d'importantes contributions à des « théorèmes de réciprocité plus profonds » avant sa disparition prématurée à l'âge de 29 ans. Sa preuve se limite à compter des points dans un réseau !

Soient p et q des entiers premiers impairs et considérons $(\frac{q}{p})$. Supposons que kq est un multiple de q qui se réduit au représentant négatif $r_k < 0$ selon les notations adoptées précédemment pour le lemme de Gauss. Cela signifie qu'il existe un entier unique j tel que $-\frac{p}{2} < kq - jp < 0$. Remarquons que $0 < j < \frac{q}{2}$ car $0 < k < \frac{p}{2}$. En d'autres termes, $(\frac{q}{p}) = (-1)^s$, où s est le nombre de points du réseau constitué des couples (x, y) d'entiers tels que :

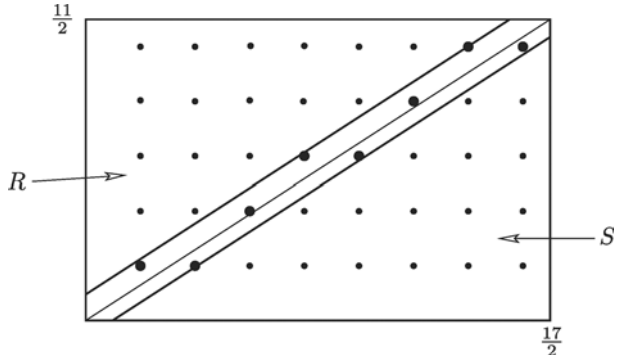
$$0 < py - qx < \frac{p}{2},\ 0 < x < \frac{p}{2},\ 0 < y < \frac{q}{2}. \tag{3}$$

De même, $(\frac{p}{q}) = (-1)^t$ où t est le nombre de points (x, y) du réseau défini par :

$$0 < qx - py < \frac{q}{2},\ 0 < x < \frac{p}{2},\ 0 < y < \frac{q}{2}. \tag{4}$$

Considérons à présent le rectangle de longueur $\frac{p}{2}$ et de largeur $\frac{q}{2}$ et traçons les deux lignes, toutes deux parallèles à la diagonale d'équation $py = qx$, d'équations $y = \frac{q}{p}x + \frac{1}{2}$ ou $py - qx = \frac{p}{2}$ d'une part, et $y = \frac{q}{p}(x - \frac{1}{2})$ ou $qx - py = \frac{q}{2}$ d'autre part.

La figure suivante montre la situation pour les valeurs $p = 17$ et $q = 11$.



$$p = 17, \quad q = 11,$$
$$s = 5, \quad t = 3,$$
$$\left(\frac{q}{p}\right) = (-1)^5 = -1,$$
$$\left(\frac{p}{q}\right) = (-1)^3 = -1.$$

On achève rapidement la démonstration en tenant compte des trois remarques suivantes :

1. Il n'y a pas de points du réseau sur la diagonale ni sur les deux parallèles évoquées. Il en est ainsi car $py = qx$ impliquerait $p \mid x$, ce qui est impossible. Pour ce qui concerne les parallèles, on observe que $py - qx$ est entier alors que $\frac{p}{2}$ et $\frac{q}{2}$ ne le sont pas.

2. Les points du réseau qui satisfont à l'équation (3) sont précisément ceux qui sont situés dans la bande supérieure (la bande délimitée par la diagonale et la parallèle évoquée précédemment située juste au-dessus de cette diagonale) $0 < py - qx < \frac{p}{2}$ et ceux qui satisfont à (4) se trouvent dans la bande inférieure $0 < qx - py < \frac{q}{2}$. Ainsi, le nombre de points du réseau que l'on trouve dans les deux bandes est égal à $s + t$.

3. Les deux parties extérieures R, définie par $py - qx > \frac{p}{2}$, et S, définie par $qx - py > \frac{q}{2}$ (les deux triangles respectivement situés au-dessus de la sur-diagonale (R) et en dessous de la sous-diagonale (S)), contiennent le *même nombre* de points. Pour s'en convaincre, il suffit de considérer l'application $\varphi : R \to S$ qui envoie le couple (x, y) sur $(\frac{p+1}{2} - x, \frac{q+1}{2} - y)$ et de vérifier que φ est involutive.

Comme le nombre total de points du réseau situés dans le rectangle considéré est $\frac{p-1}{2} \cdot \frac{q-1}{2}$, on en déduit que $s+t$ et $\frac{p-1}{2} \cdot \frac{q-1}{2}$ sont de même parité. Il en résulte que :

$$\left(\frac{q}{p}\right)\left(\frac{p}{q}\right) = (-1)^{s+t} = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}. \qquad \square$$

Seconde démonstration. Cette fois la preuve ne repose pas sur le lemme de Gauss. Au contraire, elle utilise ce que l'on appelle des *sommes de Gauss* dans les corps finis. Gauss les introduisit dans le cadre de l'étude de l'équation $x^p - 1 = 0$ et des propriétés arithmétiques de l'extension cyclotomique $\mathbb{Q}(\zeta)$ de $\mathbb{Q}$ (appelée *corps cyclotomique*), où ζ désigne une racine n-ième

de l'unité. Ces sommes sont le point de départ de raisonnements conduisant à l'obtention de lois de réciprocité plus générales.

Rappelons d'abord quelques résultats importants concernant les corps finis.

A. Soient p et q deux entiers premiers impairs distincts et considérons le corps fini F à q^{p-1} éléments. Son corps premier est $\mathbb{Z}_q$, donc $qa = 0$ pour tout $a \in F$. Cela entraîne que $(a+b)^q = a^q + b^q$ puisque tout coefficient binomial $\binom{q}{i}$ est un multiple de q pour $0 < i < q$ et est donc nul dans F. Remarquons que le critère d'Euler s'écrit sous la forme de *l'égalité* $(\frac{p}{q}) = p^{\frac{q-1}{2}}$ dans le corps premier $\mathbb{Z}_q$.

B. Le groupe multiplicatif $F^* = F \setminus \{0\}$ est cyclique de cardinal $q^{p-1}-1$ (voir l'encadré de la page suivante). Selon le petit théorème de Fermat, p divise $q^{p-1}-1$ et donc il existe un élément $\zeta \in F$ d'ordre p, c'est-à-dire tel que $\zeta^p = 1$. Cet élément ζ engendre le sous-groupe $\{\zeta, \zeta^2, \ldots, \zeta^p = 1\}$ de F^*. Remarquons que n'importe quel ζ^i $(i \not\equiv p)$ est lui aussi un générateur de ce sous-groupe. On obtient ainsi la factorisation :

$$x^p - 1 = (x-\zeta)(x-\zeta^2)\cdots(x-\zeta^p)\,.$$

Muni de ces résultats, on peut se mettre au travail. Considérons la *somme de Gauss* suivante :

$$G := \sum_{i=1}^{p-1} (\frac{i}{p})\,\zeta^i \in F,$$

où $(\frac{i}{p})$ désigne le symbole de Legendre. Pour les besoins de la démonstration qui suit, on va calculer deux expressions différentes de G^q.

Première expression. On constate que :

$$G^q = \sum_{i=1}^{p-1} (\frac{i}{p})^q \zeta^{iq} = \sum_{i=1}^{p-1} (\frac{i}{p})\zeta^{iq} = (\frac{q}{p}) \sum_{i=1}^{p-1} (\frac{iq}{p})\zeta^{iq} = (\frac{q}{p})G\,. \quad (5)$$

Exemple : on prend $p = 3$ et $q = 5$; on trouve $G = \zeta - \zeta^2$ et $G^5 = \zeta^5 - \zeta^{10} = \zeta^2 - \zeta = -(\zeta - \zeta^2) = -G$, ce qui conduit à $(\frac{5}{3}) = (\frac{2}{3}) = -1$.

La première égalité vient de $(a+b)^q = a^q + b^q$. La deuxième égalité utilise $(\frac{i}{p})^q = (\frac{i}{p})$ puisque q est impair. La troisième égalité résulte de l'application de la règle des produits (2), qui conduit notamment à $(\frac{i}{p}) = (\frac{q}{p})(\frac{iq}{p})$. Enfin, la dernière égalité s'obtient en constatant que iq, comme i, parcourt tout l'ensemble des entiers non nuls modulo p.

Seconde expression. Admettons que l'on ait établi l'égalité :

$$G^2 = (-1)^{\frac{p-1}{2}} p\,. \quad (6)$$

On a alors presque terminé. En effet :

$$G^q = G(G^2)^{\frac{q-1}{2}} = G(-1)^{\frac{p-1}{2}\frac{q-1}{2}} p^{\frac{q-1}{2}} = G(\frac{p}{q})(-1)^{\frac{p-1}{2}\frac{q-1}{2}}\,. \quad (7)$$

En comparant les expressions obtenues en (5) et en (7) puis en simplifiant par G, qui est non nul d'après (6), on trouve $(\frac{q}{p}) = (\frac{p}{q})(-1)^{\frac{p-1}{2}\frac{q-1}{2}}$ et donc finalement :

$$(\frac{q}{p})(\frac{p}{q}) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}.$$

Le groupe multiplicatif d'un corps fini est cyclique

Soit F^* le groupe multiplicatif du corps F, avec $|F^*| = n$. On désigne par $\mathrm{ord}(a)$ l'ordre de l'élément a, c'est-à-dire le plus petit entier k tel que $a^k = 1$. On veut trouver un élément $a \in F^*$ avec $\mathrm{ord}(a) = n$. Si pour un élément b de F^* on observe $\mathrm{ord}(b) = d$, alors, selon le théorème de Lagrange, d divise n (voir note de marge en page 4). En classant les éléments de F^* en fonction de leur ordre, on trouve :

$$n = \sum_{d|n} \psi(d), \text{ où } \psi(d) = \mathrm{Card}\{b \in F^* : \mathrm{ord}(b) = d\}. \quad (8)$$

Si $\mathrm{ord}(b) = d$, alors tout élément b^i ($i \in \{1, \ldots, d\}$) vérifie $(b^i)^d = 1$ et constitue donc une racine du polynôme $x^d - 1$. Cependant, comme F est un corps, $x^d - 1$ possède au plus d racines, si bien que les éléments $b, b^2, \ldots, b^d = 1$ sont exactement ces racines. En particulier, tout élément d'ordre d est de la forme b^i.
Par ailleurs, on vérifie facilement que $\mathrm{ord}(b^i) = \frac{d}{(i,d)}$, où (i, d) désigne le plus grand commun diviseur de i et d. Ainsi, $\mathrm{ord}(b^i) = d$ si et seulement si $(i, d) = 1$, c'est-à-dire si i et d sont premiers entre eux. En définissant la *fonction d'Euler* par :

$$\varphi(d) = \mathrm{Card}\{i : 1 \leqslant i \leqslant d, (i, d) = 1\},$$

on obtient $\psi(d) = \varphi(d)$ dès que $\psi(d) > 0$. Tenant compte de (8), on trouve :

$$n = \sum_{d|n} \psi(d) \leqslant \sum_{d|n} \varphi(d).$$

Toutefois, nous allons voir que :

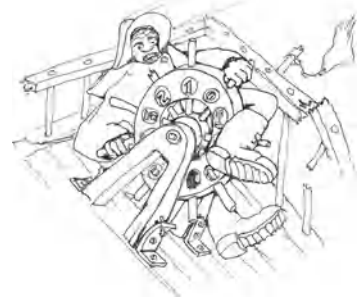
$$\sum_{d|n} \varphi(d) = n, \quad (9)$$

ce qui implique que $\psi(d) = \varphi(d)$ pour tout d. En particulier, on constate que $\psi(n) = \varphi(n) \geqslant 1$ et donc il existe au moins un élément d'ordre n.

La preuve de (9) qui suit se doit de figurer, elle aussi, dans le Grand Livre. Considérons les n fractions :

$$\frac{1}{n}, \frac{2}{n}, \ldots, \frac{k}{n}, \ldots, \frac{n}{n}.$$

En réduisant chacune d'elles à son représentant irréductible sous la forme $\frac{k}{n} = \frac{i}{d}$ avec $1 \leqslant i \leqslant d$, $(i, d) = 1$ et $d \mid n$, on vérifie que le dénominateur d apparaît exactement $\varphi(d)$ fois.



« Même dans le chaos le plus total on peut compter sur le groupe cyclique ».

– Que se passe-t-il ?

– Je transporte 196 preuves de la loi de réciprocité quadratique !

Il reste à montrer l'égalité (6). Dans ce but, nous faisons les deux observations simples suivantes :

- $\sum_{i=1}^{p} \zeta^i = 0$ et donc $\sum_{i=1}^{p-1} \zeta^i = -1$. En effet $-\sum_{i=1}^{p} \zeta^i$ est le coefficient de x^{p-1} dans $x^p - 1 = \prod_{i=1}^{p}(x - \zeta^i)$ et vaut donc 0.
- $\sum_{k=1}^{p-1}(\frac{k}{p}) = 0$ (car il y a autant de résidus que de non-résidus quadratiques) et donc $\sum_{k=1}^{p-2}(\frac{k}{p}) = -(\frac{-1}{p})$.

On constate que :

$$G^2 = \Big(\sum_{i=1}^{p-1}(\frac{i}{p})\,\zeta^i\Big)\Big(\sum_{j=1}^{p-1}(\frac{j}{p})\,\zeta^j\Big) = \sum_{i,j}(\frac{ij}{p})\,\zeta^{i+j}.$$

En écrivant $j \equiv ik \pmod p$, on trouve :

$$G^2 = \sum_{i,k}(\frac{k}{p})\,\zeta^{i(1+k)} = \sum_{k=1}^{p-1}(\frac{k}{p})\sum_{i=1}^{p-1}\zeta^{(1+k)i}.$$

Pour $k = p-1 \equiv -1 \pmod p$, cela donne $(\frac{-1}{p})(p-1)$, car $\zeta^{1+k} = 1$. En écrivant le terme obtenu pour $k = p-1$ séparément, on obtient :

Critère d'Euler : $(\frac{-1}{p}) = (-1)^{\frac{p-1}{2}}$.

$$G^2 = (\frac{-1}{p})(p-1) + \sum_{k=1}^{p-2}(\frac{k}{p})\sum_{i=1}^{p-1}\zeta^{(1+k)i}.$$

Comme ζ^{1+k} est un générateur du groupe pour tout $k \neq p-1$, la somme intérieure est égale à $\sum_{i=1}^{p-1}\zeta^i$, pour tout $k \neq p-1$, et donc à -1 d'après la première observation. La somme extérieure devient $-\sum_{k=1}^{p-2}(\frac{k}{p}) = (\frac{-1}{p})$ en tenant compte de la deuxième observation. Il s'ensuit que $G^2 = (\frac{-1}{p})p$ et donc, à l'aide du critère d'Euler, que $G^2 = (-1)^{\frac{p-1}{2}}p$, ce qui termine la démonstration. □

Pour $p = 3$, $q = 5$, $G^2 = (\zeta - \zeta^2)^2 = \zeta^2 - 2\zeta^3 + \zeta^4 = \zeta^2 - 2 + \zeta = -3 = (-1)^{\frac{3-1}{2}}3$, puisque $1 + \zeta + \zeta^2 = 0$.

Bibliographie

[1] A. BAKER : *A Concise Introduction to the Theory of Numbers,* Cambridge University Press, Cambridge 1984.

[2] F. G. EISENSTEIN : *Geometrischer Beweis des Fundamentaltheorems für die quadratischen Reste,* J. Reine Angewandte Mathematik **28** (1844), 186-191.

[3] C. F. GAUSS : *Theorema arithmetici demonstratio nova,* Comment. Soc. regiae sci. Göttingen **XVI** (1808), 69 ; Werke II, 1-8 (inclut la troisième démonstration).

[4] C. F. GAUSS : *Theorematis fundamentalis in doctrina de residuis quadraticis demonstrationes et amplicationes novae (1818),* Werke II, 47-64 (inclut la sixième démonstration).

[5] F. LEMMERMEYER : *Reciprocity Laws,* Springer-Verlag, Berlin 2000.

Tout corps fini est commutatif

Chapitre 6

La structure d'anneau joue un rôle important en algèbre moderne. Si un anneau R a un élément unité 1 pour la multiplication et si tout élément non nul admet un inverse pour la multiplication, on dit que R est un *corps*. L'exemple le plus connu de corps non commutatif est le corps des quaternions découvert par Hamilton. Toutefois, comme l'affirme le titre du chapitre, un tel corps est nécessairement infini. Si R est fini, les axiomes forcent la multiplication à être commutative.

Ce résultat désormais classique a stimulé l'imagination de nombreux mathématiciens. Comme l'écrit Herstein : « Il est très inattendu de mettre en relation deux objets apparemment sans rapport, le nombre d'éléments d'un certain système algébrique et la loi de multiplication de ce système. »

Ernst Witt

Théorème. *Tout corps fini est commutatif.*

Ce beau théorème, habituellement attribué à MacLagan Wedderburn, a été démontré par de nombreux mathématiciens, en utilisant une grande variété d'idées. Wedderburn lui-même en a donné trois preuves différentes en 1905 ; une autre preuve a été proposée par Leonard E. Dickson la même année. D'autres encore ont été données plus tard par Emil Artin, Hans Zassenhaus, Nicolas Bourbaki et bien d'autres. L'une d'elles s'impose par sa simplicité et son élégance. Elle a été trouvée par Ernst Witt en 1931 et combine deux idées élémentaires qui mènent à une conclusion brillante.

■ **Preuve.** Notre premier ingrédient est issu de l'algèbre linéaire. Étant donné un élément $s \in R$ arbitraire, soit C_s l'ensemble $\{x \in R : xs = sx\}$ des éléments permutables avec s ; C_s s'appelle le *centralisateur* de s. Il est clair que C_s contient 0 et 1 ; c'est un sous-corps de R. Le *centre* Z de R est l'ensemble des éléments qui commutent avec tous les éléments de R donc $Z = \bigcap_{s \in R} C_s$. En particulier, tous les éléments de Z sont permutables, 0 et 1 appartiennent à Z. Z est donc un *corps fini*. Posons $|Z| = q$.

On peut considérer R et C_s comme des espaces vectoriels sur le corps Z et en déduire que $|R| = q^n$, où n est la dimension de l'espace vectoriel R sur Z. De même, $|C_s| = q^{n_s}$ pour des entiers $n_s \geqslant 1$ convenables.

Supposons maintenant que R ne soit pas commutatif. Cela signifie que pour *un certain* $s \in R$ le centralisateur C_s n'est pas R tout entier, ou, ce qui revient au même, que $n_s < n$.

Sur l'ensemble $R^* := R\backslash\{0\}$, considérons la relation :

$$r' \sim r \quad :\Longleftrightarrow \quad r' = x^{-1}rx \ \text{ pour un certain } \ x \in R^*.$$

Il est facile de vérifier que $\sim$ est une relation d'équivalence. Soit :

$$A_s \ := \ \{x^{-1}sx : x \in R^*\}$$

la classe d'équivalence contenant s. Remarquons que $|A_s| = 1$, précisément lorsque s appartient au centre Z. Par hypothèse, il y a donc des classes A_s telles que $|A_s| \geqslant 2$. Si $s \in R^*$, considérons maintenant l'application, de R^* sur A_s, $f_s : x \longmapsto x^{-1}sx$. On a :

$$\begin{aligned} x^{-1}sx = y^{-1}sy \quad &\Longleftrightarrow \quad (yx^{-1})s = s(yx^{-1}) \\ &\Longleftrightarrow \quad yx^{-1} \in C_s^* \\ &\Longleftrightarrow \quad y \in C_s^*x \end{aligned}$$

où $C_s^*x = \{zx : z \in C_s^*\}$ a pour cardinal $|C_s^*|$. Par conséquent, tout élément $x^{-1}sx$ est l'image d'exactement $|C_s^*| = q^{n_s} - 1$ éléments de R^* par l'application f_s ; on en déduit que $|R^*| = |A_s|\,|C_s^*|$. En particulier, on remarque que :

$$\frac{|R^*|}{|C_s^*|} \ = \ \frac{q^n - 1}{q^{n_s} - 1} \ = \ |A_s| \quad \text{est un } \textit{entier} \text{ pour tout } s.$$

On sait que les classes d'équivalence constituent une partition de R^*. Regroupons maintenant tous les éléments centraux Z^* et désignons par A_1, …, A_t les classes d'équivalence qui contiennent plus d'un élément. Par hypothèse, nous savons que $t \geqslant 1$. Puisque $|R^*| = |Z^*| + \sum_{i=1}^{t} |A_i|$, nous obtenons le résultat suivant, connu sous le nom de *formule des classes* :

$$q^n - 1 \ = \ q - 1 + \sum_{i=1}^{t} \frac{q^n - 1}{q^{n_i} - 1} \tag{1}$$

où $1 < \frac{q^n-1}{q^{n_i}-1} \in \mathbb{N}$ pour tout i.
L'égalité (1) nous fait passer de l'algèbre abstraite à l'arithmétique. Nous allons maintenant montrer que si $q^{n_i} - 1 \,|\, q^n - 1$ alors $n_i \,|\, n$. En effet, écrivons $n = an_i + r$ avec $0 \leqslant r < n_i$. Alors $q^{n_i} - 1 \,|\, q^{an_i+r} - 1$ implique :

$$q^{n_i} - 1 \,|\, (q^{an_i+r} - 1) - (q^{n_i} - 1) = q^{n_i}(q^{(a-1)n_i+r} - 1)$$

et donc $q^{n_i} - 1 \,|\, q^{(a-1)n_i+r} - 1$, puisque q^{n_i} et $q^{n_i} - 1$ sont premiers entre eux. En poursuivant dans cette voie, nous trouvons que $q^{n_i} - 1 \,|\, q^r - 1$ si $0 \leqslant r < n_i$, ce qui n'est possible que si $r = 0$, donc $n_i \,|\, n$. Finalement, on constate que :

$$n_i \,|\, n \quad \text{pour tout } \ i \tag{2}$$

Le deuxième ingrédient intervient maintenant : l'ensemble des nombres complexes $\mathbb{C}$. Considérons le polynôme $x^n - 1$. Ses racines dans $\mathbb{C}$ sont

appelées les *racines n-ièmes de l'unité*. Puisque $\lambda^n = 1$, toutes ces racines λ vérifient $|\lambda| = 1$ et appartiennent par conséquent au cercle unité du plan complexe. En fait, ce sont précisément les nombres $\lambda_k = e^{\frac{2k\pi i}{n}} = \cos(2k\pi/n) + i\sin(2k\pi/n)$, $0 \leqslant k \leqslant n-1$ (voir encadré). Certaines racines λ vérifient $\lambda^d = 1$ avec $d < n$; par exemple, la racine $\lambda = -1$ vérifie $\lambda^2 = 1$. Pour une racine λ, soit d le plus petit exposant positif tel que $\lambda^d = 1$. En d'autres termes, d est l'ordre de λ dans le groupe des racines de l'unité. Alors $d \mid n$, d'après le théorème de Lagrange (« l'ordre de chaque élément d'un groupe divise l'ordre du groupe » ; voir encadré du chapitre 1). Notons aussi qu'il y a des racines d'ordre n, comme $\lambda_1 = e^{\frac{2\pi i}{n}}$.

Racines de l'unité

Tout nombre complexe (non nul) $z = x + iy$ peut s'écrire sous forme *polaire*

$$z = re^{i\varphi} = r(\cos\varphi + i\sin\varphi)$$

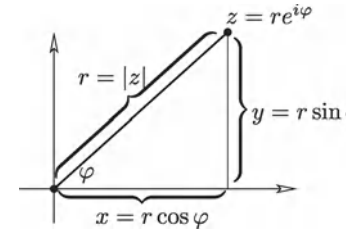
où $r = |z| = \sqrt{x^2+y^2}$ est la distance de z à l'origine et φ est l'angle mesuré à partir de l'axe x. Les racines n-ièmes de l'unité sont par conséquent de la forme :

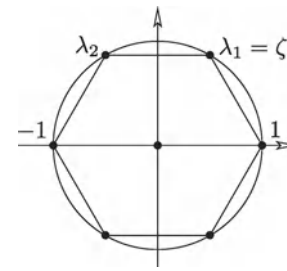
$$\lambda_k = e^{\frac{2k\pi i}{n}} = \cos(2k\pi/n) + i\sin(2k\pi/n) \qquad 0 \leqslant k \leqslant n-1$$

puisque pour tout k :

$$\lambda_k^n = e^{2k\pi i} = \cos(2k\pi) + i\sin(2k\pi) = 1$$

On obtient géométriquement ces racines en inscrivant un polygone régulier à n côtés sur le cercle unité. Notons que $\lambda_k = \zeta^k$ pour tout k, où $\zeta = e^{\frac{2\pi i}{n}}$. Ainsi, les racines n-ièmes de l'unité forment un groupe cyclique $\{\zeta, \zeta^2, \ldots, \zeta^{n-1}, \zeta^n = 1\}$ d'ordre n.





Les racines de l'unité lorsque $n = 6$.

Regroupons maintenant toutes les racines d'ordre d et posons :

$$\phi_d(x) := \prod_{\lambda \text{ d'ordre } d} (x - \lambda)$$

Remarquons que la définition de $\phi_d(x)$ est indépendante de n. Puisque chaque racine a un certain ordre d, on peut affirmer que :

$$x^n - 1 = \prod_{d \mid n} \phi_d(x) \tag{3}$$

Voici l'observation cruciale : les *coefficients* des polynômes $\phi_n(x)$ sont *entiers* (c'est-à-dire que $\phi_n(x) \in \mathbb{Z}[x]$ pour tout n) et, en outre, le coefficient constant de ces polynômes vaut 1 ou -1.

Vérifions soigneusement cette affirmation. Pour $n = 1$, l'entier 1 est la seule racine ; donc $\phi_1(x) = x - 1$. Procédons maintenant par récurrence, en

supposant que $\phi_d(x) \in \mathbb{Z}[x]$ pour tout $d < n$ et que le coefficient constant de $\phi_d(x)$ est 1 ou -1. D'après la relation (3) :

$$x^n - 1 \;=\; p(x)\,\phi_n(x) \tag{4}$$

où $p(x) = \sum_{i=0}^{\ell} p_i x^i$, $\phi_n(x) = \sum_{j=0}^{n-\ell} a_j x^j$, avec $p_0 = 1$ ou $p_0 = -1$.

Puisque $-1 = p_0 a_0$, nous voyons que $a_0 \in \{1, -1\}$. Supposons que nous sachions déjà que $a_0, a_1, \ldots, a_{k-1} \in \mathbb{Z}$. En calculant le coefficient de x^k dans les deux membres de (4), nous trouvons :

$$\sum_{i=0}^{k} p_i a_{k-i} \;=\; \sum_{i=1}^{k} p_i a_{k-i} + p_0 a_k \;\in\; \mathbb{Z}$$

Par hypothèse, tous les $a_0, \ldots, a_{k-1}$ (et tous les p_i) appartiennent à $\mathbb{Z}$. Ainsi, $p_0 a_k$ et par conséquent a_k doivent être aussi des entiers, puisque p_0 est égal à 1 ou -1.

Nous sommes prêts à donner *le coup de grâce* [1]. Soit $n_i \mid n$ l'un des nombres apparaissant dans (1). Alors :

$$x^n - 1 \;=\; \prod_{d \mid n} \phi_d(x) \;=\; (x^{n_i} - 1)\phi_n(x). \prod_{d \mid n,\, d \nmid n_i,\, d \neq n} \phi_d(x)$$

On a donc dans $\mathbb{Z}$ les relations de divisibilité suivantes :

$$\phi_n(q) \mid q^n - 1 \qquad \text{et} \qquad \phi_n(q) \,\Big|\, \frac{q^n - 1}{q^{n_i} - 1} \tag{5}$$

Puisque (5) est vérifiée pour tout i, nous déduisons de la formule des classes (1) que :

$$\phi_n(q) \mid q - 1$$

mais c'est impossible. En effet, nous savons que $\phi_n(x) = \prod(x - \lambda)$ où λ parcourt toutes les racines de $x^n - 1$ d'ordre n. Soit $\widetilde{\lambda} = a + ib$ l'une de ces racines. Comme $n > 1$ (puisque $R \neq Z$), nous avons $\widetilde{\lambda} \neq 1$, ce qui implique que la partie réelle a est plus petite que 1. Maintenant $|\widetilde{\lambda}|^2 = a^2 + b^2 = 1$, par conséquent :

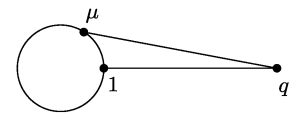
$$\begin{aligned} |q - \widetilde{\lambda}|^2 \;&=\; |q - a - ib|^2 \;=\; (q-a)^2 + b^2 \\ &=\; q^2 - 2aq + a^2 + b^2 \;=\; q^2 - 2aq + 1 \\ &>\; q^2 - 2q + 1 \qquad \text{(puisque } a < 1\text{)} \\ &=\; (q-1)^2 \end{aligned}$$

1. N.d.T. : en français dans le texte original.

et donc on a $|q-\widetilde{\lambda}| > q-1$ pour *toutes* les racines d'ordre n. Cela implique que :

$$|\phi_n(q)| \;=\; \prod_{\lambda} |q-\lambda| > q-1$$

ce qui signifie que $\phi_n(q)$ ne peut pas être un diviseur de $q-1$. □



$|q-\mu| > |q-1|\,.$

Bibliographie

[1] L. E. DICKSON : *On finite algebras,* Nachrichten der Akad. Wissenschaften Göttingen Math.-Phys. Klasse (1905), 1-36 ; Collected Mathematical Papers Vol. III, Chelsea Publ. Comp, The Bronx, NY 1975, 539-574.

[2] J. H. M. WEDDERBURN : *A theorem on finite algebras,* Trans. Amer. Math. Soc. **6** (1905), 349-352.

[3] E. WITT : *Über die Kommutativität endlicher Schiefkörper,* Abh. Math. Sem. Univ. Hamburg **8** (1931), 413.

Quelques nombres irrationnels

Chapitre 7

« π est irrationnel »

C'est ce qu'Aristote a conjecturé lorsqu'il a annoncé que le diamètre et la circonférence d'un cercle ne sont pas commensurables. La première preuve de ce résultat fondamental a été donnée par Johann Heinrich Lambert en 1766. La Preuve de notre Grand Livre est due à Ivan Niven, en 1947 : une preuve en une page, extrêmement élégante, qui ne requiert que des calculs élémentaires. Son idée est puissante et l'on peut en tirer bien davantage, comme l'ont montré Iwamoto et Koksma :

- π^2 est irrationnel ;
- e^r est irrationnel pour tout rationnel $r \neq 0$.

Toutefois, Niven a eu des précurseurs. Sa méthode trouve ses origines dans un article désormais classique de Charles Hermite, datant de 1873 ; Hermite est le premier à avoir établi que e est transcendant, c'est-à-dire que e n'est pas racine d'un polynôme à coefficients rationnels.

Charles Hermite

Suivant ainsi l'ordre chronologique dans lequel ces résultats ont été établis, nous allons nous intéresser au résultat le plus facile à établir, à savoir que e et ses puissances entières sont irrationnels, avant de traiter le cas de π.

$$\begin{aligned} e &:= 1+\tfrac{1}{1}+\tfrac{1}{2}+\tfrac{1}{6}+\tfrac{1}{24}+\dots \\ &= 2.718281828... \\ e^x &:= 1+\tfrac{x}{1}+\tfrac{x^2}{2}+\tfrac{x^4}{6}+\tfrac{x^4}{24}+\dots \\ &= \sum_{k\geqslant 0}\tfrac{x^k}{k!}\,. \end{aligned}$$

Pour commencer, on constate facilement (comme le fit Fourier en 1815) que $e = \sum_{k\geqslant 0}\frac{1}{k!}$ est irrationnel. En effet, si l'on pouvait écrire $e = \frac{a}{b}$ avec a et $b > 0$ entiers, on serait conduit à :

$$n!\,b\,e \;=\; n!\,a$$

pour *tout* $n \geqslant 0$. Une telle égalité ne peut être vraie car on obtiendrait un entier pour le membre droit de l'égalité alors que si l'on écrit :

$$e \;=\; \Big(1+\frac{1}{1!}+\frac{1}{2!}+\ldots+\frac{1}{n!}\Big) + \Big(\frac{1}{(n+1)!}+\frac{1}{(n+2)!}+\frac{1}{(n+3)!}+\dots\Big)$$

le membre gauche de l'égalité se décompose en un premier terme entier :

$$bn!\Big(1+\frac{1}{1!}+\frac{1}{2!}+\ldots+\frac{1}{n!}\Big)$$

et un second terme :

$$b\Big(\frac{1}{n+1}+\frac{1}{(n+1)(n+2)}+\frac{1}{(n+1)(n+2)(n+3)}+\dots\Big)$$

Séries géométriques

Pour une série géométrique se présentant sous la forme :

$$Q = \frac{1}{q} + \frac{1}{q^2} + \frac{1}{q^3} + \dots$$

avec $q > 1$, on trouve :

$$qQ = 1 + \frac{1}{q} + \frac{1}{q^2} + \dots = 1 + Q$$

et donc :

$$Q = \frac{1}{q-1}$$

qui vaut *à peu près* $\frac{b}{n}$, si bien que pour de grandes valeurs de n il ne peut pas être entier. On peut en effet montrer que le terme en question est plus grand que $\frac{b}{n+1}$ et plus petit que $\frac{b}{n}$ en posant :

$$\delta_n = \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \frac{1}{(n+1)(n+2)(n+3)} + \dots$$

et en effectuant une comparaison avec une série géométrique :

$$\frac{1}{n+1} < \delta_n < \frac{1}{n+1} + \frac{1}{(n+1)^2} + \frac{1}{(n+1)^3} + \dots = \frac{1}{n} \qquad \square$$

On peut alors se demander si cette astuce consistant à multiplier par $n!$ permet aussi de montrer que e^2 est irrationnel. Ce résultat est nettement plus fort que la simple irrationnalité de e : le nombre $\sqrt{2}$ est irrationnel mais son carré ne l'est pas.

John Cosgrave nous a indiqué deux astuces qui permettent chacune de prouver l'irrationnalité de e^2 et dont la combinaison conduit à l'irrationnalité de e^4. La première astuce figure dans un article d'une page rédigé par J. Liouville en 1840. La seconde est présentée dans un « *addendum* » que Liouville fit paraître sur les deux pages suivantes du même journal.

Pourquoi e^2 est-il irrationnel ? Que peut-on déduire de $e^2 = \frac{a}{b}$? Selon Liouville, il faut considérer la question de la manière suivante :

$$be = ae^{-1},$$

substituer à e et e^{-1} leurs développements en série :

$$e = 1 + \frac{1}{1} + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} + \dots$$

et :

$$e^{-1} = 1 - \frac{1}{1} + \frac{1}{2} - \frac{1}{6} + \frac{1}{24} - \frac{1}{120} + \dots,$$

puis multiplier par $n!$ pour un entier n pair suffisamment grand. On constate alors que $n!be$ est presqu'un nombre entier puisque :

$$n!b\Big(1 + \frac{1}{1} + \frac{1}{2} + \frac{1}{6} + \dots + \frac{1}{n!}\Big)$$

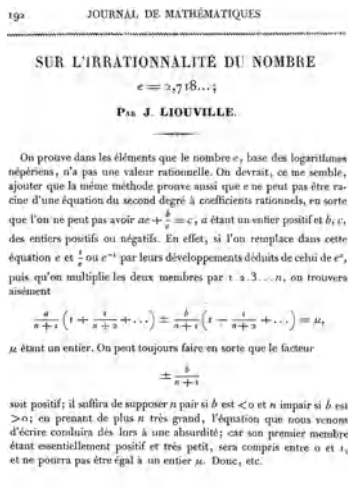
est entier et que le reste :

$$n!b\Big(\frac{1}{(n+1)!} + \frac{1}{(n+2)!} + \dots\Big)$$

vaut à peu près $\frac{b}{n}$ car ce reste est plus grand que $\frac{b}{n+1}$ et plus petit que $\frac{b}{n}$, comme nous l'avons vu précédemment.

De même, $n!ae^{-1}$ est lui aussi presqu'un nombre entier ; nous obtenons un premier terme entier et un reste :

$$(-1)^{n+1}n!a\Big(\frac{1}{(n+1)!} - \frac{1}{(n+2)!} + \frac{1}{(n+3)!} - \dots\Big),$$



192 JOURNAL DE MATHÉMATIQUES

SUR L'IRRATIONNALITÉ DU NOMBRE
$e = 2,718\dots;$

PAR J. LIOUVILLE.

On prouve dans les éléments que le nombre e, base des logarithmes néperiens, n'a pas une valeur rationnelle. On devrait, ce me semble, ajouter que la même méthode prouve aussi que e ne peut pas être racine d'une équation du second degré à coefficients rationnels, en sorte que l'on ne peut pas avoir $ae + \frac{b}{c} = c$, a étant un entier positif et b, c, des entiers positifs ou négatifs. En effet, si l'on remplace dans cette équation e et $\frac{1}{e}$ ou e^{-1} par leurs développements déduits de celui de e^x, puis qu'on multiplie les deux membres par $1.2.3\dots n$, on trouvera aisément

$$\frac{a}{n+1}\left(1 + \frac{1}{n+2} + \dots\right) \pm \frac{b}{n+1}\left(1 - \frac{1}{n+2} + \dots\right) = \mu,$$

μ étant un entier. On peut toujours faire en sorte que le facteur

$$\pm\frac{b}{n+1}$$

soit positif; il suffira de supposer n pair si b est < 0 et n impair si b est > 0; en prenant de plus n très grand, l'équation que nous venons d'écrire conduira dès lors à une absurdité; car son premier membre étant essentiellement positif et très petit, sera compris entre 0 et 1, et ne pourra pas être égal à un entier μ. Donc, etc.

L'article de Liouville.

qui vaut à peu près $(-1)^{n+1}\frac{a}{n}$. Plus précisément, pour un entier n pair, le reste est plus grand que $-\frac{a}{n}$ et plus petit que :

$$-a\Big(\frac{1}{n+1}-\frac{1}{(n+1)^2}-\frac{1}{(n+1)^3}-\dots\Big) \;=\; -\frac{a}{n+1}\Big(1-\frac{1}{n}\Big) \;<\; 0.$$

C'est alors qu'apparaît une contradiction puisque cela signifierait que pour de grands entiers n pairs, $n!ae^{-1}$ serait juste un peu plus petit qu'un entier tandis que $n!be$ serait juste un peu plus grand qu'un entier si bien que l'égalité $n!ae^{-1} = n!be$ ne pourrait pas être vérifiée. □

De manière à montrer que e^4 est irrationnel, nous raisonnons à nouveau par l'absurde, posant courageusement $e^4 = \frac{a}{b}$ et écrivons ceci sous la forme :

$$be^2 \;=\; ae^{-2}.$$

Nous pourrions essayer à nouveau de multiplier par $n!$ pour un grand entier n et regrouper les termes non entiers mais cela ne conduirait à rien d'exploitable. En effet, le reste du membre gauche vaudra à peu près $b\frac{2^{n+1}}{n}$ contre $(-1)^{n+1}a\frac{2^{n+1}}{n}$ pour le reste du membre droit, ces deux termes demeurant grands lorsque n tend vers l'infini.

Il convient donc d'examiner la situation avec un peu plus de soin et d'apporter quelques améliorations à la stratégie mise en œuvre jusqu'alors. Au lieu de prendre un grand entier n *quelconque*, on va plutôt considérer des puissances de 2, c'est-à-dire des entiers de la forme $n = 2^m$. Ensuite, au lieu de multiplier par $n!$, on va multiplier par $\frac{n!}{2^{n-1}}$. On sera alors conduit à utiliser un lemme élémentaire, cas particulier du théorème de Legendre évoqué dans cet ouvrage (voir page 8) : pour tout $n \geqslant 1$, la décomposition en facteurs premiers de l'entier $n!$ fait apparaître le facteur 2 au plus $n-1$ fois — avec égalité si et seulement si n est une puissance de 2.

Ce lemme n'est pas difficile à montrer. $\lfloor\frac{n}{2}\rfloor$ des facteurs de $n!$ sont pairs, $\lfloor\frac{n}{4}\rfloor$ d'entre eux sont multiples de 4 et fournissent donc chacun un nouveau facteur 2 et ainsi de suite. Si donc 2^k est la plus grande puissance de 2 telle que $2^k \leqslant n$, alors la décomposition de $n!$ comporte le facteur 2 exactement :

$$\Big\lfloor\frac{n}{2}\Big\rfloor+\Big\lfloor\frac{n}{4}\Big\rfloor+\dots+\Big\lfloor\frac{n}{2^k}\Big\rfloor$$

fois. Or :

$$\Big\lfloor\frac{n}{2}\Big\rfloor+\Big\lfloor\frac{n}{4}\Big\rfloor+\dots+\Big\lfloor\frac{n}{2^k}\Big\rfloor \;\leqslant\; \frac{n}{2}+\frac{n}{4}+\dots+\frac{n}{2^k} \;=\; n\Big(1-\frac{1}{2^k}\Big) \;\leqslant\; n-1$$

avec égalité dans les deux inégalités exactement si $n = 2^k$.

Revenons à $be^2 = ae^{-2}$. Observons maintenant :

$$b\frac{n!}{2^{n-1}}e^2 \;=\; a\frac{n!}{2^{n-1}}e^{-2} \qquad (1)$$

et substituons à e^2 et e^{-2} leurs développements en série :

$$e^2 = 1+\frac{2}{1}+\frac{4}{2}+\frac{8}{6}+\;\dots\;+\frac{2^r}{r!}+\dots$$

et :

$$e^{-2} = 1 - \frac{2}{1} + \frac{4}{2} - \frac{8}{6} + \ \dots \ + (-1)^r \frac{2^r}{r!} + \dots$$

Pour $r \leqslant n$ nous obtenons des termes généraux entiers dans chaque membre, à savoir :

$$b\frac{n!}{2^{n-1}}\frac{2^r}{r!} \quad \text{et} \quad (-1)^r a\frac{n!}{2^{n-1}}\frac{2^r}{r!},$$

où pour $r > 0$ le dénominateur $r!$ contient le facteur 2 *au plus* $r-1$ fois tandis que $n!$ le contient *exactement* $n-1$ fois. Ainsi, pour $r > 0$, les termes généraux sont pairs.

Comme n est pair (on a supposé $n = 2^m$), les restes que l'on obtient pour $r \geqslant n+1$ sont :

$$2b\Big(\frac{2}{n+1} + \frac{4}{(n+1)(n+2)} + \frac{8}{(n+1)(n+2)(n+3)} + \ \dots\Big)$$

et :

$$2a\Big(-\frac{2}{n+1} + \frac{4}{(n+1)(n+2)} - \frac{8}{(n+1)(n+2)(n+3)} + \ \dots\Big).$$

Pour un grand entier n, ces restes valent à peu près $\frac{4b}{n}$ et $-\frac{4a}{n}$, toujours en comparant à une série géométrique. Pour un entier $n = 2^m$, cela signifie que le membre gauche de (1) est *un peu* plus grand qu'un entier tandis que le membre droit de (1) est *un peu* plus petit qu'un entier : contradiction ! □

Nous savons donc maintenant que e^4 est irrationnel. Pour montrer que e^3, e^5, etc. sont aussi irrationnels, il faut recourir à un argumentaire plus élaboré, c'est-à-dire se tourner vers le calcul différentiel et intégral tout en s'appuyant sur une nouvelle idée — qui remonte pour l'essentiel à Charles Hermite et dont la clé réside dans le simple lemme suivant.

Lemme. *Pour un entier $n \geqslant 1$ fixé, on pose :*

$$f(x) \ = \ \frac{x^n(1-x)^n}{n!}$$

Alors :
(i) *La fonction f est une fonction polynôme qui peut s'écrire :*

$$f(x) = \frac{1}{n!}\sum_{i=n}^{2n} c_i x^i$$

où les coefficients c_i sont entiers.
(ii) *Pour $0 < x < 1$, on a $0 < f(x) < \frac{1}{n!}$.*
(iii) *Les dérivées $f^{(k)}(0)$ et $f^{(k)}(1)$ sont entières pour tout $k \geqslant 0$.*

■ **Preuve.** Les résultats (i) et (ii) sont évidents.
Pour (iii), notons que (i) implique que la k-ième dérivée $f^{(k)}$ s'annule en $x = 0$ sauf si $n \leqslant k \leqslant 2n$; dans ce cas $f^{(k)}(0) = \frac{k!}{n!}c_k$ est un entier. De $f(x) = f(1-x)$ on tire que $f^{(k)}(x) = (-1)^k f^{(k)}(1-x)$ pour tout x, et par conséquent que $f^{(k)}(1) = (-1)^k f^{(k)}(0)$ est un entier. □

Théorème 1. *e^r est irrationnel pour tout $r \in \mathbb{Q}\backslash\{0\}$.*

■ **Preuve.** Il suffit de montrer que e^s ne peut être rationnel pour aucun entier positif s (si $e^{\frac{s}{t}}$ était rationnel, $\left(e^{\frac{s}{t}}\right)^t = e^s$ le serait aussi). Supposons que $e^s = \frac{a}{b}$, où $a, b > 0$ sont des entiers, et soit n suffisamment grand pour que $n! > as^{2n+1}$. Posons :

La majoration $n! > e(\frac{n}{e})^n$ fournit explicitement un n qui est « suffisamment grand ».

$$F(x) := s^{2n}f(x) - s^{2n-1}f'(x) + s^{2n-2}f''(x) - \ldots + f^{(2n)}(x)$$

où f désigne la fonction définie au lemme précédent. $F(x)$ peut aussi s'écrire sous la forme :

$$F(x) = s^{2n}f(x) - s^{2n-1}f'(x) + s^{2n-2}f''(x) - \ldots$$

puisque les dérivées $f^{(k)}(x)$ d'ordres $k > 2n$ sont nulles. On en déduit que le polynôme $F(x)$ vérifie l'identité :

$$F'(x) = -s\,F(x) + s^{2n+1}f(x)$$

En dérivant, on obtient :

$$\frac{d}{dx}\left[e^{sx}F(x)\right] = se^{sx}F(x) + e^{sx}F'(x) = s^{2n+1}e^{sx}f(x)$$

et par conséquent :

$$N := b\int_0^1 s^{2n+1}e^{sx}f(x)dx = b\left[e^{sx}F(x)\right]_0^1 = aF(1) - bF(0)$$

N est un entier puisque la partie (iii) du lemme implique que $F(0)$ et $F(1)$ sont des entiers. Cependant, la partie (ii) du lemme donne un encadrement de N :

$$0 < N = b\int_0^1 s^{2n+1}e^{sx}f(x)dx < bs^{2n+1}e^s\frac{1}{n!} = \frac{as^{2n+1}}{n!} < 1$$

qui montre que N ne peut pas être un entier : il y a contradiction. □

Comme cette astuce a si bien fonctionné, nous allons l'utiliser encore une fois.

Théorème 2. *π^2 est irrationnel.*

■ **Preuve.** Supposons que $\pi^2 = \frac{a}{b}$ pour des entiers $a, b > 0$. Introduisons maintenant le polynôme :

$$F(x) := b^n\left(\pi^{2n}f(x) - \pi^{2n-2}f^{(2)}(x) + \pi^{2n-4}f^{(4)}(x) - \ldots\right)$$

qui vérifie $F''(x) = -\pi^2F(x) + b^n\pi^{2n+2}f(x)$.
De la partie (iii) du lemme, on déduit que $F(0)$ et $F(1)$ sont des entiers. Les règles élémentaires de dérivation conduisent à :

$$\begin{aligned}\frac{d}{dx}\left[F'(x)\sin\pi x - \pi F(x)\cos\pi x\right] &= \left(F''(x) + \pi^2F(x)\right)\sin\pi x \\ &= b^n\pi^{2n+2}f(x)\sin\pi x \\ &= \pi^2a^nf(x)\sin\pi x\end{aligned}$$

π n'est pas rationnel, mais il admet de « bonnes approximations » par des rationnels. Certaines d'entre elles sont connues depuis l'antiquité :

$$\begin{aligned} \tfrac{22}{7} &= 3.142857142857... \\ \tfrac{355}{113} &= 3.141592920353... \\ \tfrac{104348}{33215} &= 3.141592653921... \\ \pi &= 3.141592653589... \end{aligned}$$

et l'on obtient ainsi :

$$\begin{aligned} N := \pi \int_0^1 a^n f(x) \sin \pi x \, dx &= \Big[\frac{1}{\pi} F'(x) \sin \pi x - F(x) \cos \pi x\Big]_0^1 \\ &= F(0) + F(1) \end{aligned}$$

qui est un entier. En outre, N est positif puisqu'il est défini comme l'intégrale d'une fonction qui est positive (sauf aux bornes). Cependant, si n est choisi suffisamment grand pour que $\frac{\pi a^n}{n!} < 1$, alors la partie (ii) du lemme permet d'écrire :

$$0 < N = \pi \int_0^1 a^n f(x) \sin \pi x \, dx < \frac{\pi a^n}{n!} < 1$$

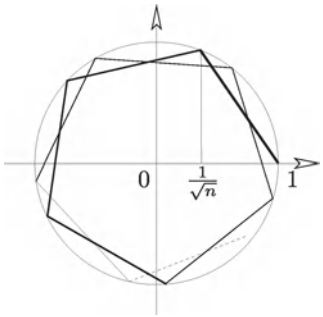
ce qui contredit le résultat précédent. □

Voici un dernier résultat sur les nombres irrationnels.

Théorème 3. *Pour tout entier impair $n \geqslant 3$, le nombre*

$$A(n) := \frac{1}{\pi} \arccos\left(\frac{1}{\sqrt{n}}\right)$$

est irrationnel.



Nous aurons besoin de ce résultat pour traiter le troisième problème de Hilbert (voir chapitre 9) lorsque $n = 3$ et $n = 9$. Si $n = 2$ ou $n = 4$, nous avons $A(2) = \frac{1}{4}$ et $A(4) = \frac{1}{3}$, la restriction aux entiers impairs est donc essentielle. Ces résultats se voient sur la figure représentée dans la marge. L'affirmation « $\frac{1}{\pi} \arccos\left(\frac{1}{\sqrt{n}}\right)$ est irrationnel » se traduit par le fait que l'arc polygonal construit à partir de $\frac{1}{\sqrt{n}}$ et dont toutes les cordes ont la même longueur, ne se referme jamais.

Nous laissons comme exercice au lecteur le soin de montrer que $A(n)$ est rationnel *seulement* si $n \in \{1, 2, 4\}$. À cet effet, il convient de distinguer les cas où $n = 2^r$ de ceux où n n'est pas une puissance de 2.

■ **Preuve.** On utilise la formule d'addition :

$$\cos \alpha + \cos \beta = 2 \cos \tfrac{\alpha+\beta}{2} \cos \tfrac{\alpha-\beta}{2}$$

issue de la trigonométrie élémentaire, qui donne pour $\alpha = (k+1)\varphi$ et $\beta = (k-1)\varphi$:

$$\cos(k+1)\varphi = 2 \cos\varphi \cos k\varphi - \cos(k-1)\varphi \qquad (2)$$

Pour l'angle $\varphi_n = \arccos\left(\frac{1}{\sqrt{n}}\right)$ (défini par $\cos\varphi_n = \frac{1}{\sqrt{n}}$ et $0 \leqslant \varphi_n \leqslant \pi$), on obtient des représentations de la forme :

$$\cos k\varphi_n = \frac{A_k}{\sqrt{n}^k}$$

où A_k est un entier qui n'est divisible par n pour aucun $k \geqslant 0$. En effet, on a de telles représentations pour $k = 0, 1$ puisque $A_0 = A_1 = 1$ et, par récurrence sur k en utilisant (2), on obtient, si $k \geqslant 1$:

$$\cos(k+1)\varphi_n \;=\; 2\frac{1}{\sqrt{n}}\frac{A_k}{\sqrt{n}^k} - \frac{A_{k-1}}{\sqrt{n}^{k-1}} \;=\; \frac{2A_k - nA_{k-1}}{\sqrt{n}^{k+1}}$$

Ainsi, $A_{k+1} = 2A_k - nA_{k-1}$. Si $n \geqslant 3$ est impair et si A_k n'est pas divisible par n, alors A_{k+1} ne peut pas non plus être divisible par n.

Supposons maintenant que :

$$A(n) \;=\; \frac{1}{\pi}\varphi_n \;=\; \frac{k}{\ell}$$

soit rationnel ($k, \ell > 0$ étant des entiers). Alors $\ell\varphi_n = k\pi$ implique que :

$$\pm 1 \;=\; \cos k\pi \;=\; \frac{A_\ell}{\sqrt{n}^\ell}$$

Donc $\sqrt{n}^\ell = \pm A_\ell$ est un entier, avec $\ell \geqslant 2$ et par conséquent $n \mid \sqrt{n}^\ell$. Comme $\sqrt{n}^\ell \mid A_\ell$, n divise A_ℓ, ce qui est contradictoire. □

Bibliographie

[1] C. HERMITE : *Sur la fonction exponentielle,* Comptes rendus de l'Académie des Sciences (Paris) **77** (1873), 18-24 ; Œuvres de Charles Hermite, Vol. III, Gauthier-Villars, Paris 1912, pp. 150-181.

[2] Y. IWAMOTO : *A proof that π^2 is irrational,* J. Osaka Institute of Science and Technology **1** (1949), 147-148.

[3] J. F. KOKSMA : *On Niven's proof that π is irrational,* Nieuw Archief voor Wiskunde (2) **23** (1949), 39.

[4] J. LIOUVILLE : *Sur l'irrationalité du nombre e = 2,718...,* Journal de Mathématiques Pures et Appl. (1) **5** (1840), 192 ; *Addition*, 193-194.

[5] I. NIVEN : *A simple proof that π is irrational,* Bulletin Amer. Math. Soc. **53** (1947), 509.

Trois méthodes pour calculer $\pi^2/6$

Chapitre 8

Nous savons que la série $\sum_{n\geqslant 1}\frac{1}{n}$ n'est pas convergente et nous avons vu au chapitre 1 que même la série $\sum_{p\in\mathbb{P}}\frac{1}{p}$ diverge.
Toutefois, la série des inverses des carrés converge (bien que très lentement comme nous le verrons) vers une valeur intéressante.

Série d'Euler.

$$\sum_{n\geqslant 1}\frac{1}{n^2} = \frac{\pi^2}{6}.$$

Cette formule célèbre, établie par Leonhard Euler en 1734, a une conséquence fondamentale : elle fournit la première valeur non triviale, $\zeta(2)$, de la fonction Zêta de Riemann (voir appendice en page 56). Cette valeur est irrationnelle, comme nous l'avons vu au chapitre 7.

Ce résultat occupe une place particulière dans l'histoire des mathématiques ; il a été démontré à plusieurs reprises et de différentes façons. Certaines preuves particulièrement élégantes ou astucieuses ont été découvertes et redécouvertes par plusieurs auteurs. Dans ce chapitre nous en présentons trois.

■ **Preuve.** La première démonstration apparaît comme exercice dans l'ouvrage de théorie des nombres que publia William J. LeVeque en 1956. Il dit cependant : « Je ne sais pas du tout d'où vient cette idée mais je suis convaincu que je n'en suis pas l'auteur ».
La preuve consiste à évaluer de deux façons différentes l'intégrale double [1] :

$$I := \iint_{[0,1]\times[0,1]} \frac{1}{1-xy}\,dx\,dy$$

Pour la première, on développe $\frac{1}{1-xy}$ en série géométrique, avant de décomposer l'intégrale double en produit d'intégrales simples qui se calculent

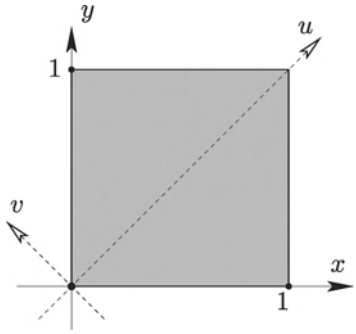
1. N.d.T. : il s'agit de l'intégrale d'une fonction positive sur le domaine considéré, son existence dans $\overline{\mathbb{R}}$ en résulte.

Timbre suisse de 1957 à l'effigie de Leonhard Euler.

1	$=$	$1,000000$
$1+\frac{1}{4}$	$=$	$1,250000$
$1+\frac{1}{4}+\frac{1}{9}$	$=$	$1,361111$
$1+\frac{1}{4}+\frac{1}{9}+\frac{1}{16}$	$=$	$1,423611$
$1+\frac{1}{4}+\frac{1}{9}+\frac{1}{16}+\frac{1}{25}$	$=$	$1,463611$
$1+\frac{1}{4}+\frac{1}{9}+\frac{1}{16}+\frac{1}{25}+\frac{1}{36}$	$=$	$1,491388$
$\pi^2/6$	$=$	$1,644934$

sans effort :

$$\begin{aligned} I &= \iint_{[0,1]\times[0,1]} \sum_{n\geqslant 0} (xy)^n \, dx\, dy = \sum_{n\geqslant 0} \iint_{[0,1]\times[0,1]} x^n y^n \, dx\, dy \\ &= \sum_{n\geqslant 0} \left(\int_0^1 x^n dx \right) \left(\int_0^1 y^n dy \right) = \sum_{n\geqslant 0} \frac{1}{n+1} \frac{1}{n+1} \\ &= \sum_{n\geqslant 0} \frac{1}{(n+1)^2} = \sum_{n\geqslant 1} \frac{1}{n^2} = \zeta(2) \end{aligned}$$



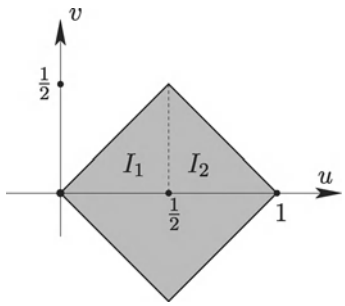
Ce calcul montre aussi que cette intégrale double (d'une fonction positive ayant un pôle en $x = y = 1$) est finie. Notons que le calcul est aussi facile et direct si nous le lisons à l'envers : l'évaluation de $\zeta(2)$ conduit à l'intégrale double I.

La deuxième manière d'évaluer I repose sur le changement de variables suivant :

$$u = \frac{y+x}{2} \quad \text{et} \quad v = \frac{y-x}{2}$$

Le nouveau domaine d'intégration est un carré de côté $\frac{1}{2}\sqrt{2}$, transformé de l'ancien domaine par la composée d'une rotation d'angle $45°$ et d'une homothétie de rapport $\frac{1}{\sqrt{2}}$. En substituant ces nouvelles coordonnées, on obtient :

$$\frac{1}{1-xy} = \frac{1}{1-u^2+v^2}$$



Pour transformer l'intégrale, il convient de remplacer $dx\, dy$ par $2\, du\, dv$ afin de compenser le fait que le changement de variables réduit les aires dans un facteur constant 2 (qui est en fait la valeur absolue du déterminant jacobien de la transformation ; voir encadré).

Le nouveau domaine d'intégration et la fonction à intégrer sont symétriques relativement à l'axe des u, il suffit donc simplement de doubler la valeur de l'intégrale sur la moitié supérieure du domaine, que nous divisons naturellement en deux parties (ce qui fait apparaître un nouveau facteur 2) :

$$I = 4 \int_0^{\frac{1}{2}} \left(\int_0^u \frac{dv}{1-u^2+v^2} \right) du + 4 \int_{\frac{1}{2}}^1 \left(\int_0^{1-u} \frac{dv}{1-u^2+v^2} \right) du$$

Sachant que $\displaystyle\int \frac{dx}{a^2+x^2} = \frac{1}{a} \arctan\left(\frac{x}{a}\right) + C$, on trouve :

$$\begin{aligned} I = \quad & 4 \int_0^{\frac{1}{2}} \frac{1}{\sqrt{1-u^2}} \arctan\left(\frac{u}{\sqrt{1-u^2}} \right) du \\ + \quad & 4 \int_{\frac{1}{2}}^1 \frac{1}{\sqrt{1-u^2}} \arctan\left(\frac{1-u}{\sqrt{1-u^2}} \right) du \end{aligned}$$

Ces expressions peuvent être simplifiées. En effet, les valeurs de la fonction arctan qui apparaissent dans les intégrales sont des angles θ dans

l'intervalle $0 \leqslant \theta < \pi/2$. Si $u = \sin\theta$, alors $\sqrt{1-u^2} = \cos\theta$ donc $\frac{u}{\sqrt{1-u^2}} = \tan\theta$. De même, si $u = \cos\theta$, alors $1-u = 1-\cos\theta = 2\sin^2\frac{\theta}{2}$ et $\sqrt{1-u^2} = \sin\theta = 2\cos\frac{\theta}{2}\sin\frac{\theta}{2}$, ainsi, $\frac{1-u}{\sqrt{1-u^2}} = \tan\frac{\theta}{2}$. Nous faisons ainsi d'une pierre deux coups puisque les intégrales à calculer se simplifient en :

$$I \;=\; 4\int_0^{\frac{1}{2}} \frac{\arcsin(u)}{\sqrt{1-u^2}}\,du \;+\; 4\int_{\frac{1}{2}}^1 \frac{\frac{1}{2}\arccos(u)}{\sqrt{1-u^2}}\,du$$

et que l'on reconnaît des intégrales de la forme $\int f(u)f'(u)du$ qui s'intègrent donc en $\frac{1}{2}f^2(u)$:

$$\begin{aligned} I = \quad & 4\left[\tfrac{1}{2}\arcsin(u)^2\right]_0^{1/2} \;-\; 2\left[\tfrac{1}{2}\arccos(u)^2\right]_{1/2}^{1} \\ = \quad & 2\left(\frac{\pi}{6}\right)^2 \;+\; \left(\frac{\pi}{3}\right)^2 \;=\; \frac{\pi^2}{6} \end{aligned}$$ □

Cette démonstration tire la valeur de la série d'Euler d'une intégrale à l'aide d'un changement de variables relativement simple. Une autre version ingénieuse de la démonstration fondée sur le même principe mais reposant cette fois sur un changement de variables non trivial, a été découverte plus tard par Beukers, Calabi et Kolk. Le point de départ de leur preuve consiste à séparer les termes pairs des termes impairs dans la somme $\sum_{n\geqslant 1}\frac{1}{n^2}$. Manifestement, les termes pairs $\frac{1}{2^2}+\frac{1}{4^2}+\frac{1}{6^2}+\ldots = \sum_{k\geqslant 1}\frac{1}{(2k)^2}$ contribuent au résultat à hauteur de $\frac{1}{4}\zeta(2)$, de sorte que les termes impairs $\frac{1}{1^2}+\frac{1}{3^2}+\frac{1}{5^2}+\ldots = \sum_{k\geqslant 0}\frac{1}{(2k+1)^2}$ contribuent aux trois quarts de ce même résultat. Ainsi, le théorème précédent est équivalent à :

$$\sum_{k\geqslant 0}\frac{1}{(2k+1)^2} \;=\; \frac{\pi^2}{8}.$$

Comme précédemment, on fait appel à une intégrale double, en l'occurrence :

$$J \;=\; \iint_{[0,1]\times[0,1]} \frac{1}{1-x^2y^2}\,dx\,dy \;=\; \sum_{k\geqslant 0}\frac{1}{(2k+1)^2}$$

On cherche donc à calculer cette intégrale J. À cet effet, Beukers, Calabi et Kolk ont proposé le changement de variables suivant :

$$u \;:=\; \arccos\sqrt{\frac{1-x^2}{1-x^2y^2}} \qquad v \;:=\; \arccos\sqrt{\frac{1-y^2}{1-x^2y^2}}$$

Pour le calcul de cette intégrale double, on peut négliger les bords du domaine d'intégration et considérer que x et y appartiennent à l'intérieur du rectangle *ouvert* $0 < x < 1$ et $0 < y < 1$. Par suite, u et v vont appartenir

La formule du changement de variables

Si $S, T \subseteq \mathbb{R}^2$ sont des domaines plans, et si $\Phi : T \longrightarrow S$ est une application bijective, de classe C^1 sur l'intérieur de T et dont la réciproque est de classe C^1 sur l'intérieur de S définie par :

$$\begin{array}{rccc} \Phi: & T & \longrightarrow & S \\ & (u,v) & \longmapsto & (x(u,v), y(u,v)) \end{array}$$

alors pour calculer l'intégrale :

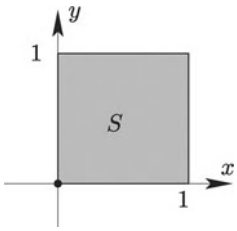
$$I = \iint_S f(x,y)\,dx\,dy$$

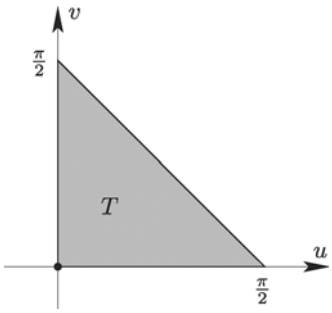
on peut effectuer le changement de variables défini par Φ ; l'intégrale I s'écrit encore :

$$\iint_T f(x(u,v), y(u,v))\,|J_\Phi(u,v)|\,du\,dv$$

où $J_\Phi(u,v)$ désigne le déterminant jacobien associé à la transformation Φ :

$$J_\Phi(u,v) \;=\; \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}.$$





à l'intérieur du triangle ouvert $u > 0$, $v > 0$, $u + v < \pi/2$. Ce changement de variables peut être inversé explicitement et l'on trouve :

$$x = \frac{\sin u}{\cos v} \quad \text{and} \quad y = \frac{\sin v}{\cos u}$$

On vérifie facilement que ces formules définissent une transformation bijective (et continûment différentiable ainsi que sa réciproque) de l'intérieur du carré unité $S = \{(x, y) : 0 \leqslant x, y \leqslant 1\}$ sur l'intérieur du triangle $T = \{(u, v) : u, v \geqslant 0,\ u + v \leqslant \pi/2\}$.

Il reste à calculer le déterminant jacobien associé au changement de variables. Sa valeur se révèle être :

$$\det \begin{pmatrix} \frac{\cos u}{\cos v} & \frac{\sin u \sin v}{\cos^2 v} \\ \frac{\sin u \sin v}{\cos^2 u} & \frac{\cos v}{\cos u} \end{pmatrix} = 1 - \frac{\sin^2 u \sin^2 v}{\cos^2 u \cos^2 v} = 1 - x^2 y^2$$

L'intégrale initiale se ramène donc finalement à :

$$J = \iint_T 1 \, du \, dv$$

qui n'est autre que l'aire $\frac{1}{2}(\frac{\pi}{2})^2 = \frac{\pi^2}{8}$ du triangle T. □

Splendide ! Et ce d'autant plus que la même méthode se généralise au calcul de $\zeta(2k)$ comme intégrale $2k$-dimensionnelle pour tous les $k \geqslant 1$. Le lecteur peut se reporter à l'article original de Beuker, Calabi et Kolk [2] ainsi qu'au chapitre 23 où l'on établira le même résultat d'une autre manière en ayant recours à l'astuce de Herglotz et à la méthode initialement développée par Euler.

Après ces deux démonstrations qui utilisent le changement de variables dans le calcul d'une intégrale, nous ne pouvons résister à la tentation de présenter une preuve complètement élémentaire de $\sum_{n \geqslant 1} \frac{1}{n^2} = \frac{\pi^2}{6}$. Elle apparaît dans une série d'exercices du livre écrit par les frères jumeaux Akiva et Isaak Yaglom dont l'édition originale russe parut en 1954. Des versions de cette preuve ont été redécouvertes et présentées par F. Holme (1970), I. Papadimitriou (1973) et par Ransford (1982) qui l'attribua à John Scholes.

■ **Preuve.** La première étape consiste à établir une relation remarquable entre (les carrés de) certaines valeurs de la fonction cotangente. Pour tout $m \geqslant 1$, on peut affirmer :

$$\cot^2\left(\tfrac{\pi}{2m+1}\right) + \cot^2\left(\tfrac{2\pi}{2m+1}\right) + \ldots + \cot^2\left(\tfrac{m\pi}{2m+1}\right) = \tfrac{2m(2m-1)}{6} \quad (1)$$

Pour établir ce résultat, on part de la relation $e^{ix} = \cos x + i \sin x$. En élevant la relation précédente à la puissance n, on trouve $e^{inx} = (e^{ix})^n$ et on est conduit à :

$$\cos nx + i \sin nx = (\cos x + i \sin x)^n$$

Pour $m = 1, 2, 3$, on trouve :
$\cot^2 \frac{\pi}{3} = \frac{1}{3}$,
$\cot^2 \frac{\pi}{5} + \cot^2 \frac{2\pi}{5} = 2$,
$\cot^2 \frac{\pi}{7} + \cot^2 \frac{2\pi}{7} + \cot^2 \frac{3\pi}{7} = 5$.

On extrait la partie imaginaire de chaque membre :

$$\sin nx = \binom{n}{1}\sin x\cos^{n-1}x - \binom{n}{3}\sin^3 x\cos^{n-3}x + \ldots \qquad (2)$$

Posons maintenant $n = 2m+1$ et faisons prendre à x les m valeurs distinctes $x = \frac{r\pi}{2m+1}$, avec $r = 1, 2, \ldots, m$. Pour chacune de ces valeurs, nous avons $nx = r\pi$ et donc $\sin nx = 0$, alors que $0 < x < \frac{\pi}{2}$ implique que $\sin x$ prend m valeurs strictement positives distinctes.

En particulier, nous pouvons diviser les deux membres de (2) par $\sin^n x$, ce qui donne :

$$0 = \binom{n}{1}\cot^{n-1}x - \binom{n}{3}\cot^{n-3}x + \ldots$$

soit encore :

$$0 = \binom{2m+1}{1}\cot^{2m}x - \binom{2m+1}{3}\cot^{2m-2}x + \ldots$$

pour chacune des m valeurs distinctes de x. Ainsi, le polynôme de degré m

$$p(t) := \binom{2m+1}{1}t^m - \binom{2m+1}{3}t^{m-1} + \ldots + (-1)^m\binom{2m+1}{2m+1}$$

possède les m racines distinctes :

$$a_r = \cot^2\left(\tfrac{r\pi}{2m+1}\right) \quad \text{pour} \quad r = 1, 2, \ldots, m$$

Par conséquent, le polynôme coïncide avec :

$$p(t) = \binom{2m+1}{1}\left(t - \cot^2\left(\tfrac{\pi}{2m+1}\right)\right)\cdot\ldots\cdot\left(t - \cot^2\left(\tfrac{m\pi}{2m+1}\right)\right)$$

La comparaison entre les coefficients de t^{m-1} dans $p(t)$ implique maintenant que la somme des racines est :

$$a_1 + \ldots + a_r = \frac{\binom{2m+1}{3}}{\binom{2m+1}{1}} = \frac{2m(2m-1)}{6}$$

Comparaison de coefficients :
Si $p(t) = c(t-a_1)\cdots(t-a_m)$, le coefficient de t^{m-1} est $-c(a_1 + \ldots + a_m)$.

ce qui prouve (1).

Nous avons encore besoin d'une deuxième identité du même type :

$$\csc^2\left(\tfrac{\pi}{2m+1}\right) + \csc^2\left(\tfrac{2\pi}{2m+1}\right) + \ldots + \csc^2\left(\tfrac{m\pi}{2m+1}\right) = \frac{2m(2m+2)}{6} \qquad (3)$$

où la fonction cosécante est définie par $\csc x = \frac{1}{\sin x}$. Comme :

$$\csc^2 x = \frac{1}{\sin^2 x} = \frac{\cos^2 x + \sin^2 x}{\sin^2 x} = \cot^2 x + 1$$

nous pouvons déduire (3) de (1) en ajoutant m aux deux membres de l'équation.

Tout est maintenant en place. Dans l'intervalle $0 < y < \frac{\pi}{2}$ on observe que :

$0 < a < b < c$
implique :
$0 < \frac{1}{c} < \frac{1}{b} < \frac{1}{a}$.

$$0 \;<\; \sin y \;<\; y \;<\; \tan y,$$

et donc :

$$0 \;<\; \cot y \;<\; \frac{1}{y} \;<\; \csc y$$

ce qui implique :

$$\cot^2 y \;<\; \frac{1}{y^2} \;<\; \csc^2 y$$

Appliquons cette inégalité à chacune des m valeurs distinctes de x et additionnons membre à membre les inégalités obtenues. En utilisant (1) pour le membre gauche et (3) pour le membre droit, on trouve :

$$\tfrac{2m(2m-1)}{6} \;<\; \left(\tfrac{2m+1}{\pi}\right)^2 + \left(\tfrac{2m+1}{2\pi}\right)^2 + \ldots + \left(\tfrac{2m+1}{m\pi}\right)^2 \;<\; \tfrac{2m(2m+2)}{6}$$

c'est-à-dire :

$$\tfrac{\pi^2}{6}\,\tfrac{2m}{2m+1}\,\tfrac{2m-1}{2m+1} \;<\; \tfrac{1}{1^2} + \tfrac{1}{2^2} + \ldots + \tfrac{1}{m^2} \;<\; \tfrac{\pi^2}{6}\,\tfrac{2m}{2m+1}\,\tfrac{2m+2}{2m+1}$$

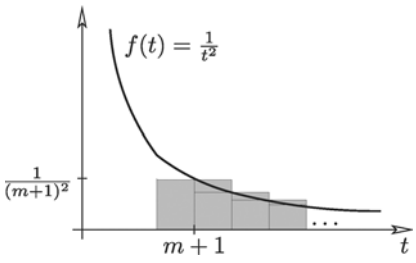
Majorant et minorant convergent vers $\frac{\pi^2}{6}$ lorsque $m \longrightarrow \infty$ d'où la conclusion. □

À quelle vitesse $\sum \frac{1}{n^2}$ converge-t-elle vers $\pi^2/6$? Pour le savoir, il faut évaluer la différence :

$$\frac{\pi^2}{6} - \sum_{n=1}^{m} \frac{1}{n^2} \;=\; \sum_{n=m+1}^{\infty} \frac{1}{n^2}.$$

Cela est très facile si l'on « compare à une intégrale », comme nous l'avons fait dans l'appendice du chapitre 2 (page 11). On obtient de la sorte un majorant du reste :

$$\sum_{n=m+1}^{\infty} \frac{1}{n^2} \;<\; \int_m^{\infty} \frac{1}{t^2}dt \;=\; \frac{1}{m}$$



ainsi qu'un minorant :

$$\sum_{n=m+1}^{\infty} \frac{1}{n^2} \;>\; \int_{m+1}^{\infty} \frac{1}{t^2}dt \;=\; \frac{1}{m+1}\,.$$

Si l'on veut être un peu plus précis, on a même :

$$\sum_{n=m+1}^{\infty} \frac{1}{n^2} \;>\; \int_{m+\frac{1}{2}}^{\infty} \frac{1}{t^2}dt \;=\; \frac{1}{m+\frac{1}{2}}$$

en utilisant la convexité de la fonction $t \mapsto \frac{1}{t^2}$.

Cela signifie que la série ne converge pas très vite. Si l'on calcule la somme partielle des mille premiers termes, on peut espérer une valeur approchée de la somme à 10^{-3} près. De même, si l'on calcule la somme partielle du premier million de termes, on peut espérer une valeur approchée à 10^{-6} près. C'est alors que se présente un phénomène très surprenant : on peut obtenir pour le même prix une précision à 10^{-45} près.

$$\begin{aligned} \pi^2/6 &= 1,64493406684822643647241516664602518921894990 1 \\ \sum_{n=1}^{10^6} \frac{1}{n^2} &= 1,64493306684872643630574849997939185588561654 4 \end{aligned}$$

L'observation des deux valeurs approchées reproduites ci-dessus permet de constater que le sixième chiffre après la virgule de la somme partielle est erroné (inférieur à la bonne valeur de 1) mais *les six chiffres suivants sont corrects* ! Puis à nouveau le chiffre suivant est erroné (supérieur à la bonne valeur de 5) puis de nouveau les cinq chiffres suivants sont corrects. Cette découverte surprenante très récente est due à Roy D. North de Colorado Springs en 1988. En 1982, Martin R. Powell, un professeur de lycée d'Amersham, Bucks, en Angleterre, passa à côté de cet effet remarquable faute de moyens informatiques suffisants à l'époque. Ce phénomène est trop étrange pour n'être qu'une coïncidence... L'observation de la valeur approchée du reste avec 45 chiffres significatifs :

$$\sum_{n=10^6+1}^{\infty} \frac{1}{n^2} = 0,000000999999500000166666666666633333333333357$$

fait clairement apparaître une forme régulière. On peut encore écrire ce nombre sous la forme :

$$+10^{-6} - \tfrac{1}{2}10^{-12} + \tfrac{1}{6}10^{-18} - \tfrac{1}{30}10^{-30} + \tfrac{1}{42}10^{-42} + \ldots$$

où les coefficients $(1, -\frac{1}{2}, \frac{1}{6}, 0, -\frac{1}{30}, 0, \frac{1}{42})$ des 10^{-6i} constituent le début de la suite des *nombres de Bernoulli* que nous rencontrerons à nouveau au chapitre 23. Nous renvoyons le lecteur à l'article de Borwein, Borwein & Dilcher [3] pour d'autres « coïncidences » surprenantes — et pour les démonstrations qui leur sont associées.

Appendice - La fonction Zêta de Riemann

La *fonction Zêta de Riemann* $\zeta(s)$ est définie pour tout réel $s > 1$ par :

$$\zeta(s) \ := \ \sum_{n \geqslant 1} \frac{1}{n^s}$$

Les estimations établies pour H_n (voir page 11) impliquent que la série $\sum_{n \geqslant 1} \frac{1}{n^s}$ diverge pour $s = 1$ et converge pour tout réel $s > 1$ ce qui entraîne l'existence de $\zeta(s)$ pour $s > 1$. La fonction Zêta admet un prolongement canonique sur le plan complexe privé du point $s = 1$ (la fonction présente un pôle simple en ce point), qui peut être construit en utilisant des développements en séries entières. La fonction complexe qui en résulte est de la plus haute importance pour la théorie des nombres premiers. Mentionnons quatre résultats différents :

(1) L'identité remarquable :

$$\zeta(s) \ = \ \prod_{p} \frac{1}{1 - p^{-s}} \, ,$$

établie par Euler, contient le fait que chaque entier naturel admet une décomposition unique (!) en facteurs premiers ; en effet, à partir de la décomposition des entiers, le résultat d'Euler est une simple conséquence du développement en série entière :

$$\frac{1}{1 - p^{-s}} \ = \ 1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \frac{1}{p^{3s}} + \cdots$$

(2) La méthode merveilleuse suivante, due à Don Zagier, permet de calculer $\zeta(4)$ à partir de $\zeta(2)$. Considérons la fonction :

$$f(m,n) \ = \ \frac{2}{m^3 n} + \frac{1}{m^2 n^2} + \frac{2}{m n^3}$$

définie pour les entiers $m, n \geqslant 1$. On vérifie facilement que pour tout m et pour tout n :

$$f(m,n) - f(m+n,n) - f(m,m+n) \ = \ \frac{2}{m^2 n^2}.$$

Additionnons terme à terme cette égalité pour tous les $m, n \geqslant 1$. Pour $i \neq j$, le couple (i,j) est soit de la forme $(m+n,n)$, soit de la forme $(m,m+n)$, avec $m, n \geqslant 1$. Ainsi, dans la somme des termes du membre de gauche, tous les termes de la forme $f(i,j)$ avec $i \neq j$ s'éliminent et donc le membre de gauche se simplifie en :

$$\sum_{n \geqslant 1} f(n,n) \ = \ \sum_{n \geqslant 1} \frac{5}{n^4} \ = \ 5\zeta(4)$$

Pour le membre de droite on trouve :

$$\sum_{m,n\geqslant 1} \frac{2}{m^2 n^2} \;=\; 2\sum_{m\geqslant 1}\frac{1}{m^2}\cdot\sum_{n\geqslant 1}\frac{1}{n^2} \;=\; 2\zeta(2)^2,$$

si bien que l'on est conduit à l'égalité :

$$5\zeta(4) \;=\; 2\zeta(2)^2.$$

Connaissant $\zeta(2)=\frac{\pi^2}{6}$ on trouve donc $\zeta(4)=\frac{\pi^4}{90}$.
Une démonstration de ce résultat, à l'aide des nombres de Bernoulli, est présentée au chapitre 23.

(3) On sait depuis longtemps que, si s est un entier *pair* $s\geqslant 2$, $\zeta(s)$ est un multiple rationnel de π^s, et qu'il est par conséquent irrationnel (voir chapitre 23). En revanche, l'irrationalité de $\zeta(3)$ n'a été prouvée qu'en 1979 par Roger Apéry. En dépit d'efforts considérables, les connaissances sur $\zeta(s)$ sont bien incomplètes pour les autres entiers impairs $s=2t+1\geqslant 5$. Très récemment, Keith Ball et Tanguy Rivoal ont montré qu'une infinité de valeurs $\zeta(2t+1)$ sont irrationnelles. Mieux, même si l'on n'a pas réussi à établir le caractère irrationnel d'un $\zeta(s)$ particulier pour $s\geqslant 5$ impair, Wadim Zudilin a réussi à montrer qu'au moins l'un des quatre nombres $\zeta(5)$, $\zeta(7)$, $\zeta(9)$ ou $\zeta(11)$ est irrationnel. Nous renvoyons le lecteur au magnifique article de synthèse rédigé par Fischler [4].

(4) La localisation des zéros complexes de la fonction Zêta fait l'objet de « l'hypothèse de Riemann » : c'est l'une des conjectures non résolues parmi les plus célèbres et les plus importantes de toutes les mathématiques. Elle affirme que les zéros non triviaux $s\in\mathbb{C}$ de la fonction Zêta vérifient :

$$\mathrm{Re}(s) \;=\; \frac{1}{2}$$

(la fonction Zêta s'annule pour tous les entiers pairs négatifs, appelés *zéros triviaux* de la fonction Zêta).
Très récemment, Jeff Lagarias a montré que l'hypothèse de Riemann est curieusement équivalente à l'affirmation élémentaire suivante :

$$\text{pour tout } n\geqslant 1,\quad \sum_{d\,|\,n} d \;\leqslant\; H_n \;+\; \exp(H_n)\ln(H_n)$$

où H_n désigne toujours le n-ième nombre harmonique, l'égalité n'ayant lieu que si $n=1$.

Bibliographie

[1] K. BALL & T. RIVOAL : *Irrationalité d'une infinité de valeurs de la fonction zêta aux entiers impairs,* Inventiones math. **146** (2001), 193-207.

[2] F. BEUKERS, J. A. C. KOLK & E. CALABI : *Sums of generalized harmonic series and volumes,* Nieuw Archief voor Wiskunde (4) **11** (1993), 217-224.

[3] J. M. BORWEIN, P. B. BORWEIN & K. DILCHER : *Pi, Euler numbers, and asymptotic expansions,* Amer. Math. Monthly **96** (1989), 681-687.

[4] S. FISCHLER : *Irrationalité de valeurs de zêta (d'après Apéry, Rivoal, ...),* Bourbaki Seminar, No. 910, November 2002 ; Astérisque **294** (2004), 27-62.

[5] J. C. LAGARIAS : *An elementary problem equivalent to the Riemann hypothesis,* Amer. Math. Monthly **109** (2002), 534-543.

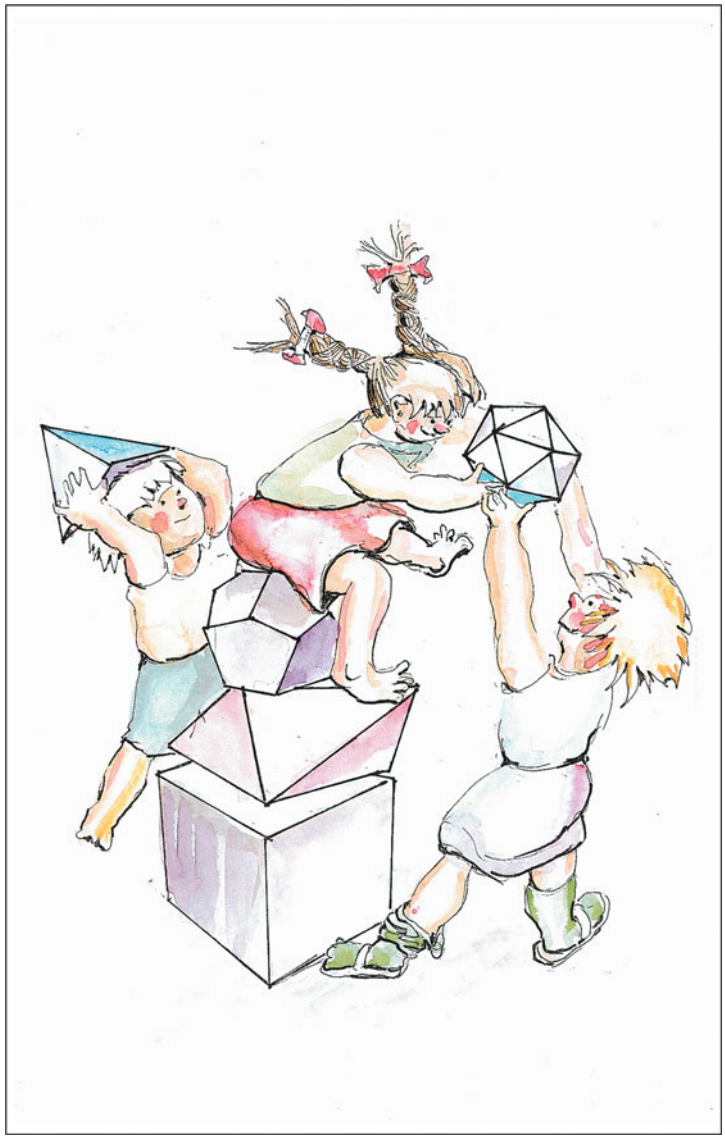
[6] W. J. LEVEQUE : *Topics in Number Theory,* Vol. I, Addison-Wesley, Reading MA 1956.

[7] A. M. YAGLOM & I. M. YAGLOM : *Challenging mathematical problems with elementary solutions,* Vol. II, Holden-Day, Inc., San Francisco, CA 1967.

[8] D. ZAGIER : *Values of zeta functions and their applications,* Proc. First European Congress of Mathematics, Vol. II (Paris 1992), Progress in Math. **120**, Birkhäuser, Basel 1994, pp. 497-512.

[9] W. ZUDILIN : *Arithmetic of linear forms involving odd zeta values,* J. Théorie Nombres Bordeaux **16** (2004), 251-291.

Géométrie



« Les solides platoniciens : un jeu d'enfant ! »

Le troisième problème de Hilbert : la décomposition des polyèdres

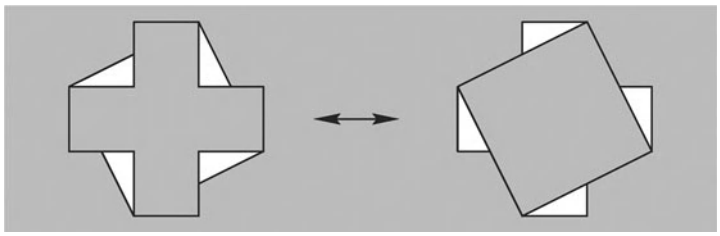
Chapitre 9

Lors de sa conférence légendaire au cours du Congrès international des mathématiciens à Paris en 1900, David Hilbert demanda — il s'agissait du troisième de ses vingt-trois problèmes — de déterminer :

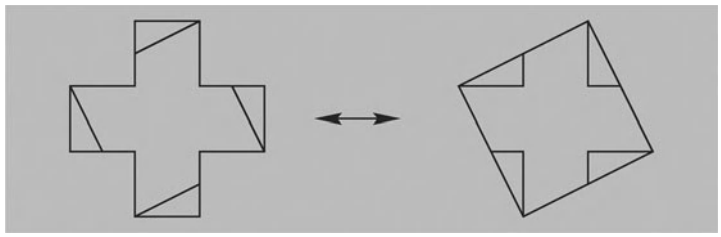
> *« deux tétraèdres de bases égales et de hauteurs égales qui ne puissent d'aucune manière être décomposés en tétraèdres congruents, et qui ne puissent pas être combinés avec des tétraèdres congruents pour former deux polyèdres qui seraient eux-mêmes décomposables en tétraèdres congruents. »*

David Hilbert

On trouve mention de ce problème dans deux lettres de Carl Friedrich Gauss datant de 1844 (publiées en 1900 dans le recueil des œuvres de Gauss). Si des tétraèdres de volume égal pouvaient être décomposés en parties congruentes, cela fournirait une preuve « élémentaire » du théorème XII.5 d'Euclide selon lequel des pyramides de même base et de même hauteur ont le même volume. On obtiendrait ainsi une définition élémentaire du volume des polyèdres (ne reposant pas sur des arguments issus de l'analyse et, en particulier, d'arguments de continuité).



La croix est équicomplémentaire à un carré de même aire. En considérant quatre fois le même triangle, on peut observer la construction qui prouve que la croix et le carré sont équidécomposables.



En fait, la croix et le carré sont même équidécomposables.

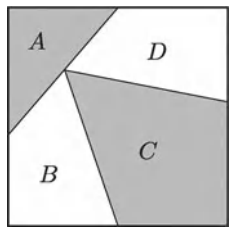
Il existe un énoncé similaire en géométrie plane : le théorème de Bolyai-Gerwien [1, Sect. 2.7] affirme que les polygones plans sont à la fois *équi-*

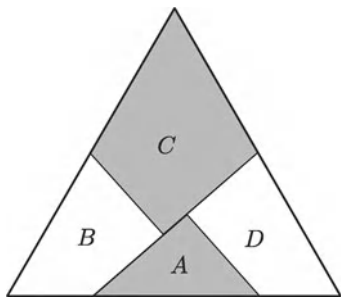
décomposables (on peut les découper en triangles congruents) et *équicomplémentaires* (on peut voir qu'ils sont congruents en considérant des triangles congruents), si et seulement s'ils ont la même aire. Hilbert, comme le montre sa formulation du problème, s'attendait à ce qu'il n'y ait pas de théorème analogue en dimension 3 et il avait raison. En fait, le problème fut complètement résolu par Max Dehn, un étudiant de Hilbert, dans deux publications : la première, exhibant des tétraèdres non-équidécomposables de base et hauteur égales, parut dès 1900, la seconde, traitant aussi le problème d'équicomplémentarité, parut en 1902. Cependant, les publications de Dehn ne sont pas faciles à comprendre et il fallut beaucoup d'efforts pour s'assurer que Dehn n'était pas tombé dans un piège subtil dans lequel d'autres étaient tombés : des preuves très élégantes mais fausses ont été proposées par Bricard (en 1896 !), par Meschkowski (en 1960) ; on peut penser que d'autres encore ont essayé en vain.

Néanmoins, la preuve de Dehn a été améliorée et vérifiée à plusieurs reprises. Grâce aux efforts conjugués de différents auteurs on en est arrivé à la « preuve classique » telle qu'elle est rédigée dans l'ouvrage de Boltianskii à propos du troisième problème de Hilbert ou telle qu'elle figure dans une précédente édition du présent ouvrage.

Dans la suite, nous allons tirer profit d'une simplification décisive trouvée par V. F. Kagan d'Odessa dès 1903. Sa démonstration repose sur ce que l'on présente ici sous le nom de « lemme du cône », qui lui-même conduit au « lemme des perles » (dont l'énoncé est présenté ici sous une forme moderne due à Benko). À l'aide de ces éléments, nous pouvons établir une preuve correcte et complète au « critère de Bricard » (énoncé par Bricard dans un article de 1896). En appliquant ce critère à différents exemples nous obtiendrons facilement la solution au troisième problème de Hilbert.

L'appendice qui clôt ce chapitre fournit quelques notions élémentaires sur les polyèdres.





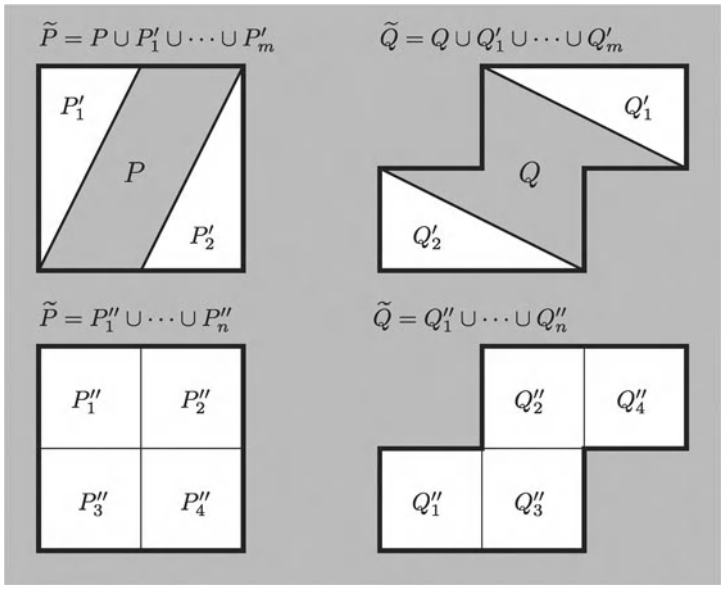
Cette équidécomposition d'un carré et d'un triangle équilatéral en quatre pièces a été proposée par Henry Dudeney (1902).
Le petit segment au milieu du triangle équilatéral est le support de l'intersection des pièces A et C mais ce n'est l'arête d'aucune de ces deux pièces.

Comme précédemment, on dit que deux polyèdres P et Q sont *équidécomposables* s'ils peuvent être respectivement décomposés en un nombre fini de polyèdres $P_1, \dots, P_n$ et $Q_1, \dots, Q_n$ tels que P_i et Q_i soient congruents pour tout i $(1 \leqslant i \leqslant n)$. Deux polyèdres sont *équicomplémentaires* s'il existe deux polyèdres équidécomposables $\widetilde{P} = P_1'' \cup \dots \cup P_n''$ et $\widetilde{Q} = Q_1'' \cup \dots \cup Q_n''$ qui admettent aussi des décompositions comportant P et Q de la forme $\widetilde{P} = P \cup P_1' \cup P_2' \cup \dots \cup P_m'$ et $\widetilde{Q} = Q \cup Q_1' \cup Q_2' \cup \dots \cup Q_m'$, où P_k' est congruent à Q_k' pour tout k. Un théorème de Gerling de 1844 [1, §12] a pour conséquence qu'il est indifférent de prendre en compte ou non les réflexions lorsqu'on considère les congruences.

Pour des polygones du plan, on définit l'équidécomposabilité et l'équicomplémentarité de manière analogue.

Il est clair que des polyèdres équidécomposables sont équicomplémentaires (c'est le cas $m = 0$) mais la réciproque est loin d'être évidente. Nous allons utiliser le « critère de Bricard » pour trouver, comme l'a proposé Hilbert, des tétraèdres de même volume qui ne sont pas équicomplémentaires et donc pas équidécomposables.

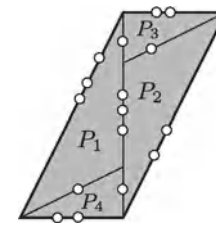
Avant de travailler véritablement avec des polyèdres tridimensionnels, commençons par établir le « lemme des perles » qui est intéressant aussi pour les décompositions planaires. Ce lemme utilise les *segments* d'une décomposition. Dans toute décomposition, les arêtes d'une pièce peuvent être subdivisées par des sommets ou des arêtes d'autres pièces. Nous appelons segments les intervalles d'une telle subdivision. Ainsi, dans le cas de la dimension 2, l'extrémité de tout segment est délimitée par un sommet. En dimension 3, l'extrémité d'un segment peut aussi être définie par l'intersection de deux arêtes. Néanmoins, dans tous les cas, tous les points intérieurs à un segment appartiennent au même ensemble d'arêtes des pièces considérées.

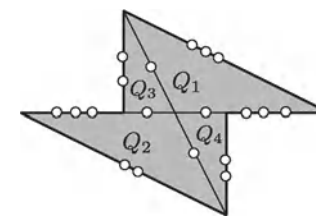


La figure ci-contre montre les quatre décompositions qui interviennent dans la définition pour un parallélogramme P et un hexagone non convexe Q qui sont équicomplémentaires.

Lemme des perles. *Si P et Q sont équidécomposables, alors on peut disposer des perles (c'est-à-dire affecter des nombres entiers positifs) sur tous les segments des décompositions $P = P_1 \cup \cdots \cup P_n$ et $Q = Q_1 \cup \cdots \cup Q_n$ de manière à ce que chaque arête d'une pièce P_k reçoive le même nombre de perles que l'arête qui lui correspond dans Q_k.*

Les polygones P et Q considérés dans la figure ci-dessus sont équidécomposables. La figure présentée ci-contre en fait l'illustration et montre une répartition possible pour les perles.





■ **Preuve.** Affectons une variable x_i (respectivement y_i) à chaque segment apparaissant dans la décomposition de P (respectivement de Q). Il nous faut trouver une valeur *entière* positive pour les variables x_i et y_j de sorte que les variables x_i correspondant à des segments de n'importe quelle arête d'un quelconque P_k produise la même somme que les variables y_j affectés aux segments de l'arête correspondante de Q_k. Cela conduit à des conditions qui signifient que la somme de certains x_i donne le même résultat que

la somme de certains y_j, soit encore que :

$$\sum_{i:s_i \subseteq e} x_i \;-\; \sum_{j:s'_j \subseteq e'} y_j \;=\; 0$$

où l'arête $e \subset P_k$ se décompose en les segments s_i, tandis que l'arête $e' \subset Q_k$ se décompose en les segments s'_j. Cette équation est linéaire et à coefficients entiers.

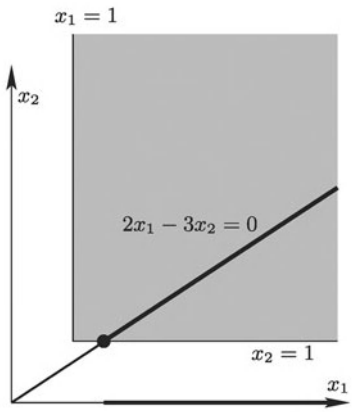
Nous remarquons néanmoins que certaines valeurs positives *réelles* satisfont à ces contraintes : ce sont les longueurs (réelles) des segments de la décomposition ! Le lemme qui suit permet de conclure. □

Lemme du cône. *Si un système d'équations linéaires homogènes à coefficients entiers admet une solution positive* ***réelle*** *alors il admet aussi une solution positive* ***entière****.*

■ **Preuve.** Le nom du lemme vient de ce que l'ensemble

$$C \;:=\; \{\mathbf{x} \in \mathbb{R}^N : A\mathbf{x} = \mathbf{0},\ \mathbf{x} > \mathbf{0}\}$$

où $A \in \mathbb{Z}^{M \times N}$ est une matrice à coefficients entiers est un cône rationnel (ouvert). Il faut montrer que si cet ensemble n'est pas vide alors il contient aussi des points entiers : $C \cap \mathbb{N}^N \neq \emptyset$. Si C n'est pas vide, alors il en va de même pour $\bar{C} := \{\mathbf{x} \in \mathbb{R}^N : A\mathbf{x} = \mathbf{0},\ \mathbf{x} \geqslant \mathbf{1}\}$ car pour tout vecteur positif, un multiple bien choisi de ce vecteur aura toutes ses coordonnées supérieures ou égales à 1 (on note ici $\mathbf{1}$ le vecteur dont toutes les coordonnées sont égales à 1). Il suffit de vérifier que $\bar{C} \subseteq C$ contient au moins un point à coordonnées *rationelles* car en multipliant ce vecteur par un dénominateur commun à ses coordonnées on obtiendra un point entier de $\bar{C} \subseteq C$.



Exemple : ici $\bar{C}$ est défini par $2x_1 - 3x_2 = 0$, $x_i \geqslant 1$. L'élimination de x_2 conduit à $x_1 \geqslant \frac{3}{2}$. La solution minimale au sens lexicographique du système est $(\frac{3}{2}, 1)$.

Il y a différente façons de prouver ce résultat. Nous allons suivre un sentier bien balisé qui a été emprunté pour la première fois par Fourier et Motzkin [8, Conférence 1]. En procédant à une *élimination de Fourier–Motzkin*, nous allons montrer que le système :

$$A\mathbf{x} = \mathbf{0},\ \mathbf{x} \geqslant \mathbf{1}$$

admet une plus petite solution au sens lexicographique et que cette solution est rationnelle si la matrice A est entière.

L'équation linéaire $\mathbf{a}^T\mathbf{x} = 0$ est équivalente aux deux inéquations $\mathbf{a}^T\mathbf{x} \geqslant 0$, $-\mathbf{a}^T\mathbf{x} \geqslant 0$ (on note $\mathbf{a}$ un vecteur colonne et $\mathbf{a}^T$ son transposé). Ainsi, il suffit de montrer que tout système de la forme :

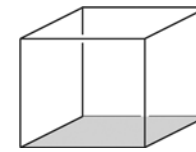
$$A\mathbf{x} \geqslant \mathbf{b},\ \mathbf{x} \geqslant \mathbf{1}$$

(où A et $\mathbf{b}$ sont à coefficients entiers) admet une plus petite solution (au sens lexicographique) qui est rationnelle pourvu que le système admette au moins une solution réelle.

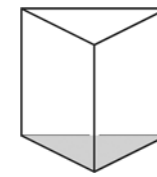
Nous allons raisonner par récurrence sur N. Le cas $N = 1$ est évident. Pour $N > 1$, on considère les inégalités dans lesquelles apparaît x_N. Si le vecteur $\mathbf{x}' = (x_1, \ldots, x_{N-1})$ est fixé, ces inégalités fournissent des minorants pour x_N (parmi lesquels $x_N \geqslant 1$) et éventuellement aussi des majorants. On forme alors un nouveau système $A'\mathbf{x}' \geqslant \mathbf{b}$, $\mathbf{x}' \geqslant \mathbf{1}$ à $N-1$ variables, comportant toutes les inégalités du système $A\mathbf{x} \geqslant \mathbf{b}$ dans lesquelles x_N n'apparaît pas, ainsi que toutes les inégalités obtenues en contraignant toutes les majorants de x_N (s'il y en a) sont supérieurs ou égaux à tous les minorants de x_N (dont $x_N \geqslant 1$). Ce système à $N-1$ inconnues a une solution et par récurrence il admet une solution minimale au sens lexicographique x'_* qui est rationnelle. On trouve facilement le plus petit x_N compatible avec cette solution x'_* ; il est déterminé par une équation ou une inéquation à coefficients entiers. x_N est donc rationnel lui aussi.

□

À présent nous nous concentrons sur la décomposition des polyèdres en dimension 3. Les angles dièdres, c'est-à-dire les angles entre deux faces adjacentes, jouent un rôle décisif dans le théorème suivant.



Dans un cube, tous les angles dièdres sont égaux à $\frac{\pi}{2}$.



Dans un prisme dont la base est un triangle équilatéral, on trouve des angles dièdres égaux à $\frac{\pi}{3}$ et à $\frac{\pi}{2}$.

Théorème. Critère de Bricard
On se place en dimension 3. Si les polyèdres P et Q d'angles dièdres respectifs $\alpha_1, \ldots, \alpha_r$ et $\beta_1, \ldots, \beta_s$ sont équidécomposables, alors il existe des entiers strictement positifs $m_1, \ldots, m_r$ et $n_1, \ldots, n_s$ et un entier k tels que :

$$m_1\alpha_1 + \ldots + m_r\alpha_r \;=\; n_1\beta_1 + \ldots + n_s\beta_s \;+\; k\pi.$$

De manière plus générale, la même condition est vérifiée si P et Q sont équicomplémentaires.

■ **Preuve.** Supposons pour commencer que P et Q sont équidécomposables avec les décompositions $P = P_1 \cup \cdots \cup P_n$ et $Q = Q_1 \cup \cdots \cup Q_n$, où P_i est congruent à Q_i. On affecte un nombre (positif) de perles à chaque segment dans les deux décompositions, selon les termes du lemme des perles.

Soit Σ_1 la somme de tous les angles dièdres en chaque endroit où se trouve une perle dans la décomposition de P. Si une arête d'une pièce composant P_i est affectée de plusieurs perles alors l'angle dièdres en cette arête est comptabilisé plusieurs fois dans la somme Σ_1.

Si une perle est contenue dans plusieurs pièces, alors plusieurs angles sont additionnés dans la somme pour cette perle mais ils sont tous mesurés dans le plan passant par la perle orthogonal au segment auquel elle correspond. Si un segment est contenu dans une arête de P, l'addition comptabilise l'angle dièdre intérieur α_j en cette arête. L'addition comptabilise l'angle π dans le cas où le segment se trouve sur un bord de P mais pas sur une arête. Si la perle/le segment se trouve à l'intérieur de P, alors la somme des angles dièdre comptabilise $2/pi$ ou π (le dernier cas se produit dans le cas où la perle se trouve à l'intérieur d'une face d'une pièce P_i).

Ainsi, on obtient :

$$\Sigma_1 \;=\; m_1\alpha_1 + \ldots + m_r\alpha_r \;+\; k_1\pi$$

pour des entiers strictement positifs m_j ($1 \leqslant j \leqslant r$) et un entier positif ou nul k_1. De même, pour la somme Σ_2 de tous les angles pour toutes les perles de la décomposition de Q on trouve :

$$\Sigma_2 \;=\; n_1\beta_1 + \ldots + n_s\beta_s \;+\; k_2\pi$$

pour des entiers strictement positifs n_j ($1 \leqslant j \leqslant s$) et un entier strictement positif k_2.

Toutefois, on peut aussi obtenir les sommes Σ_1 et Σ_2 en comptabilisant les contributions individuelles de chaque pièce P_i et Q_i. Comme P_i et Q_i sont congruents, ils présentent les mêmes angles dièdres sur les arêtes qui se correspondent ; le lemme des perles assure alors que l'on dispose du même nombre de perles sur les arêtes qui se correspondent dans chacune des décompositions considérées de P et de Q. On obtient donc $\Sigma_1 = \Sigma_2$, ce qui constitue le critère de Bricard (avec $k = k_2 - k_1 \in \mathbb{Z}$) pour le cas de l'équidécomposabilité.

Considérons à présent le cas où P et Q sont équicomplémentaires, c'est-à-dire supposons que nous disposons des décompositions suivantes :

$$\widetilde{P} \;=\; P \cup P'_1 \cup \cdots \cup P'_m \qquad \text{and} \qquad \widetilde{Q} \;=\; Q \cup Q'_1 \cup \cdots \cup Q'_m,$$

où P'_i et Q'_i sont congruents et où $\widetilde{P}$ et $\widetilde{Q}$ sont équidécomposables en :

$$\widetilde{P} \;=\; P''_1 \cup \cdots \cup P''_n \qquad \text{et} \qquad \widetilde{Q} \;=\; Q''_1 \cup \cdots \cup Q''_n,$$

où les P''_i et les Q''_i sont congruents (comme montré sur la figure de la page 63). En appliquant de nouveau le lemme des perles, nous plaçons des perles sur chacun des segments dans les quatre décompositions et on impose une condition supplémentaire : chaque arête de $\widetilde{P}$ (respectivement de $\widetilde{Q}$) reçoit le même nombre total de perles dans les deux décompositions (la démonstration du lemme des perles fondée sur le lemme du cône nous autorise à requérir de telles conditions supplémentaires). Nous calculons aussi la somme des angles en les perles Σ'_1 et Σ'_2 comme Σ''_1 et Σ''_2.

Les sommes d'angles Σ''_1 et Σ''_2 correspondent à des décompositions de polyèdres différents, $\widetilde{P}$ et $\widetilde{Q}$, en *le même ensemble de pièces* ; on trouve donc $\Sigma''_1 = \Sigma''_2$ comme précédemment.

Les sommes d'angles Σ'_1 et Σ''_1 correspondent à des décompositions différentes *du même polyèdre*, $\widetilde{P}$. Comme nous avons placé le même nombre de perles sur les arêtes dans les deux décompositions, l'argument précédent conduit à : $\Sigma'_1 = \Sigma''_1 + \ell_1\pi$ pour un entier $\ell_1 \in \mathbb{Z}$. De manière analogue, on trouve : $\Sigma'_2 = \Sigma''_2 + \ell_2\pi$ pour un entier $\ell_2 \in \mathbb{Z}$. On en conclut que :

$$\Sigma'_2 \;=\; \Sigma'_1 + \ell\pi \quad \text{for } \ell = \ell_2 - \ell_1 \in \mathbb{Z}.$$

Toutefois, Σ'_1 et Σ'_2 correspondent à des décompositions de $\widetilde{P}$ et $\widetilde{Q}$ respectivement en les mêmes pièces, *mais* la première décomposition utilise P

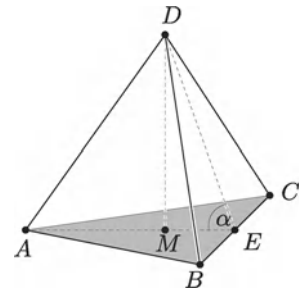
en tant que pièce, alors que la deuxième utilise Q. Ainsi en retranchant les contributions des P_i' et des Q_i' dans chaque membre on obtient la conclusion désirée : les contributions de P et Q aux sommes d'angles respectives,

$$m_1\alpha_1 + \ldots + m_r\alpha_r \quad \text{et} \quad n_1\beta_1 + \ldots + n_s\beta_s,$$

(où m_j dénombre les perles sur les arêtes d'angle dièdre α_j dans P et n_j dénombre les perles sur les arêtes d'angle dièdre β_j dans Q) diffèrent d'un multiple entier de π, en l'occurrence de $\ell\pi$. □

Le critère de Bricard nous permet de donner une solution complète au troisième problème de Hilbert : il suffit de calculer les angles dièdres sur quelques exemples.

Exemple 1. Étant donné un tétraèdre régulier T_0 de longueur d'arête ℓ, nous calculons l'angle dièdre à l'aide de la figure ci-contre. Le centre M du triangle de base divise la hauteur AE du triangle de base dans un rapport $1 : 2$ et, puisque $|AE| = |DE|$, nous trouvons $\cos\alpha = \frac{1}{3}$; ainsi



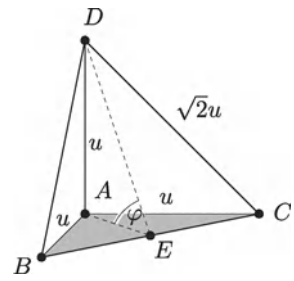
$$\alpha \;=\; \arccos\tfrac{1}{3}$$

Ainsi, *un tétraèdre régulier et un cube ne peuvent pas être équidécomposables ou équicomplémentaires* En effet, tous les angles dièdres dans un cube sont égaux à $\frac{\pi}{2}$ et la condition de Bricard impose que :

$$m_1 \arccos\tfrac{1}{3} \;=\; n_1\tfrac{\pi}{2} + k\pi$$

pour certains entiers strictement positifs m_1 et n_1 et un entier k. Or ceci est impossible car on sait depuis le chapitre 7 (prendre $n = 9$ dans le théorème 3) que $\frac{1}{\pi}\arccos\frac{1}{3}$ est irrationnel.

Exemple 2. Soit T_1 un tétraèdre engendré par trois arêtes orthogonales AB, AC, AD de longueur u. Trois angles dièdres de ce tétraèdre sont droits et trois autres sont égaux à un angle φ, que nous calculons à l'aide de la figure ci-contre :



$$\cos\varphi \;=\; \frac{|AE|}{|DE|} \;=\; \frac{\frac{1}{2}\sqrt{2}u}{\frac{1}{2}\sqrt{3}\,\sqrt{2}u} \;=\; \frac{1}{\sqrt{3}}$$

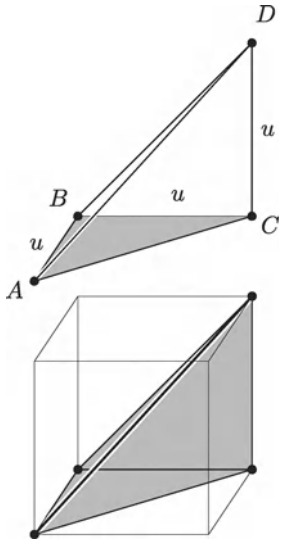
et donc :

$$\varphi \;=\; \arccos\frac{1}{\sqrt{3}}\,.$$

Ainsi les seuls angles dièdres que l'on puisse trouver dans T_1 sont π, $\frac{\pi}{2}$ et $\arccos\frac{1}{\sqrt{3}}$. Selon le critère de Bricard, ce tétraèdre et le cube de même volume ne sont ni équidécomposables, ni équicomplémentaires, cette fois parce que

$$\tfrac{1}{\pi}\arccos\tfrac{1}{\sqrt{3}}$$

est irrationnel comme on l'a démontré au chapitre 7 (prendre $n = 3$ dans le théorème 3).



Exemple 3. Soit T_2 un tétraèdre ayant trois arêtes consécutives AB, BC et CD mutuellement orthogonales et de même longueur u (ce volume est appelé un *orthoschème*). Il est facile de calculer les angles dièdres d'un tel tétraèdre (trois d'entre eux sont égaux à $\frac{\pi}{2}$, deux sont égaux à $\frac{\pi}{4}$ et le dernier est égal à $\frac{\pi}{6}$) en utilisant le fait que le cube de côté u peut être décomposé en six tétraèdres de ce type (trois copies congruentes et trois copies symétriques). Ainsi tous les angles dièdres dans T_2 sont des multiples rationnels de π et en raisonnant comme précédemment (en particulier en tenant compte des résultats d'irrationalité établis au chapitre 7) le critère de Bricard implique que T_2 et T_0 (ou T_1) ne sont ni équidécomposables, ni équicomplémentaires.

Cela résout le troisième problème de Hilbert puisque T_1 et T_2 ont des bases congruentes et la même hauteur.

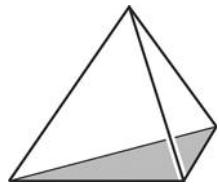
Appendice - Polytopes et polyèdres

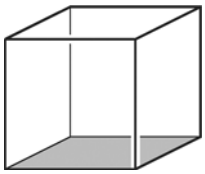
Un *polytope convexe* de $\mathbb{R}^d$ est l'enveloppe convexe d'un ensemble fini de points $S = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, c'est-à-dire, l'ensemble :

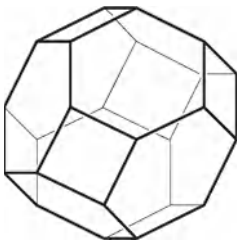
$$P \;=\; \text{conv}(S) \;:=\; \Big\{ \sum_{i=1}^{n} \lambda_i \boldsymbol{s}_i : \lambda_i \geqslant 0,\; \sum_{i=1}^{n} \lambda_i = 1 \Big\}$$

Les polytopes sont évidemment des objets familiers : les premiers exemples sont les *polygones* convexes (polytopes convexes de dimension 2) et les *polyèdres* convexes (polytopes convexes de dimension 3).

Il y a plusieurs types de polyèdres qui généralisent naturellement ce concept en dimension plus grande. Par exemple, si l'ensemble S est affinement libre et de cardinal $d+1$, alors $\text{conv}(S)$ est un *simplexe* de dimension d (ou un *d-simplexe*). Si $d = 2$, c'est un triangle, si $d = 3$, un tétraèdre. De même, les carrés et les cubes sont des cas particuliers de d-cubes, comme le *d-cube unité* défini par $C_d = [0,1]^d \subseteq \mathbb{R}^d$.







Quelques polytopes familiers : tétraèdre, cube et permutaèdre.

Les polytopes généraux sont définis comme des réunions finies de polytopes convexes. Dans ce livre, les polyèdres non-convexes apparaîtront dans le contexte du théorème de rigidité de Cauchy au chapitre 13, les polygones non-convexes dans le contexte du théorème de Pick au chapitre 12 et encore lorsque nous travaillerons sur le théorème de la galerie d'art au chapitre 35.

Les polytopes convexes peuvent être définis de façon équivalente comme les ensembles de solutions bornées de systèmes finis d'inégalités linéaires. Ainsi, tout polytope convexe $P \subseteq \mathbb{R}^d$ a une représentation de la forme :

$$P \;=\; \{\boldsymbol{x} \in \mathbb{R}^d : A\boldsymbol{x} \leqslant \boldsymbol{b}\}$$

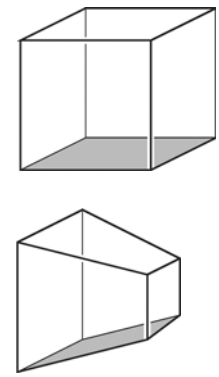
où $A \in \mathbb{R}^{m \times d}$ est une matrice et $\boldsymbol{b} \in \mathbb{R}^m$ un vecteur. En d'autres termes, P est l'ensemble des solutions d'un système de m inégalités linéaires $\boldsymbol{a}_i^T \boldsymbol{x} \leqslant b_i$, où $\boldsymbol{a}_i^T$ est la i-ième ligne de A. Réciproquement, un tel ensemble borné de solutions est toujours un polytope convexe et peut donc être représenté comme l'enveloppe convexe d'un ensemble fini de points.

Pour les polygones et les polyèdres, on a les concepts familiers de *sommets*, *arêtes* et 2-*faces*. En ce qui concerne les polytopes convexes de plus grande dimension, on peut définir leurs faces comme suit : une *face* de P est un sous-ensemble $F \subseteq P$ de la forme $P \cap \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{a}^T \boldsymbol{x} = b\}$, où $\boldsymbol{a}^T \boldsymbol{x} \leqslant b$ est une inégalité linéaire réalisée en tout point $\boldsymbol{x} \in P$.

Toutes les faces d'un polytope sont elles-mêmes des polytopes. L'ensemble V des sommets (faces de dimension 0) d'un polytope convexe est aussi l'ensemble minimal pour l'inclusion satisfaisant $\mathrm{conv}(V) = P$. Si l'on suppose que $P \subseteq \mathbb{R}^d$ est un polytope convexe de dimension d, les *facettes* (les faces de dimension $(d-1)$) déterminent un ensemble minimal d'hyperplans, donc de demi-espaces contenant P, dont l'intersection est P. En particulier, cela implique le fait suivant, dont nous aurons besoin plus tard : soit F une facette de P ; notons H_F l'hyperplan qu'elle détermine, H_F^+ et H_F^- les deux demi-espaces bornés par H_F. Alors l'un de ces deux demi-espaces contient P (et l'autre non).

Le *graphe* $G(P)$ d'un polytope convexe P est constitué de l'ensemble V des sommets et de l'ensemble E des arêtes (les faces de dimension 1). Si la dimension de P est 3, alors ce graphe est planaire et donne naissance à la fameuse « formule d'Euler des polyèdres » (voir chapitre 12).

Deux polytopes $P, P' \subseteq \mathbb{R}^d$ sont *congruents* s'il existe une application affine préservant les longueurs, qui envoie P sur P'. Une telle application peut inverser l'orientation de l'espace, comme le fait une réflexion relativement à un hyperplan, qui envoie P sur un *symétrique* de P. Ils sont *combinatoirement équivalents* s'il existe une bijection qui envoie les faces de P sur les faces de P' et qui conserve la dimension et les inclusions entre les faces. Cette notion d'équivalence combinatoire est beaucoup plus faible que la congruence : par exemple, notre figure montre un cube unité et un cube déformé qui sont combinatoirement équivalents (nous pourrions appeler l'un ou l'autre « un cube ») mais certainement pas congruents.



Polytopes combinatoirement équivalents

On dit qu'un polytope (ou un sous-ensemble plus général de $\mathbb{R}^d$) a une *symétrie centrale* s'il existe un point $\boldsymbol{x}_0 \in \mathbb{R}^d$ tel que :

$$\boldsymbol{x}_0 + \boldsymbol{x} \in P \iff \boldsymbol{x}_0 - \boldsymbol{x} \in P$$

Dans ce cas, $\boldsymbol{x}_0$ s'appelle le *centre* de P.

Bibliographie

[1] V. G. BOLTIANSKII : *Hilbert's Third Problem,* V. H. Winston & Sons (Halsted Press, John Wiley & Sons), Washington DC 1978.

[2] D. BENKO : *A new approach to Hilbert's third problem,* Amer. Math. Monthly, **114** (2007), 665-676.

[3] M. DEHN : *Ueber raumgleiche Polyeder,* Nachrichten von der Königl. Gesellschaft der Wissenschaften, Mathematisch-physikalische Klasse (1900), 345-354.

[4] M. DEHN : *Ueber den Rauminhalt,* Mathematische Annalen **55** (1902), 465-478.

[5] C. F. GAUSS : *"Congruenz und Symmetrie" : Briefwechsel mit Gerling,* pp. 240-249 in : Werke, Band VIII, Königl. Gesellschaft der Wissenschaften zu Göttingen ; B. G. Teubner, Leipzig 1900.

[6] D. HILBERT : *Mathematical Problems,* Lecture delivered at the International Congress of Mathematicians at Paris in 1900, Bulletin Amer. Math. Soc. **8** (1902), 437-479.

[7] B. KAGAN : *Über die Transformation der Polyeder,* Mathematische Annalen **57** (1903), 421-424.

[8] G. M. ZIEGLER : *Lectures on Polytopes,* Graduate Texts in Mathematics **152**, Springer, New York 1995/1998.

Droites du plan et décompositions de graphes

Chapitre 10

Peut-être le plus connu des problèmes sur les configurations de droites fut-il soulevé par Sylvester, en 1893, dans un recueil de problèmes mathématiques :

QUESTIONS FOR SOLUTION.

11851. (Professor SYLVESTER.)—Prove that it is not possible to arrange any finite number of real points so that a right line through every two of them shall pass through a third, unless they all lie in the same right line.

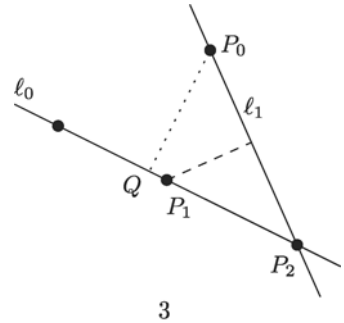
Montrer qu'il n'est pas possible de disposer un ensemble fini de points du plan de sorte que les droites définies par les différents couples de points passent aussi par un troisième point sans que tous les points de cet ensemble soient alignés.

James J. Sylvester

On ne sait pas si Sylvester lui-même avait établi une preuve. Une démonstration correcte fut donnée par Tibor Gallai [Grünwald] quelques quarante années plus tard. C'est pourquoi on attribue communément le théorème suivant à Sylvester et Gallai. À la suite de la preuve de Gallai, plusieurs autres démonstrations apparurent mais on peut considérer que l'argument suivant, dû à L. M. Kelly, est tout simplement le meilleur.

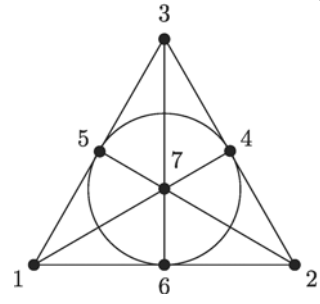
Théorème 1. *Pour toute configuration de n points non alignés du plan, il existe une droite qui contient exactement deux de ces points.*

■ **Preuve.** Soit $\mathcal{P}$ l'ensemble des points. Considérons l'ensemble $\mathcal{L}$ de toutes les droites qui passent par deux points au moins de $\mathcal{P}$. Parmi toutes les paires $(P, \ell) \in \mathcal{L} \times \mathcal{P}$ telles que P ne soit pas sur ℓ, choisissons une paire (P_0, ℓ_0) telle que P_0 soit le point le plus proche de ℓ_0 ; soit Q la projection orthogonale de P_0 sur ℓ_0.



Proposition. *Cette droite ℓ_0 répond à la question !*

Sinon ℓ_0 contiendrait au moins trois points de $\mathcal{P}$; deux d'entre eux, appelons-les P_1 et P_2, seraient du même côté de Q. Supposons que P_1 se trouve entre Q et P_2, (P_1 coïncidant éventuellement avec Q). La figure de droite montre cette configuration. Alors la distance de P_1 à la droite ℓ_1 déterminée par P_0 et P_2 est inférieure à la distance de P_0 à ℓ_0, ce qui contredit notre choix de ℓ_0 et P_0. □



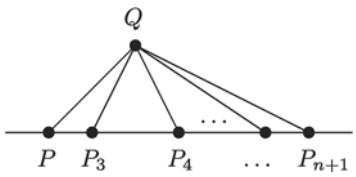
Dans la preuve, nous avons utilisé des axiomes métriques (plus petite distance) et des axiomes d'ordre (P_1 se trouve entre Q et P_2) du plan réel. Avons-nous réellement besoin de ces propriétés en plus des axiomes habituels d'incidence de points et droites ? En fait, une condition supplémentaire est requise, comme le montre le célèbre plan de Fano représenté dans

la marge. Ici, $\mathcal{P} = \{1, 2, \ldots, 7\}$ et $\mathcal{L}$ se compose de 7 droites constituées de trois points comme le montre la figure, la « droite » $\{4, 5, 6\}$ comprise. Tout couple de points détermine une unique droit donc les axiomes d'incidence sont satisfaits. Cependant il n'y a pas de droite engendrée par 2 points. Ainsi, le théorème de Sylvester-Gallai montre que la configuration de Fano ne peut être plongée dans le plan réel de sorte que les sept triplets de droites colinéaires appartiennent à des droites réelles : il doit toujours y avoir une droite « tordue ».

Cependant, Coxeter a montré que les axiomes d'ordre suffisent à démontrer le théorème de Sylvester-Gallai. On peut donc imaginer une preuve qui n'utilise aucune propriété métrique (voir aussi la preuve utilisant la formule d'Euler donnée au chapitre 12).

Le théorème de Sylvester-Gallai implique directement un autre résultat célèbre sur les points et droites du plan, dû à Paul Erdős et Nicolaas G. de Bruijn. On a en fait un résultat plus général avec des systèmes arbitraires de points-droites, comme l'avaient déjà observé Erdős et de Bruijn. Nous allons étudier ce résultat plus général ultérieurement.

Théorème 2. *Soit $\mathcal{P}$ un ensemble de $n \geqslant 3$ points non alignés du plan. Alors l'ensemble $\mathcal{L}$ des droites passant au moins par deux points contient au moins n droites.*



■ **Preuve.** Si $n = 3$, il n'y a rien à montrer. Nous allons maintenant procéder par récurrence sur n. Soit $|\mathcal{P}| = n + 1$. En utilisant le théorème précédent, il existe une droite $\ell_0 \in \mathcal{L}$ contenant exactement deux points P et Q de $\mathcal{P}$. Considérons l'ensemble $\mathcal{P}' = \mathcal{P} \backslash Q$ et l'ensemble $\mathcal{L}'$ des droites déterminées par $\mathcal{P}'$. Si les points de $\mathcal{P}'$ ne sont pas alignés, alors un raisonnement par récurrence montre que $|\mathcal{L}'| \geqslant n$ et donc $|\mathcal{L}| \geqslant n + 1$ à cause de la droite supplémentaire ℓ_0 de $\mathcal{L}$. Si, d'autre part, les points de $\mathcal{P}'$ sont alignés, alors nous obtenons le faisceau qui se compose d'exactement $n+1$ droites. □

Voici, comme promis, le résultat qui s'applique à des « géométries d'incidence » beaucoup plus générales.

Théorème 3. *Soit X un ensemble de $n \geqslant 3$ éléments et soient $A_1, \ldots, A_m$ des sous-ensembles propres de X tels que chaque paire d'éléments de X soit contenue dans un ensemble A_i exactement. Alors $m \geqslant n$.*

■ **Preuve.** La preuve suivante, attribuée selon les cas à Motzkin ou à Conway, est tout à fait admirable et tient presque en une ligne. Si $x \in X$, soit r_x le nombre de sous-ensembles A_i contenant x. (Notons que $2 \leqslant r_x < m$ par hypothèse.) Si $x \notin A_i$, alors $r_x \geqslant |A_i|$ parce que les ensembles $|A_i|$ contenant x et un élément de A_i doivent être distincts. Si $m < n$, alors $m|A_i| < n\, r_x$ donc $m(n - |A_i|) > n(m - r_x)$ si $x \notin A_i$; par suite :

$$1 = \sum_{x \in X} \frac{1}{n} = \sum_{x \in X} \sum_{A_i : x \notin A_i} \frac{1}{n(m - r_x)} > \sum_{A_i} \sum_{x : x \notin A_i} \frac{1}{m(n - |A_i|)} = \sum_{A_i} \frac{1}{m} = 1$$

ce qui est absurde. □

Il existe une autre preuve très courte de ce théorème qui utilise l'algèbre linéaire. Soit B la *matrice d'incidence* de $(X; A_1, \ldots, A_m)$, en d'autres termes les lignes de B sont indexées par les éléments de X, les colonnes par $A_1, \ldots, A_m$, avec :

$$B_{xA} := \begin{cases} 1 & \text{si} \quad x \in A \\ 0 & \text{si} \quad x \notin A \end{cases}$$

Considérons le produit BB^T. Si $x \neq x'$ nous avons : $(BB^T)_{xx'} = 1$, puisque x et x' sont contenus dans un seul ensemble A_i. Ainsi :

$$BB^T = \begin{pmatrix} r_{x_1}-1 & 0 & \ldots & 0 \\ 0 & r_{x_2}-1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & r_{x_n}-1 \end{pmatrix} + \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & & \vdots \\ \vdots & & \ddots & 1 \\ 1 & \ldots & 1 & 1 \end{pmatrix}$$

où r_x est défini comme indiqué ci-dessus. Puisque la première matrice est définie positive (elle admet des valeurs propres toutes strictement positives) et la deuxième matrice est semi-définie positive (elle admet les valeurs propres n et 0), BB^T est définie positive et donc en particulier inversible ce qui implique que $\text{rang}(BB^T) = n$. On en déduit que le rang de la $(n \times m)$-matrice B est au moins n et on en conclut que $n \leqslant m$, puisque le rang ne peut pas excéder le nombre de colonnes.

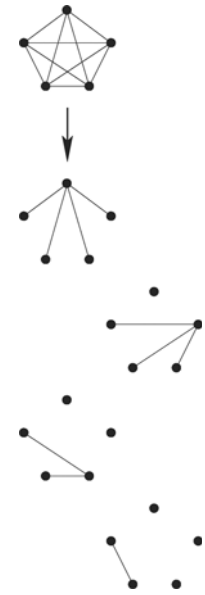
Allons un peu plus loin et tournons-nous vers la théorie des graphes (nous renvoyons à l'appendice de ce chapitre pour un rappel des notions élémentaires sur les graphes). Un moment de réflexion montre que l'énoncé suivant est en fait identique au théorème 3 :

> *Si l'on décompose un graphe complet K_n en m cliques*[1] *distinctes de K_n, telles que chaque arête appartienne à une unique clique, alors $m \geqslant n$.*

En effet, si nous faisons correspondre à X l'ensemble des sommets de K_n et aux ensembles A_i les ensembles des sommets des cliques, alors les énoncés sont identiques. Notre tâche suivante est de décomposer K_n en graphes bipartis complets tels qu'à nouveau, chaque arête se trouve exactement dans un de ces graphes. Il y a une façon simple de s'y prendre. Numérotons les sommets $\{1, 2, \ldots, n\}$. Prenons d'abord le graphe biparti complet joignant 1 à tous les autres sommets. Nous obtenons ainsi le graphe $K_{1,n-1}$ appelé *étoile*. Ensuite, joignons 2 à $3, \ldots, n$, ce qui donne une étoile $K_{1,n-2}$. En continuant, nous décomposons K_n en étoiles $K_{1,n-1}, K_{1,n-2}, \ldots, K_{1,1}$. Cette décomposition utilise $n-1$ graphes bipartis complets. Peut-on faire mieux, c'est-à-dire utiliser moins de graphes ? Non, comme l'affirme le résultat de Ron Graham et Henry O. Pollak :

Théorème 4. *Si K_n se décompose en sous-graphes bipartis complets H_1, $H_2, \ldots, H_m$, alors $m \geqslant n-1$.*

1. N.D.T. : se reporter à l'appendice en fin de chapitre pour une définition de ce terme.



Une décomposition de K_5 en 4 sous-graphes bipartis complets.

Contrairement au théorème d'Erdős-de Bruijn, aucune preuve combinatoire de ce résultat n'est connue ! Toutes les démonstrations utilisent l'algèbre linéaire d'une manière ou d'une autre. Parmi diverses idées plus ou moins équivalentes, examinons la preuve de Tverberg, qui semble être la plus transparente.

■ **Preuve.** Soit $\{1, \ldots, n\}$ l'ensemble des sommets de K_n et soit L_j, R_j les ensembles de sommets du graphe biparti complet H_j, $j = 1, \ldots, m$. À chaque sommet i nous associons une variable x_i. Puisque $H_1, \ldots, H_m$ décompose K_n, nous avons :

$$\sum_{i<j} x_i x_j = \sum_{k=1}^{m} \Big(\sum_{a \in L_k} x_a \cdot \sum_{b \in R_k} x_b \Big) \tag{1}$$

Supposons maintenant le théorème faux, c'est-à-dire que $m < n-1$. Alors, le système d'équations linéaires :

$$\begin{aligned} x_1 + \ldots + x_n &= 0 \\ \sum_{a \in L_k} x_a &= 0 \qquad (k = 1, \ldots, m) \end{aligned}$$

a moins d'équations que de variables et admet une solution non triviale $c_1, \ldots, c_n$. De (1) on déduit :

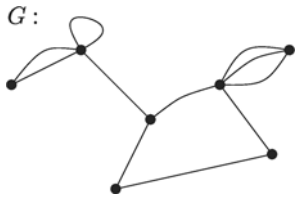
$$\sum_{i<j} c_i c_j = 0$$

Mais cela implique :

$$0 = (c_1 + \ldots + c_n)^2 = \sum_{i=1}^{n} c_i^2 + 2 \sum_{i<j} c_i c_j = \sum_{i=1}^{n} c_i^2 > 0$$

ce qui est contradictoire. □

Appendice - Notions élémentaires sur les graphes

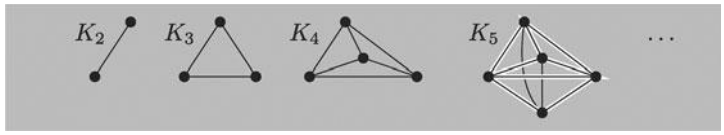


Un graphe G avec 7 sommets et 11 arêtes. Il possède une boucle, une arête double et une arête triple.

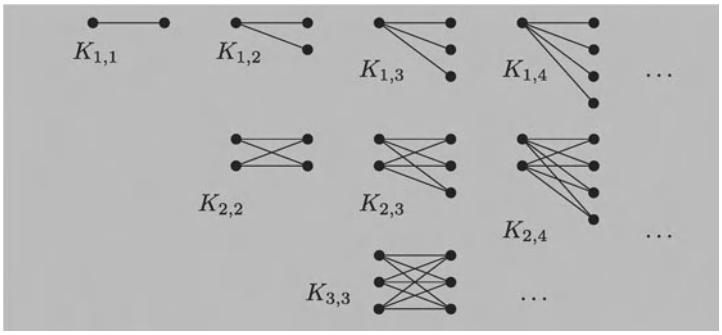
Les graphes font partie des structures mathématiques les plus fondamentales. Il en existe plusieurs versions, représentations et incarnations différentes. De façon abstraite, un *graphe* est un couple $G = (V, E)$, où V est l'ensemble des *sommets*, E est l'ensemble des *arêtes*, chaque arête $e \in E$ « reliant » deux sommets $v, w \in V$. Nous considérons seulement les graphes finis, ceux pour lesquels V et E sont finis.

Habituellement, nous travaillons avec des *graphes simples* : nous excluons donc les *boucles*, c'est-à-dire les arêtes dont les deux extrémités coïncident, et les *arêtes multiples* qui partagent les mêmes extrémités. Les sommets d'un graphe sont dits *adjacents* ou *voisins* si ce sont les extrémités d'une arête. Un sommet et une arête sont dits *incidents* si l'arête admet le sommet en question pour extrémité.

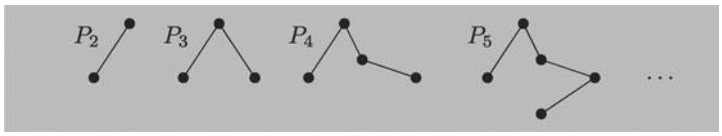
Voici une petite galerie de graphes (simples) importants :



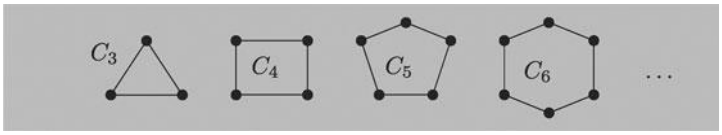
Les *graphes complets* K_n à n sommets et $\binom{n}{2}$ arêtes.



Les *graphe bipartis complets* $K_{m,n}$ à $m+n$ sommets et mn arêtes.



Les *chemins* P_n à n sommets.



Les *cycles* C_n à n sommets.

Deux graphes $G = (V, E)$ et $G' = (V', E')$ sont *isomorphes* s'il existe des bijections $V \to V'$ et $E \to E'$ qui conservent les incidences entre les arêtes et leurs extrémités. Un problème majeur non résolu consiste à savoir s'il existe un test efficace qui permette de décider si deux graphes donnés sont isomorphes. Cette notion d'isomorphisme nous permet de parler *du* graphe complet K_5 à 5 sommets, etc.

$G' = (V', E')$ est un *sous-graphe* de $G = (V, E)$ si $V' \subseteq V, E' \subseteq E$ et si chaque arête $e \in E'$ a les mêmes extrémités dans G' et dans G. G' est un *sous-graphe induit* si, en outre, *toutes* les arêtes de G qui relient les sommets de G' sont aussi des arêtes de G'.

est un sous-graphe de :

De nombreuses notions sur les graphes sont tout à fait intuitives : par exemple, un graphe G est *connexe* si tout couple de sommets est relié par une chaîne de G (c'est-à-dire une suite finie d'arêtes consécutives) ou, de façon équivalente, si G ne peut pas se décomposer en deux sous-graphes non-vides dont les ensembles de sommets sont disjoints. Tout graphe est la réunion disjointe de ses *composantes connexes*.

Nous terminons ce survol des notions élémentaires sur les graphes par quelques autres éléments de terminologie. Une *clique* dans G est un sous-

graphe complet. Un *ensemble indépendant* dans G est un sous-graphe induit sans arêtes, c'est-à-dire un sous-ensemble de l'ensemble des sommets tel qu'aucun couple de sommets ne soit relié par une arête de G. Un graphe est une *forêt* s'il ne contient aucun cycle. Un *arbre* est une forêt connexe. Finalement, un graphe $G = (V, E)$ est un *graphe biparti* s'il est isomorphe à un sous-graphe d'un graphe biparti complet, c'est-à-dire si l'ensemble des sommets peut être décrit comme la réunion $V = V_1 \cup V_2$ de deux ensembles indépendants.

Bibliographie

[1] N. G. DE BRUIJN & P. ERDŐS : *On a combinatorial problem,* Proc. Kon. Ned. Akad. Wetensch. **51** (1948), 1277-1279.

[2] H. S. M. COXETER : *A problem of collinear points,* Amer. Math. Monthly **55** (1948), 26-28 (contient la preuve de Kelly).

[3] P. ERDŐS : *Problem 4065 — Three point collinearity,* Amer. Math. Monthly **51** (1944), 169-171 (contient la preuve de Gallai).

[4] R. L. GRAHAM & H. O. POLLAK : *On the addressing problem for loop switching,* Bell System Tech. J. **50** (1971), 2495-2519.

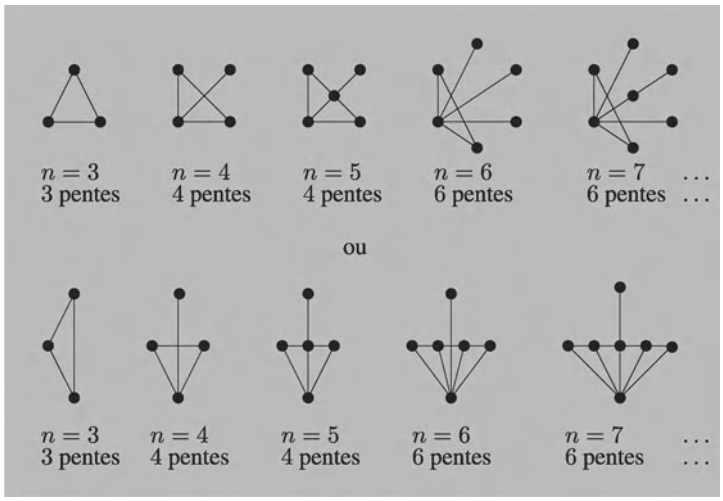
[5] J. J. SYLVESTER : *Mathematical Question 11851,* The Educational Times **46** (1893), 156.

[6] H. TVERBERG : *On the decomposition of K_n into complete bipartite graphs,* J. Graph Theory **6** (1982), 493-494.

Le problème des pentes

Chapitre 11

Nous suggérons au lecteur, avant de poursuivre la lecture de ce chapitre, d'essayer de construire des configurations de points du plan qui déterminent « relativement peu » de pentes. À cet effet, on suppose bien sûr que les $n \geqslant 3$ points ne sont pas alignés. Rappelons le théorème d'Erdős et de Bruijn, (voir chapitre 10) : les n points vont déterminer au moins n droites différentes. Cependant, beaucoup de ces droites peuvent être parallèles et déterminent donc la même pente.

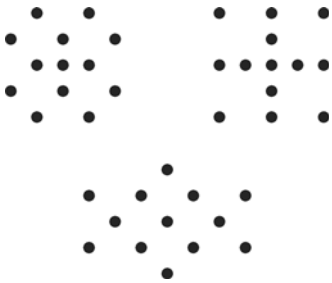


Une rapide expérimentation pour une petite valeur de n conduira probablement le lecteur à une suite semblable aux deux suites représentées ici.

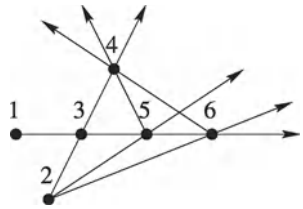
Après quelques essais pour trouver des configurations ayant un nombre de pentes plus faible, on peut conjecturer, comme Scott en 1970, le théorème suivant :

> **Théorème.** *Si $n \geqslant 3$ points du plan ne sont pas alignés, alors ils déterminent au moins $n - 1$ pentes différentes, l'égalité étant réalisée seulement dans le cas où n est impair et $n \geqslant 5$.*

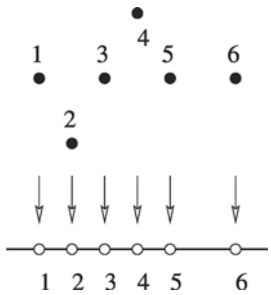
Les figures ci-dessus représentent quelques-unes des premières configurations appartenant à deux suites infinies d'exemples et montrent que cet énoncé est le *meilleur possible* : pour tout entier impair $n \geqslant 5$ il existe



Trois jolis exemples sporadiques du catalogue de Jamison-Hill.



Cette configuration de $n = 6$ points détermine $t = 6$ pentes différentes.



Ici, une direction de départ verticale conduit à $\pi_0 = 123456$.

une configuration de n points qui détermine exactement $n-1$ pentes différentes, et pour tout autre entier $n \geqslant 3$, il existe une configuration ayant exactement n pentes.

Cependant, les configurations représentées ci-dessus sont loin d'être les seules. Par exemple, Jamison et Hill ont décrit quatre familles infinies de configurations, chacune d'entre elles étant constituée de configurations ayant un nombre n impair de points qui ne déterminent que $n-1$ pentes (« configurations critiques de pentes »). En outre, ils ont énuméré 102 exemples « sporadiques » qui ne semblent pas s'inscrire dans une famille infinie, la plupart d'entre eux ayant été découverts après de longues recherches sur ordinateur.

Le bon sens conduit à penser que les problèmes extrémaux sont d'autant plus difficiles que les configurations extrêmes sont diverses et irrégulières. Il y a en effet beaucoup à dire sur la structure des configurations de pentes critiques (voir [2]) mais une classification semble complètement hors de portée. Cependant, le théorème précédent admet une preuve simple qui repose sur deux ingrédients principaux : une réduction du problème à un modèle combinatoire efficace, due à Eli Goodman et Ricky Pollack, et un bel argument à propos de ce modèle grâce auquel Peter Ungar a pu établir la preuve en 1982.

■ **Preuve.** **(1)** Remarquons d'abord qu'il suffit de montrer que chaque ensemble de points du plan de cardinal pair $n = 2m$ $(m \geqslant 2)$ détermine au moins n pentes. En effet, le cas $n = 3$ est trivial et pour tout ensemble de $n = 2m+1 \geqslant 5$ points (non tous alignés) on peut trouver un sous-ensemble de $n-1 = 2m$ points, non tous alignés, qui déterminent déjà $n-1$ pentes.

Ainsi, dans la suite, nous considérons une configuration de $n = 2m$ points du plan qui déterminent $t \geqslant 2$ pentes différentes.

(2) Le modèle combinatoire est obtenu en construisant une suite périodique de permutations. Pour cela, on commence avec une direction du plan qui n'est pas l'une des pentes de la configuration et l'on numérote les points $1, \ldots, n$ selon l'ordre dans lequel ils apparaissent lorsqu'on les projette sur une droite parallèlement à cette direction. Ainsi, la permutation $\pi_0 = 123...n$ représente l'ordre des points relatif à la direction de départ.

Ensuite, on fait tourner la direction de projection en sens inverse des aiguilles d'une montre et l'on observe comment change l'ordre des projetés. Les changements dans l'ordre des points projetés se produisent exactement au moment où la direction correspond à l'une des pentes de la configuration.

Ces changements sont loin d'être aléatoires ou arbitraires : en effectuant une rotation de 180° de la direction, nous obtenons une suite de permutations :

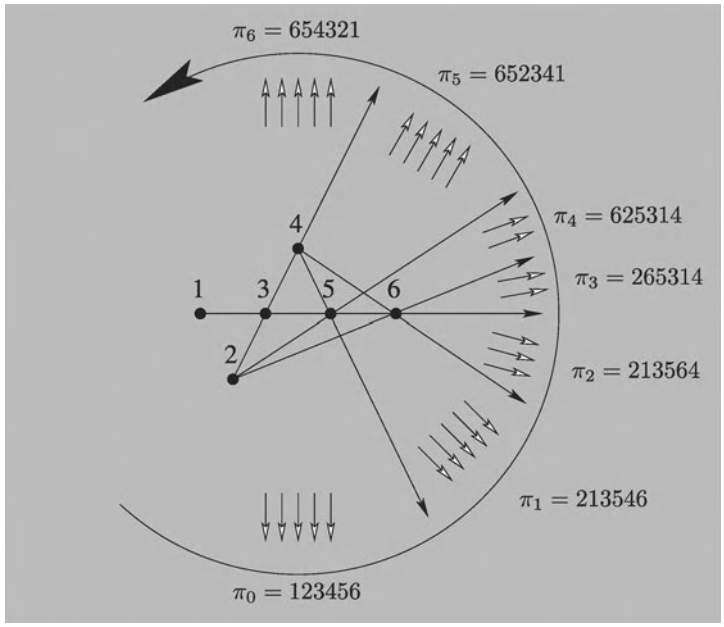
$$\pi_0 \to \pi_1 \to \pi_2 \to \ldots \to \pi_{t-1} \to \pi_t$$

dont on peut énoncer quelques propriétés remarquables.

- La suite commence avec $\pi_0 = 123...n$ et finit avec $\pi_t = n...321$.
- La longueur t de la suite correspond au nombre de pentes de la configu-

ration de points.

- Au cours de la suite, chaque paire $i < j$ est inversée une fois exactement. Cela signifie qu'entre $\pi_0 = 123...n$ et $\pi_t = n...321$, seules des sous-chaînes *croissantes* sont inversées.
- Tout mouvement est constitué de l'inversion d'une ou plusieurs sous-chaînes croissantes disjointes (correspondant à une ou plusieurs droites parallèles à la direction donnée).



Voici la suite des permutations rencontrées dans l'exemple étudié.

En poursuivant ce mouvement circulaire autour de la configuration, on peut considérer la suite comme une partie d'une suite périodique infinie des deux côtés :

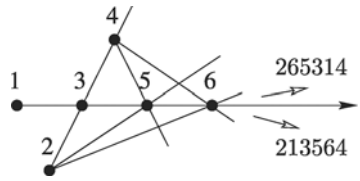
$$\ldots \to \pi_{-1} \to \pi_0 \to \ldots \to \pi_t \to \pi_{t+1} \to \ldots \to \pi_{2t} \to \ldots$$

où π_{i+t} est l'inverse de π_i pour tout i donc $\pi_{i+2t} = \pi_i$ pour tout $i \in \mathbb{Z}$.

Nous allons montrer que *chaque* suite possédant les propriétés ci-dessus (et $t \geqslant 2$) doit avoir une longueur $t \geqslant n$.

(3) La clé de la preuve consiste à scinder chaque permutation en une « moitié gauche » et une « moitié droite » de même longueur $m = \frac{n}{2}$ puis à compter les numéros qui traversent la *barrière* imaginaire délimitant la moitié gauche de la moitié droite.

$\pi_i \to \pi_{i+1}$ est appelé *un mouvement traversant* si l'une des sous-chaînes qu'elle inverse met en cause des numéros qui se trouvent des deux côtés de la barrière. Le mouvement traversant a un *ordre* d s'il déplace $2d$ numéros à travers la barrière, c'est-à-dire si la chaîne traversante comporte exactement d numéros d'un côté et au moins d numéros de l'autre. Ainsi, dans



Un mouvement traversant.

l'exemple :

$$\pi_2 \;=\; 2\underline{13:56}4 \longrightarrow 2\overline{65:31}4 \;=\; \pi_3$$

est un mouvement traversant d'ordre $d = 2$ car il déplace $1, 3, 5, 6$ à travers la barrière (qui est représentée par « : ») et :

$$65\underline{2:34}1 \longrightarrow 65\overline{4:32}1$$

est un mouvement traversant d'ordre $d_2 = 1$, tandis que, par exemple :

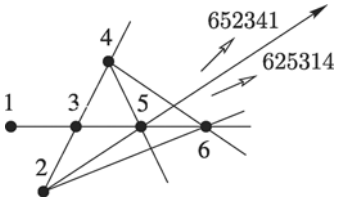
$$6\underline{25}:3\underline{14} \longrightarrow 6\overline{52}:3\overline{41}$$

n'est pas un mouvement traversant.

Au cours de la suite $\pi_0 \to \pi_1 \to \ldots \to \pi_t$, chacun des numéros $1, 2, \ldots, n$ doit traverser la barrière au moins une fois. Si $d_1, \ldots, d_c$ désignent les ordres des c mouvements traversant, cela implique :

$$\sum_{i=1}^{c} 2d_i \;=\; \#\{\text{numéros qui traversent la barrière}\} \;\geqslant\; n$$

Cela implique également que l'on a au moins deux mouvements traversants, puisque l'on a un mouvement traversant tel que $2d_i = n$ seulement si tous les points sont sur une même droite, c'est-à-dire pour $t = 1$. Géométriquement, un mouvement traversant correspond à la pente d'une droite de la configuration qui a moins de m points de chaque côté.

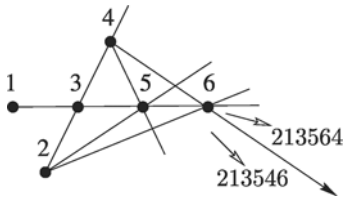


Un mouvement touchant.

(4) Un *mouvement touchant* est un mouvement qui inverse une chaîne adjacente à la barrière centrale mais qui ne la traverse pas. Par exemple :

$$\pi_4 \;=\; 6\underline{25}:3\underline{14} \longrightarrow 6\overline{52}:3\overline{41} \;=\; \pi_5$$

est un mouvement touchant. Géométriquement, un mouvement touchant correspond à la pente d'une droite de la configuration qui possède exactement m points d'un côté et donc au plus $m - 2$ points de l'autre côté.



Un mouvement ordinaire.

Des mouvements qui ne sont ni des mouvements touchants ni des mouvements traversants seront appelés des *mouvements ordinaires*. En voici un exemple :

$$\pi_1 \;=\; 213{:}5\underline{46} \longrightarrow 213{:}5\overline{64} \;=\; \pi_2$$

Ainsi, tout mouvement est soit un mouvement traversant, soit un mouvement touchant, soit un mouvement ordinaire. Utilisons les lettres T *(touchant)*, C *(traversant)* et O *(ordinaire)* pour désigner chaque type de mouvement. $C(d)$ désignera un mouvement traversant d'ordre d. Ainsi, pour l'exemple étudié :

$$\pi_0 \xrightarrow{T} \pi_1 \xrightarrow{O} \pi_2 \xrightarrow{C(2)} \pi_3 \xrightarrow{O} \pi_4 \xrightarrow{T} \pi_5 \xrightarrow{C(1)} \pi_6$$

On peut même désigner cette suite plus brièvement par $T, O, C(2), O, T, C(1)$.

(5) Pour établir la preuve, nous avons besoin des résultats qui suivent.

> *Entre deux mouvements traversants, il y a au moins un mouvement touchant ;*
>
> *Entre tout mouvement traversant d'ordre d et le mouvement touchant suivant, il y a au moins $d-1$ mouvements ordinaires.*

En effet, après un mouvement traversant d'ordre d, la barrière est contenue dans une sous-chaîne symétrique décroissante de longueur $2d$, avec d numéros de chaque côté de la barrière. Avant le mouvement traversant suivant, la barrière centrale doit être placée dans une sous-chaîne croissante de longueur 2 au moins. Toutefois, seul un mouvement touchant affecte le fait que la barrière se trouve dans une sous-chaîne croissante. Cela implique le premier résultat. Quant au second, remarquons qu'à chaque mouvement ordinaire, renversant une sous-chaîne *croissante*, la $2d$-chaîne décroissante ne peut être raccourcie que d'un numéro seulement de chaque côté. Tant que la chaîne décroissante comporte au moins 4 numéros, un mouvement touchant est impossible. Cela implique le second résultat.

Si nous construisons la suite de permutations en commençant par la même projection initiale mais en utilisant la rotation dans le sens des aiguilles d'une montre, nous obtenons alors la suite inversée des permutations. Ainsi, la suite enregistrée doit aussi posséder la propriété réciproque du second résultat, à savoir :

> *Entre un mouvement touchant et le mouvement traversant suivant d'ordre d, il y a au moins $d-1$ mouvements ordinaires.*

(6) Le modèle T-O-C de la suite infinie de permutations, déduit de **(2)**, est obtenu en répétant indéfiniment le modèle T-O-C de longueur t de la suite $\pi_0 \longrightarrow \ldots \longrightarrow \pi_t$. Ainsi, en utilisant les résultats établis en **(5)**, on trouve que dans la suite infinie de mouvements, chaque mouvement traversant d'ordre d se trouve dans un modèle T-O-C de type :

$$T, \underbrace{O, O, \ldots, O}_{\geqslant d-1}, C(d), \underbrace{O, O, \ldots, O}_{\geqslant d-1} \qquad (*)$$

de longueur $1 + (d-1) + 1 + (d-1) = 2d$.

Dans la suite infinie, nous pouvons considérer un segment fini de longueur t qui commence avec un mouvement touchant. Ce segment est constitué de sous-chaînes de type $(*)$, dans lesquelles peuvent être insérés des T supplémentaires. Cela implique que sa longueur t vérifie :

$$t \geqslant \sum_{i=1}^{c} 2d_i \geqslant n$$

ce qui termine la preuve. □

Bibliographie

[1] J. E. GOODMAN & R. POLLACK : *A combinatorial perspective on some problems in geometry,* Congressus Numerantium **32** (1981), 383-394.

[2] R. E. JAMISON & D. HILL : *A catalogue of slope-critical configurations,* Congressus Numerantium **40** (1983), 101-125.

[3] P. R. SCOTT : *On the sets of directions determined by n points,* Amer. Math. Monthly **77** (1970), 502-505.

[4] P. UNGAR : *$2N$ noncollinear points determine at least $2N$ directions,* J. Combinatorial Theory Ser. A **33** (1982), 343-347.

Trois applications de la formule d'Euler

Chapitre 12

On dit qu'un graphe est *planaire* si on peut le représenter dans le plan $\mathbb{R}^2$ (ou, ce qui est équivalent, sur la sphère S^2 de dimension 2) sans que ses arêtes ne se coupent. On dit qu'un graphe est *plan* si l'on dispose déjà d'une telle représentation. Une représentation de ce type décompose le plan ou la sphère en un nombre fini de régions connexes, en incluant la région extérieure non bornée, qui sont appelées *faces*. La formule d'Euler met en évidence une belle relation entre les nombres de sommets, d'arêtes et de faces, valide pour tout graphe plan. Euler a mentionné ce résultat pour la première fois dans une lettre à son ami Goldbach en 1750 mais il n'en avait pas de preuve complète à l'époque. Parmi les nombreuses démonstrations de la formule d'Euler, nous en présentons une particulièrement élégante et auto-duale, qui ne fait pas appel à un raisonnement par récurrence. On peut la trouver dès 1847 dans le livre de von Staudt intitulé *Geometrie der Lage*.

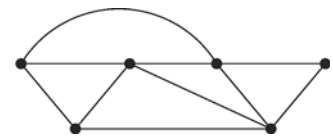
Leonhard Euler

Formule d'Euler. *Si G est un graphe plan connexe à n sommets, e arêtes et f faces, alors*

$$n - e + f = 2$$

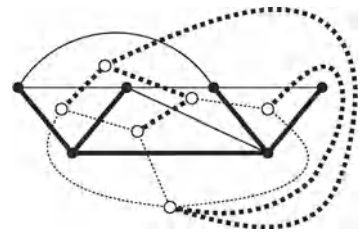
■ **Preuve.** Soit $T \subseteq E$ l'ensemble des arêtes d'un arbre recouvrant de G, c'est à dire d'un sous-graphe minimal qui relie tous les sommets de G. Ce graphe ne contient pas de cycle à cause de la condition de minimalité.

Nous avons besoin du *graphe dual* G^* de G : pour le construire, on place un sommet à l'intérieur de chaque face de G et l'on relie deux sommets de G^* par une arête lorsqu'ils appartiennent à des faces qui ont une arête bordante commune. S'il y a plusieurs arêtes bordantes communes, alors on dessine plusieurs arêtes de liaison dans le graphe dual. Ainsi, G^* peut avoir des arêtes multiples même si le graphe original G est simple.



Un graphe plan G : $n = 6$, $e = 10$, $f = 6$.

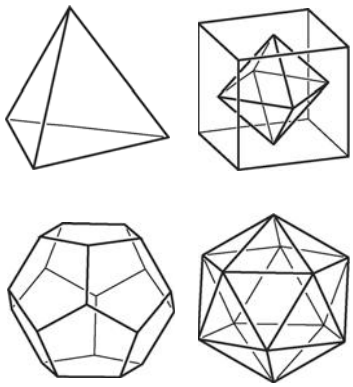
Considérons la collection $T^* \subseteq E^*$ des arêtes du graphe dual correspondant aux arêtes de $E \backslash T$. Les arêtes de T^* relient toutes les faces puisque T n'a pas de cycle. T^* aussi ne comporte pas de cycle, sinon il séparerait certains sommets de G à l'intérieur du cycle de sommets à l'extérieur (ce qui est impossible puisque T est un sous-graphe générateur et que les arêtes de T et de T^* ne se coupent pas). Ainsi, T^* est un arbre recouvrant de G^*.



Arbres recouvrants duaux dans G et dans G^*.

Pour un arbre, le nombre de sommets est égal au nombre d'arêtes plus 1. Pour s'en convaincre, on choisit un sommet comme racine et l'on dirige toutes les arêtes vers l'extérieur de la racine. Cela définit une bijection entre

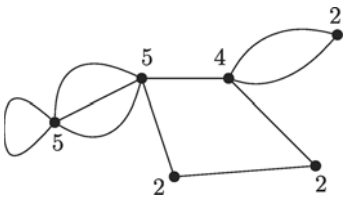
les sommets autres que la racine et les arêtes qui associe à chaque arête le sommet vers lequel elle pointe. En appliquant ce résultat à l'arbre T on trouve $n = e_T + 1$, tandis que pour l'arbre T^* on trouve $f = e_{T^*} + 1$. En additionnant les deux équations, nous obtenons : $n + f = (e_T + 1) + (e_{T^*} + 1) = e + 2$. □



Les cinq solides platoniciens.

La formule d'Euler conduit ainsi à une conclusion *numérique* forte à partir d'une situation *géométrique* et *topologique* : les nombres de sommets, arêtes et faces d'un graphe fini G satisfont $n - e + f = 2$ chaque fois que le graphe est *ou peut être* représenté dans le plan ou sur la sphère. De nombreuses conséquences classiques et bien connues se déduisent de la formule d'Euler. Parmi elles figure la classification des polyèdres réguliers convexes (les solides platoniciens), le fait que K_5 et $K_{3,3}$ ne sont pas planaires (voir ci-dessous) et le théorème des cinq couleurs qui affirme que toute carte plane peut être coloriée avec cinq couleurs au plus de sorte que deux régions adjacentes ne soient pas de la même couleur. Cependant, pour ce dernier résultat il existe une preuve bien meilleure, qui n'utilise même pas la formule d'Euler (voir chapitre 34).

Ce chapitre rassemble trois autres belles preuves qui reposent sur la formule d'Euler. Les deux premières — une preuve du théorème de Sylvester-Gallai et un théorème sur les configurations de points 2-coloriés — utilisent la formule d'Euler astucieusement combinée à d'autres relations arithmétiques qui font intervenir certains paramètres fondamentaux de la théorie des graphes. Examinons d'abord ces paramètres.



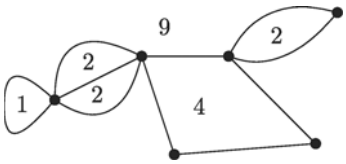
Le degré est précisé à côté de chaque sommet. En comptant les sommets d'un degré donné, on obtient : $n_2 = 3$, $n_3 = 0$, $n_4 = 1$, $n_5 = 2$ et $n_i = 0$ pour les autres valeurs de i.

Le *degré* d'un sommet est le nombre d'arêtes qui ont ce sommet pour extrémité, les arcs qui définissent une boucle comptant double. Notons n_i le nombre de sommets de degré i dans G. En dénombrant les sommets en fonction de leur degré, on obtient :

$$n = n_0 + n_1 + n_2 + n_3 + \dots \tag{1}$$

D'autre part, chaque arête a deux extrémités ; elle contribue donc deux fois à la somme de tous les degrés et ainsi :

$$2e = n_1 + 2n_2 + 3n_3 + 4n_4 + \dots \tag{2}$$



Le nombre de côtés est précisé dans chaque région. Le décompte des faces ayant un nombre donné de côtés conduit à : $f_1 = 1$, $f_2 = 3$, $f_4 = 1$, $f_9 = 1$ et $f_i = 0$ pour les autres valeurs de i.

On peut interpréter cette identité comme le dénombrement des extrémités des arêtes, c'est-à-dire les incidences arête-sommet, effectué de deux façons différentes.

Le *degré moyen* $\overline{d}$ des sommets est donc :

$$\overline{d} = \frac{2e}{n}$$

Comptons ensuite les faces d'un graphe plan en fonction de leur nombre de côtés : une *k-face* est une face qui est bordée par k arêtes (une arête qui borde une même région des deux côtés doit être comptée deux fois !). Soit f_k le nombre de k-faces.

En comptant toutes les faces, on trouve :

$$f \;=\; f_1 + f_2 + f_3 + f_4 + \ldots \qquad (3)$$

En comptant les arêtes en fonction des faces qu'elles bordent, on obtient :

$$2e \;=\; f_1 + 2f_2 + 3f_3 + 4f_4 + \ldots \qquad (4)$$

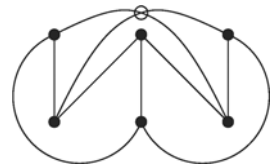
Comme précédemment, cette relation peut être interprétée comme deux manières de dénombrer les incidences arête-face. Remarquons que le nombre moyen de côtés par face est :

$$\overline{f} \;=\; \frac{2e}{f}$$

On peut en déduire rapidement à l'aide de la formule d'Euler que le graphe complet K_5 et le graphe complet biparti $K_{3,3}$ ne sont pas planaires. Relativement à une hypothétique représentation planaire de K_5 on aurait $n = 5$, $e = \binom{5}{2} = 10$, donc $f = e+2-n = 7$ et $\overline{f} = \frac{2e}{f} = \frac{20}{7} < 3$. Toutefois si le nombre moyen de côtés était plus petit que 3, le plongement comporterait une face ayant au plus deux côtés, ce qui est impossible.

K_5 représenté avec une intersection.

De même, pour $K_{3,3}$ on trouve : $n = 6$, $e = 9$ et $f = e + 2 - n = 5$, donc $\overline{f} = \frac{2e}{f} = \frac{18}{5} < 4$, ce qui est impossible puisque $K_{3,3}$ est simple et biparti, donc tous les cycles ont une longueur au moins égale à 4.



$K_{3,3}$ représenté avec une intersection.

Bien entendu, la ressemblance entre les équations (3) et (4) relatives aux f_i et les équations (1) et (2) relatives aux n_i n'est pas fortuite. Elles sont transformées les unes des autres par la relation de dualité $G \to G^*$ précédemment évoquée.

À partir des identités précédentes, on obtient les énoncés suivants. Ce sont des conséquences « locales » importantes de la formule d'Euler.

Proposition. *Soit G un graphe plan simple non vide à n sommets. Alors*
(A) *G a au plus $3n - 6$ arêtes.*

(B) *G a un sommet de degré au plus 5.*

(C) *Si les arêtes de G sont 2-coloriées, alors il existe un sommet de G qui présente au plus deux changements de couleurs relativement à l'ordre cyclique des arêtes défini autour de ce sommet.*

■ **Preuve.** Pour chacune de ces affirmations, on peut supposer que G est connexe.

(A) Chaque face a au moins 3 côtés puisque G est simple. Par suite, (3) et (4) entraînent :

$$f \;=\; f_3 + f_4 + f_5 + \ldots$$

et :

$$2e \;=\; 3f_3 + 4f_4 + 5f_5 + \ldots$$

et donc : $2e - 3f \geqslant 0$. La formule d'Euler conduit alors à :

$$3n - 6 \;=\; 3e - 3f \;\geqslant\; e.$$

(B) D'après le résultat (A), le degré moyen $\overline{d}$ vérifie :

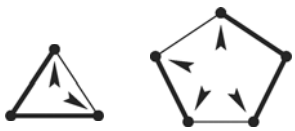
$$\overline{d} = \frac{2e}{n} \leqslant \frac{6n-12}{n} < 6.$$

Il doit donc y avoir un sommet de degré au plus 5.

Les flèches pointent vers les coins présentant un changement de couleurs.

(C) Soit c le nombre de coins où se produisent des changements de couleurs. Supposons que l'affirmation soit fausse. Alors nous avons $c \geqslant 4n$ changements de couleurs, puisqu'à chaque sommet se produit un nombre pair de changements. Comme chaque face ayant $2k$ ou $2k+1$ côtés présente au plus $2k$ coins de ce type :

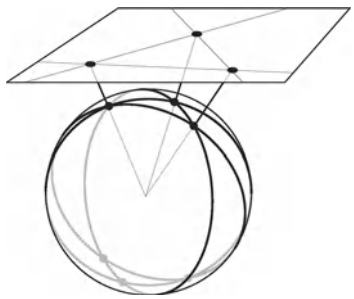
$$\begin{aligned} 4n \leqslant c &\leqslant 2f_3 + 4f_4 + 4f_5 + 6f_6 + 6f_7 + 8f_8 + \dots \\ &\leqslant 2f_3 + 4f_4 + 6f_5 + 8f_6 + 10f_7 + \dots \\ &= 2(3f_3 + 4f_4 + 5f_5 + 6f_6 + 7f_7 + \dots) \\ &\quad -4(f_3 + f_4 + f_5 + f_6 + f_7 + \dots) \\ &= 4e - 4f \end{aligned}$$



en utilisant encore (3) et (4). Ainsi, nous avons $e \geqslant n + f$, contredisant encore une fois la formule d'Euler. □

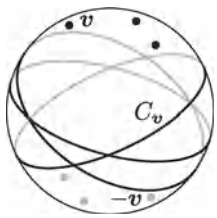
1. Le théorème de Sylvester-Gallai revisité

Norman Steenrod a, semble-t-il, remarqué le premier que la partie (A) de la proposition fournit une preuve remarquablement simple du théorème de Sylvester-Gallai (voir chapitre 10).



> **Le théorème de Sylvester-Gallai.** *Étant donné un ensemble de $n \geqslant 3$ points quelconques non alignés du plan, il existe toujours une droite qui en contient exactement deux.*

■ **Preuve.** (Sylvester-Gallai *via* Euler)
Plongeons le plan $\mathbb{R}^2$ dans $\mathbb{R}^3$ « près » de la sphère unité S^2 comme le montre la figure. Alors chaque point de $\mathbb{R}^2$ correspond à un couple de points antipodaux sur S^2 et les droites de $\mathbb{R}^2$ correspondent à des grands cercles de S^2. Ainsi, le théorème de Sylvester-Gallai est équivalent au résultat suivant :



> *Étant donné un ensemble quelconque de $n \geqslant 3$ couples de points antipodaux sur la sphère, qui ne sont pas tous sur un grand cercle, il existe toujours un grand cercle qui contient exactement deux couples de points antipodaux.*

Maintenant on passe au problème dual en remplaçant chaque paire de points antipodaux par le grand cercle correspondant sur la sphère. C'est-à-dire qu'au lieu de considérer les points $\pm\boldsymbol{v} \in S^2$ on considère les cercles orthogonaux définis par $C_{\boldsymbol{v}} := \{\boldsymbol{x} \in S^2 : \langle \boldsymbol{x}, \boldsymbol{v} \rangle = 0\}$. Ce cercle $C_{\boldsymbol{v}}$ est l'équateur si on considère $\boldsymbol{v}$ comme le pôle nord de la sphère.

Le problème de Sylvester-Gallai revient alors à montrer :

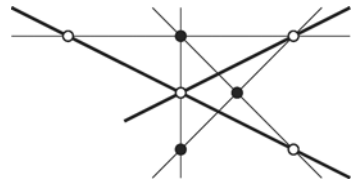
> *Étant donnée une collection quelconque de $n \geqslant 3$ grands cercles sur S^2, ne passant pas tous par un même point, il existe toujours un point qui se trouve exactement sur deux de ces grands cercles.*

La configuration de ces cercles induit un graphe plan simple sur S^2, dont les sommets sont les points d'intersection de deux des grands cercles qui divisent les grands cercles en arêtes. Par construction, tous les degrés des sommets sont pairs et au moins égaux à 4. La partie (A) de la proposition précédente implique l'existence d'un sommet de degré 4. Le théorème est démontré ! □

2. Droites monochromatiques

Le résultat suivant, analogue « coloré » du théorème de Sylvester-Gallai, est dû à Don Chakerian.

Théorème. *Étant donnée une configuration de points noirs et blancs non alignés du plan, il existe toujours une droite « monochromatique », c'est-à-dire une droite qui contient au moins deux points d'une même couleur et aucun point de l'autre.*



■ **Preuve.** Comme pour le problème de Sylvester-Gallai, on transporte le problème sur la sphère unité et l'on passe au problème dual. Il faut donc montrer :

> *Étant donnée une collection finie de grand cercles noirs et blancs sur la sphère unité, ne passant pas tous par un même point, il existe toujours un point d'intersection qui se trouve uniquement sur des grands cercles blancs ou uniquement sur des grands cercles noirs.*

Ce résultat est une simple conséquence de la partie (C) de la proposition précédente puisqu'en chaque sommet intersection de grands cercles de couleurs différentes, se trouvent toujours au moins 4 coins présentant des changements de signe. □

3. Le théorème de Pick

Le théorème de Pick, établi en 1899, est un très beau résultat, en soi surprenant, mais c'est aussi une conséquence classique de la formule d'Euler. Dans ce qui suit, on dit qu'un polygone réseau convexe $P \subseteq \mathbb{R}^2$ est *élémentaire* si ses sommets ont des coordonnées entières et s'il ne contient aucun autre point du réseau.

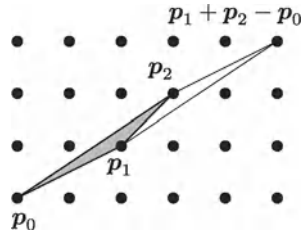
Lemme. *Tout triangle élémentaire $\Delta = \operatorname{conv}\{\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{p}_2\} \subseteq \mathbb{R}^2$ a une aire $A(\Delta) = \frac{1}{2}$.*

Bases d'un réseau

Une *base* de $\mathbb{Z}^2$ est un couple de vecteurs linéairement indépendants $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ tels que :

$$\mathbb{Z}^2 \;=\; \{\lambda_1 \boldsymbol{e}_1 + \lambda_2 \boldsymbol{e}_2 : \lambda_1, \lambda_2 \in \mathbb{Z}\}.$$

Soient $\boldsymbol{e}_1 = \binom{a}{b}$ et $\boldsymbol{e}_2 = \binom{c}{d}$. L'aire du parallélogramme engendré par $\boldsymbol{e}_1$ et $\boldsymbol{e}_2$ vérifie $A(\boldsymbol{e}_1, \boldsymbol{e}_2) = |\det(\boldsymbol{e}_1, \boldsymbol{e}_2)| = |\det \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)|$.
Si $\boldsymbol{f}_1 = \binom{r}{s}$ et $\boldsymbol{f}_2 = \binom{t}{u}$ forment une autre base, alors il existe une $\mathbb{Z}$-matrice inversible Q telle que $\left(\begin{smallmatrix} r & t \\ s & u \end{smallmatrix}\right) = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right) Q$. Puisque $QQ^{-1} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ et puisque les déterminants sont entiers, on trouve $|\det Q| = 1$ et $|\det(\boldsymbol{f}_1, \boldsymbol{f}_2)| = |\det(\boldsymbol{e}_1, \boldsymbol{e}_2)|$. Ainsi, tous les parallélogrammes formant une base ont une même aire égale à 1, puisque $A\left(\binom{1}{0}, \binom{0}{1}\right) = 1$.



■ **Preuve.** Le parallélogramme P de coins $\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_1 + \boldsymbol{p}_2 - \boldsymbol{p}_0$ et le réseau $\mathbb{Z}^2$ sont invariants par l'application :

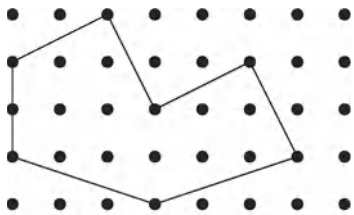
$$\sigma : \boldsymbol{x} \;\longmapsto\; \boldsymbol{p}_1 + \boldsymbol{p}_2 - \boldsymbol{x}$$

qui est la symétrie dont le centre est le milieu du segment reliant $\boldsymbol{p}_1$ à $\boldsymbol{p}_2$. Ainsi, le parallélogramme $P = \Delta \cup \sigma(\Delta)$ est également élémentaire et ses translatés entiers recouvrent le plan. $\{\boldsymbol{p}_1 - \boldsymbol{p}_0, \boldsymbol{p}_2 - \boldsymbol{p}_0\}$ est donc une base du réseau $\mathbb{Z}^2$ qui admet ± 1 pour déterminant. L'aire de P est donc 1 et celle de Δ est $\frac{1}{2}$ (pour des explications à propos de ces calculs, se reporter à l'encadré). □

Théorème. *L'aire de tout polygone $Q \subseteq \mathbb{R}^2$ (non nécessairement convexe) de sommets entiers vérifie :*

$$A(Q) \;=\; n_{int} + \frac{1}{2} n_{bd} - 1$$

où n_{int} (respectivement n_{bd}) désigne le nombre de points entiers à l'intérieur (respectivement sur le bord) de Q.

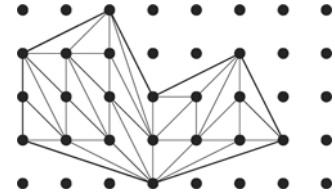


$n_{int} = 11$ et $n_{bd} = 8$ donc $A = 14$.

■ **Preuve.** Un tel polygone peut toujours être triangulé en utilisant tous les n_{int} points du réseau qui se trouvent à l'intérieur et tous les n_{bd} points du réseau qui se trouvent sur le bord de Q. Cette propriété n'est pas complètement évidente, en particulier si l'on n'impose pas à Q d'être convexe, mais l'argument donné au chapitre 35 sur le problème de la surveillance d'un musée le montre.
À présent, interprétons la triangulation comme un graphe plan qui partage le plan en une face non bornée et $f - 1$ triangles d'aire $\frac{1}{2}$. Alors :

$$A(Q) \;=\; \frac{1}{2}(f - 1)$$

Chaque triangle a trois côtés ; chacune des e_{int} arêtes intérieures borde deux triangles, tandis que les e_{bd} arêtes du bord apparaissent chacune dans un unique triangle. Par suite, $3(f-1) = 2e_{int} + e_{bd}$ et il s'ensuit que $f = 2(e-f) - e_{bd} + 3$. En outre, il y a autant d'arêtes que de sommets sur le bord, $e_{bd} = n_{bd}$. Ces deux affirmations combinées avec la formule d'Euler conduisent à :



$$\begin{aligned} f &= 2(e-f) - e_{bd} + 3 \\ &= 2(n-2) - n_{bd} + 3 = 2n_{int} + n_{bd} - 1 \end{aligned}$$

et donc :

$$A(Q) = \tfrac{1}{2}(f-1) = n_{int} + \tfrac{1}{2}n_{bd} - 1 \qquad \square$$

Bibliographie

[1] G. D. CHAKERIAN : *Sylvester's problem on collinear points and a relative,* Amer. Math. Monthly **77** (1970), 164-167.

[2] D. EPPSTEIN : *Nineteen proofs of Euler's formula : $V - E + F = 2$,* in : The Geometry Junkyard, `http://www.ics.uci.edu/~eppstein/junkyard/euler/`.

[3] G. PICK : *Geometrisches zur Zahlenlehre,* Sitzungsberichte Lotos (Prag), Natur-med. Verein für Böhmen **19** (1899), 311-319.

[4] K. G. C. VON STAUDT : *Geometrie der Lage,* Verlag der Fr. Korn'schen Buchhandlung, Nürnberg 1847.

[5] N. E. STEENROD : *Solution 4065/Editorial Note,* Amer. Math. Monthly **51** (1944), 170-171.

Le théorème de rigidité de Cauchy — Chapitre 13

Un résultat célèbre, conséquence de la formule d'Euler (précisément de la partie (C) de la proposition du chapitre précédent) est le théorème de rigidité de Cauchy pour les polyèdres de dimension 3.

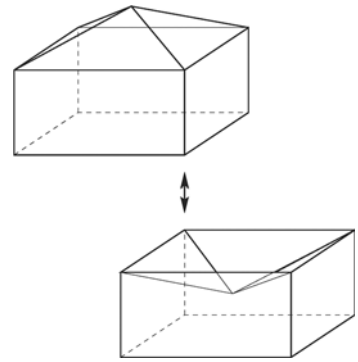
En ce qui concerne les notions de congruence et d'équivalence combinatoire utilisées dans la suite, nous nous référons à l'appendice sur les polytopes et les polyèdres du chapitre concernant le troisième problème de Hilbert (page 68).

Augustin Cauchy

> **Théorème.** *Si deux polyèdres convexes de dimension 3 appelés P et P' sont combinatoirement équivalents et si les facettes de P et P' qui se correspondent sont congruentes, alors les angles entre les paires correspondantes de facettes adjacentes sont égaux (et donc P est congru à P').*

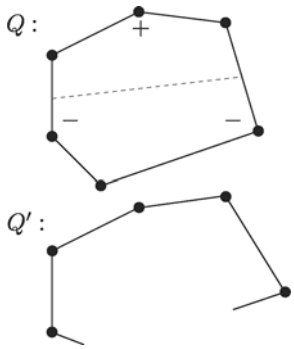
L'illustration figurant en marge montre deux polyèdres de dimension 3 qui sont combinatoirement équivalents et dont les faces correspondantes sont congruentes. Ils ne sont toutefois pas congruents, un seul des deux polyèdres étant convexe. L'hypothèse de convexité est donc essentielle pour le théorème de Cauchy !

■ **Preuve.** Pour l'essentiel, ce qui suit reprend la preuve originale de Cauchy. Soient deux polyèdres convexes P et P' dont les faces sont congruentes. Colorions les arêtes de P comme suit : une arête est noire (ou « positive ») si l'angle intérieur des deux facettes adjacentes est plus grand dans P' que dans P ; elle est blanche (ou « négative ») si l'angle correspondant est plus petit dans P' que dans P.



Les arêtes noires et blanches de P forment ensemble un graphe plan à deux couleurs sur la surface de P, qui peut être envoyé sur la sphère unité par projection radiale en supposant que $\mathbf{0}$ est à l'intérieur de P. Si P et P' ont des angles-facettes correspondants différents, alors le graphe est non vide. En utilisant la partie (C) de la proposition du chapitre précédent, on constate qu'il existe un sommet p adjacent à une arête noire ou blanche au moins, tel qu'il y ait au plus deux changements entre arêtes noires et blanches (relativement à l'ordre cyclique).

Coupons maintenant P avec une petite sphère S_ε de rayon ε et de centre le sommet p. Coupons P' avec une sphère S'_ε de même rayon ε et de centre p' (le sommet correspondant à p dans P'). Dans S_ε et S'_ε nous obtenons

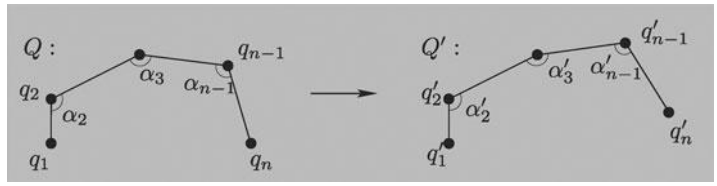


des polygones convexes sphériques Q et Q' dont les arcs correspondants ont même longueur, puisque les facettes de P et P' sont congruentes et que nous avons choisi le même rayon ε.

Marquons d'un $+$ les angles de Q pour lesquels l'angle correspondant dans Q' est plus grand et par $-$ ceux pour lesquels l'angle correspondant dans Q' est plus petit. En d'autres termes, lorsqu'on va de Q vers Q' les angles $+$ s'ouvrent, les angles $-$ se ferment, tandis que les longueurs des côtés et les angles non marqués restent constants.

Le choix de p implique qu'*un certain nombre* de signes $+$ ou $-$ apparaissent et que relativement à l'ordre cyclique il y a au plus deux changements $+/-$. Si un seul type de signes apparaît, alors le lemme qui suit implique directement une contradiction, une arête devant changer de longueur. Si les deux types de signes se produisent, alors (puisqu'il n'y a que deux changements de signe) il y a une « ligne de séparation » qui relie les milieux de deux arêtes et sépare tous les signes $+$ de tous les signes $-$. À nouveau, nous obtenons une contradiction grâce au lemme ci-dessous, puisque la ligne de séparation ne peut pas être à la fois plus grande et plus petite dans Q' que dans Q. □

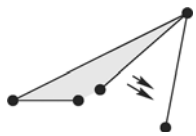
Le lemme du bras de Cauchy.
Soient Q et Q' des n-gones[1] *convexes (planaires ou sphériques), numérotés comme indiqué sur la figure suivante :*



tels que les longueurs des arêtes correspondantes vérifient l'égalité $\overline{q_i q_{i+1}} = \overline{q'_i q'_{i+1}}$ pour $1 \leqslant i \leqslant n-1$, et que les mesures des angles correspondants vérifient l'inégalité $\alpha_i \leqslant \alpha'_i$ pour $2 \leqslant i \leqslant n-1$. Alors la longueur de l'arête manquante vérifie :

$$\overline{q_1 q_n} \leqslant \overline{q'_1 q'_n}$$

avec égalité si et seulement si $\alpha_i = \alpha'_i$ pour tout i.



Il est intéressant de savoir que la preuve originale du lemme de Cauchy était fausse : un mouvement continu qui ouvre les angles et laisse fixe la longueur des côtés peut rendre la convexité caduque (voir figure). Le lemme ainsi que la preuve donnée ici sont issus d'une lettre de I. J. Schoenberg à S. K. Zaremba. Ils sont valides aussi bien pour les polygones planaires que pour les polygones sphériques.

■ **Preuve.** On raisonne par récurrence sur n. Le cas $n = 3$ est facile : si dans un triangle, on augmente l'angle γ entre deux côtés de longueurs fixes

1. N.d.T. : on désigne par *n-gone* un polygone ayant n côtés.

a et b, alors la longueur c du côté opposé augmente aussi. Analytiquement, cela résulte de la relation :

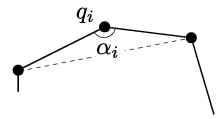
$$c^2 \;=\; a^2 + b^2 - 2ab\cos\gamma$$

dans le cas planaire, et de son analogue :

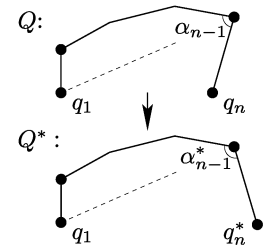
$$\cos c \;=\; \cos a\cos b + \sin a \sin b \cos\gamma$$

en trigonométrie sphérique. Ici, les longueurs a, b, c sont mesurées sur la sphère de rayon 1 ; elles prennent donc leurs valeurs dans l'intervalle $[0, \pi]$.

Supposons maintenant $n \geqslant 4$. Si, pour un certain $i \in \{2, \ldots, n-1\}$, nous avons $\alpha_i = \alpha'_i$, alors le sommet correspondant peut être supprimé en introduisant la diagonale de q_{i-1} à q_{i+1} (respectivement de q'_{i-1} à q'_{i+1}), avec $\overline{q_{i-1}q_{i+1}} = \overline{q'_{i-1}q'_{i+1}}$. Ainsi le résultat est démontré par récurrence. Nous pouvons donc supposer $\alpha_i < \alpha'_i$ pour $2 \leqslant i \leqslant n-1$.



On définit à présent un nouveau polygone Q^* à partir de Q en remplaçant α_{n-1} par le plus grand angle $\alpha^*_{n-1} \leqslant \alpha'_{n-1}$ possible tout en laissant Q^* convexe. À cet effet, q_n est remplacé par q^*_n, les autres q_i étant conservés comme les longueurs d'arêtes et les angles de Q. Bien entendu, si l'on peut choisir $\alpha^*_{n-1} = \alpha'_{n-1}$ en laissant Q^* convexe, alors : $\overline{q_1q_n} < \overline{q_1q^*_n} \leqslant \overline{q'_1q'_n}$, en utilisant le cas $n = 3$ dans un premier temps, puis une récurrence analogue à la précédente dans un deuxième temps.



Sinon, un mouvement non trivial conduit à :

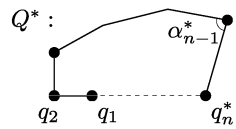
$$\overline{q_1q^*_n} \;>\; \overline{q_1q_n} \tag{1}$$

et nous nous retrouvons dans la situation où q_2, q_1 et q^*_n sont colinéaires et tels que :

$$\overline{q_2q_1} + \overline{q_1q^*_n} \;=\; \overline{q_2q^*_n} \tag{2}$$

En comparant maintenant Q^* et Q', on trouve :

$$\overline{q_2q^*_n} \;\leqslant\; \overline{q'_2q'_n} \tag{3}$$



par récurrence sur n (en ignorant les sommets q_1 et q'_1). On trouve finalement :

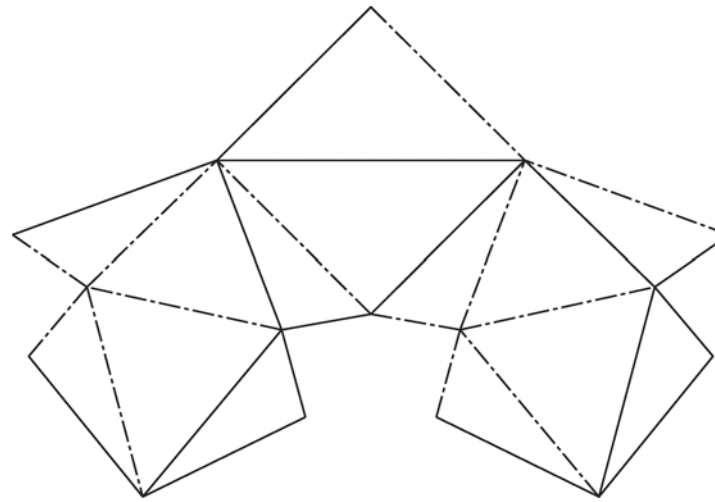
$$\overline{q'_1q'_n} \overset{(*)}{\geqslant} \overline{q'_2q'_n} - \overline{q'_1q'_2} \overset{(3)}{\geqslant} \overline{q_2q^*_n} - \overline{q_1q_2} \overset{(2)}{=} \overline{q_1q^*_n} \overset{(1)}{>} \overline{q_1q_n}$$

où $(*)$ résulte simplement de l'inégalité triangulaire, toutes les autres relations ayant déjà été démontrées. □

Nous avons vu un exemple montrant que le théorème de Cauchy n'est pas vrai pour les polyèdres *non convexes*. La particularité de cet exemple vient de ce qu'il résulte de l'application d'une déformation non continue qui envoie un polyèdre sur l'autre en conservant les facettes congruentes tandis que les angles dièdres font des « sauts ». On peut demander mieux :

Peut-il exister, pour un polyèdre non convexe, une déformation continue *qui laisse les facettes plates et congruentes ?*

Il avait été conjecturé qu'aucune surface triangulée, convexe ou non, ne pouvait admettre un tel mouvement. Ce fut donc une surprise lorsqu'en 1977, plus de cent-soixante ans après le travail de Cauchy, Robert Connelly présenta des contre-exemples : des sphères triangulées (fermées sans bord) plongées dans $\mathbb{R}^3$ (sans auto-intersections) qui sont flexibles, avec un mouvement continu qui conserve les longueurs des arêtes constantes et qui conserve les faces triangulaires congruentes.



Un bel exemple de surface flexible construite par Klaus Steffen. Les droites en pointillés représentent les arêtes qui correspondent aux dièdres rentrants dans ce modèle en papier. Il faut plier les lignes pleines comme des crêtes et les lignes en pointillés comme des vallées. Les arêtes du modèle ont des longueurs égales à 5, 10, 11, 12 et 17 unités.

La théorie de la rigidité des surfaces nous réserve bien d'autres surprises : très récemment, Sabitov a réussi à montrer que lorsqu'une telle surface flexible bouge, le *volume* qu'elle enferme doit être constant. Sa démonstration est très belle notamment dans son recours à une machinerie algébrique qui dépasse le cadre de cet ouvrage.

Bibliographie

[1] A. CAUCHY : *Sur les polygones et les polyèdres, second mémoire,* J. École Polytechnique XVIe Cahier, Tome IX (1813), 87 ; Œuvres Complètes, IIe Série, Vol. 1, Paris 1905, 26-38.

[2] R. CONNELLY : *A counterexample to the rigidity conjecture for polyèdres,* Inst. Haut. Etud. Sci., Publ. Math. **47** (1978), 333-338.

[3] R. CONNELLY : *The rigidity of polyedral surfaces,* Mathematics Magazine **52** (1979), 275-283.

[4] I. KH. SABITOV : *The volume as a metric invariant of polyhedra,* Discrete Comput. Geometry **20** (1998), 405-425.

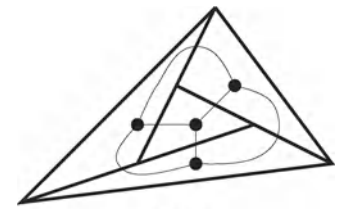
[5] J. SCHOENBERG & S.K. ZAREMBA : *On Cauchy's lemma concerning convex polygons,* Canadian J. Math. **19** (1967), 1062-1071.

Simplexes contigus

Chapitre 14

Combien de simplexes de dimension d peut-on positionner dans $\mathbb{R}^d$ de sorte qu'ils soient deux à deux contigus, c'est-à-dire que leurs intersections deux à deux soient de dimension $(d-1)$?

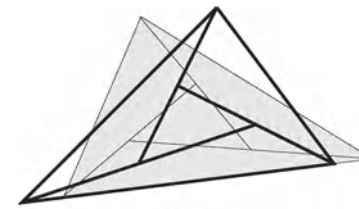
C'est un vieux problème très naturel. Appelons $f(d)$ la réponse à ce problème et notons que, trivialement, $f(1) = 2$. Si $d = 2$ la configuration de quatre triangles représentée ci-contre montre que $f(2) \geqslant 4$. Il n'y a pas de configuration similaire avec cinq triangles, parce que la construction du graphe dual, qui dans notre exemple de quatre triangles conduit à une représentation planaire de K_4, impliquerait un plongement planaire de K_5, ce qui est impossible (voir page 85). Ainsi :



$f(2) \geqslant 4$.

$$f(2) = 4$$

En dimension trois, on voit facilement que $f(3) \geqslant 8$. À cet effet, on utilise la configuration de huit triangles représentée ci-contre. On relie les quatre triangles gris à un point x qui se trouve au dessous du plan de la figure, ce qui induit quatre tétraèdres qui touchent le plan par dessous. De même, on relie les quatre triangles blancs à un point y qui se trouve au dessus du plan du dessin. On obtient ainsi une configuration de huit tétraèdres de $\mathbb{R}^3$, c'est-à-dire, $f(3) \geqslant 8$.



$f(3) \geqslant 8$.

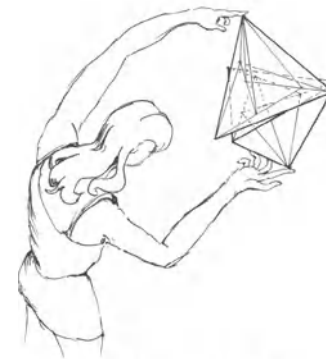
En 1965, Baston a écrit un livre dans lequel il a montré que $f(3) \leqslant 9$ et, en 1991, il fallut à Zaks un autre livre pour établir que :

$$f(3) = 8$$

Comme $f(1) = 2, f(2) = 4$ et $f(3) = 8$, il ne faut pas beaucoup d'inspiration pour formuler la conjecture suivante, énoncée en premier par Bagemihl en 1956.

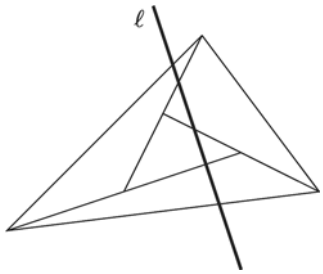
Conjecture. *Le nombre maximal de d-simplexes deux à deux contigus dans une configuration de $\mathbb{R}^d$ est :*

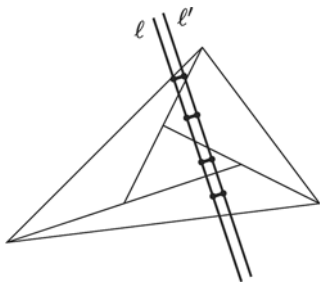
$$f(d) = 2^d$$

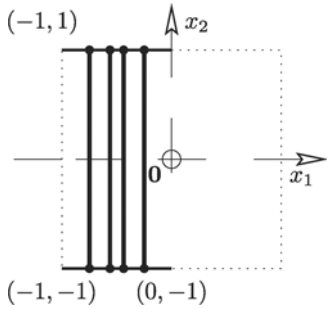


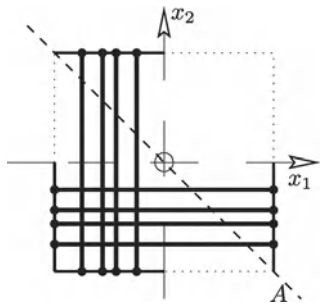
« Simplexes contigus ».

La borne inférieure $f(d) \geqslant 2^d$ est facile à vérifier, si l'on s'y prend bien. Cela demande un usage lourd de transformations de coordonnées affines et une récurrence sur la dimension. On obtient le résultat plus fort suivant, dû à Zaks [4].









Théorème 1. *Pour tout entier $d \geqslant 2$, il existe une famille de 2^d d-simplexes de $\mathbb{R}^d$ deux à deux contigus, ainsi qu'une droite transverse qui rencontre l'intérieur de chacun d'entre eux.*

■ **Preuve.** Si $d = 2$ la famille de quatre triangles que nous venons de considérer est bien traversée par une telle droite. Considérons maintenant une configuration en dimension d de simplexes contigus et qui possède une droite transverse ℓ. Toute droite parallèle voisine ℓ' est aussi une droite transverse. Si nous choisissons ℓ' et ℓ parallèles et suffisamment proches, alors chacun des simplexes contient un intervalle orthogonal (de longueur minimale) reliant les deux droites. Seule une partie bornée des droites ℓ et ℓ' est contenue dans les simplexes de la configuration. Nous pouvons donc ajouter deux segments de liaison en dehors de la configuration, tels que le rectangle engendré par les deux droites de liaisons extérieures (c'est-à-dire, leur enveloppe convexe) contienne tous les autres segments de liaison. Ainsi, nous avons placé une échelle telle que chacun des simplexes de la configuration contienne un des barreaux de l'échelle en son intérieur, tandis que les quatre bouts de l'échelle sont en dehors de la configuration.

La principale étape consiste à effectuer un changement de coordonnées (affines) qui envoie $\mathbb{R}^d$ sur $\mathbb{R}^d$ et transforme le rectangle engendré par l'échelle en un rectangle (demi-carré) défini, comme le montre la figure ci-contre, par :

$$R^1 \;=\; \{(x_1, x_2, 0, \ldots, 0) : -1 \leqslant x_1 \leqslant 0; -1 \leqslant x_2 \leqslant 1\}$$

Ainsi, la configuration des simplexes contigus Σ^1 dans $\mathbb{R}^d$ ainsi obtenus possède l'axe x_1 comme droite transversale et elle est positionnée de façon telle que chacun des simplexes contient un segment

$$S^1(\alpha) \;=\; \{(\alpha, x_2, 0, \ldots, 0) : -1 \leqslant x_2 \leqslant 1\}$$

en son intérieur (avec α tel que $-1 < \alpha < 0$), tandis que l'origine $\mathbf{0}$ est en dehors de tous les simplexes.

Fabriquons maintenant une deuxième copie de cette configuration en prenant la symétrique de la première relativement à l'hyperplan défini par $x_1 = x_2$. Cette deuxième configuration Σ^2 possède l'axe x_2 comme droite transverse et chaque simplexe contient un segment

$$S^2(\beta) \;=\; \{(x_1, \beta, 0, \ldots, 0) : -1 \leqslant x_1 \leqslant 1\}$$

en son intérieur, tel que $-1 < \beta < 0$. Cependant, chaque segment $S^1(\alpha)$ coupe chaque segment $S^2(\beta)$, si bien que l'intérieur de chaque simplexe de Σ^1 coupe chaque simplexe de Σ^2 en son intérieur. Ainsi, si nous ajoutons une nouvelle $(d+1)$-ième coordonnée x_{d+1} et si nous définissons Σ par :

$$\{\text{conv}(P_i \cup \{-\boldsymbol{e}_{d+1}\}) : P_i \in \Sigma^1\} \;\cup\; \{\text{conv}(P_j \cup \{\boldsymbol{e}_{d+1}\}) : P_j \in \Sigma^2\}$$

nous obtenons une configuration de $(d+1)$-simplexes contigus dans $\mathbb{R}^{d+1}$.

En outre, l'antidiagonale :

$$A \;=\; \{(x,-x,0,\ldots,0) : x \in \mathbb{R}\} \;\subseteq\; \mathbb{R}^d$$

coupe tous les segments $S^1(\alpha)$ et $S^2(\beta)$. Nous pouvons « l'incliner » un peu et obtenir une droite :

$$L_\varepsilon \;=\; \{(x,-x,0,\ldots,0,\varepsilon x) : x \in \mathbb{R}\} \;\subseteq\; \mathbb{R}^{d+1}$$

qui, pour tout $\varepsilon > 0$ suffisamment petit, coupe tous les simplexes de Σ. Cela achève la récurrence. □

À l'inverse de cette borne inférieure exponentielle, il est plus difficile d'obtenir des bornes supérieures fines. Un raisonnement par récurrence naïf (considérant séparément toutes les facettes hyperplanes d'une configuration contiguë) implique seulement que :

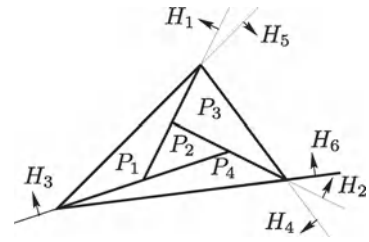
$$f(d) \;\leqslant\; \frac{2}{3}(d+1)!$$

ce qui est bien loin de la borne inférieure du théorème 1. Cependant, Micha Perles a trouvé la démonstration magique suivante, qui fournit une borne bien meilleure.

Théorème 2. *Pour tout $d \geqslant 1$, on a $f(d) < 2^{d+1}$.*

■ **Preuve.** Étant donnée une configuration de r d-simplexes contigus P_1, $P_2, \ldots, P_r$ dans $\mathbb{R}^d$, commençons par dénombrer les différents hyperplans $H_1, H_2, \ldots, H_s$ engendrés par les facettes de P_i. Pour chacune d'entre elles, choisissons arbitrairement un côté positif H_i^+ et appelons l'autre côté H_i^-.

Par exemple, pour la configuration de dimension 2 de $r = 4$ triangles dessinée à droite, nous trouvons $s = 6$ hyperplans (qui sont des droites si $d = 2$).



À partir de ces données, construisons la *B-matrice* comme suit : c'est une matrice de $r \times s$ coefficients dans $\{+1, -1, 0\}$, définie par :

$$B_{ij} := \begin{cases} +1 & \text{si } P_i \text{ a une facette dans } H_j, \text{ et } P_i \subseteq H_j^+ \\ -1 & \text{si } P_i \text{ a une facette dans } H_j, \text{ et } P_i \subseteq H_j^- \\ \;\;0 & \text{si } P_i \text{ n'a pas de facette dans } H_j \end{cases}$$

Par exemple, la configuration de dimension 2 de la figure conduit à la matrice :

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

Voici trois propriétés importantes de ce type de matrices. Premièrement, puisque chaque d-simplexe comporte $d+1$ faces, chaque ligne de B a

exactement $d+1$ coefficients non nuls et exactement $s-(d+1)$ coefficients nuls. Deuxièmement, nous travaillons avec une configuration de simplexes deux à deux contigus, donc pour chaque paire de lignes, il existe une colonne dans laquelle une ligne a un coefficient égal à $+1$, alors que le coefficient correspondant dans l'autre ligne est égal à -1. En d'autres termes, les lignes sont différentes *même si l'on ne tient pas compte des coefficients nuls*. Troisièmement, les lignes de B « représentent » les simplexes P_i selon la formule :

$$P_i = \bigcap_{j:B_{ij}=1} H_j^+ \cap \bigcap_{j:B_{ij}=-1} H_j^- \qquad (*)$$

À présent, construisons à partir de B une nouvelle matrice C, où chaque ligne de B est remplacée par tous les vecteurs lignes que l'on peut générer à partir d'elle en remplaçant tous les 0 soit par $+1$ soit par -1. Puisque chaque ligne de B a $s-d-1$ coefficients nuls et puisque B a r lignes, la matrice C a $2^{s-d-1}r$ lignes.

Dans notre exemple, la matrice C est une matrice de 32×6 coefficients qui débute comme suit :

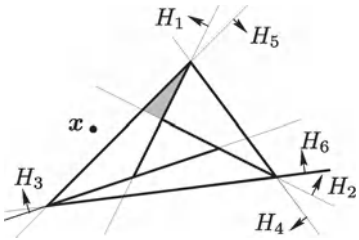
$$C = \left(\begin{array}{cccccc} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ \hline -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array}\right)$$

les huit premières lignes de C étant déduites de la première ligne de B, les huit lignes suivantes provenant de la seconde ligne de B, etc.

Il est important de constater que toutes les lignes de C sont distinctes : si deux lignes sont déduites de la même ligne de B, alors elles sont différentes puisque leurs zéros ont été remplacés de façons différentes ; si elles sont déduites de deux lignes différentes de B, elles sont différentes quelque soit la façon dont les zéros ont été remplacés. Mais les lignes de C sont des vecteurs de composantes (± 1) et de longueur s. Il y en a seulement 2^s différents. Ainsi, puisque les lignes de C sont distinctes, C a au plus 2^s lignes et donc :

$$2^{s-d-1}r \leqslant 2^s$$

Cependant, tous les (± 1)-vecteurs possibles n'apparaissent pas dans C, ce qui implique une inégalité stricte $2^{s-d-1}r < 2^s$ donc $r < 2^{d+1}$. Pour s'en convaincre, notons que chaque ligne de C représente une intersection de demi-espaces, exactement comme les lignes de B dans la formule $(*)$. L'intersection est un sous-ensemble du simplexe P_i déterminé par la ligne



La première ligne de la C-matrice représente le triangle ombré, tandis que la deuxième ligne correspond à une intersection vide de demi-espaces. Le point $\boldsymbol{x}$ induit le vecteur :

$$\begin{pmatrix} 1 & -1 & 1 & 1 & -1 & 1 \end{pmatrix}$$

qui n'apparaît pas dans la C-matrice.

de B correspondante. Prenons un point $\boldsymbol{x} \in \mathbb{R}^d$ qui ne se trouve dans aucun des hyperplans H_j et dans aucun des simplexes P_i. Avec cet $\boldsymbol{x}$, nous construisons un (± 1)-vecteur qui enregistre pour chaque j si $\boldsymbol{x} \in H_j^+$ ou si $\boldsymbol{x} \in H_j^-$. Ce (± 1)-vecteur n'apparaît pas dans C, parce que son demi-espace intersection selon $(*)$ contient $\boldsymbol{x}$ et n'est donc contenu dans aucun simplexe P_i. □

Bibliographie

[1] F. BAGEMIHL : *A conjecture concerning neighboring tetrahedra,* Amer. Math. Monthly **63** (1956) 328-329.

[2] V. J. D. BASTON : *Some Properties of Polyhedra in Euclidean Space,* Pergamon Press, Oxford 1965.

[3] M. A. PERLES : *At most* 2^{d+1} *neighborly simplexes in* E^d, Annals of Discrete Math. **20** (1984), 253-254.

[4] J. ZAKS : *Neighborly families of* 2^d *d-simplexes in* E^d, Geometriae Dedicata **11** (1981), 279-296.

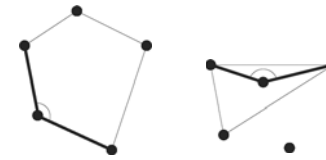
[5] J. ZAKS : *No nine neighborly tetrahedra exist,* Memoirs Amer. Math. Soc. No. 447, Vol. 91, 1991.

Tout grand ensemble de points détermine un angle obtus

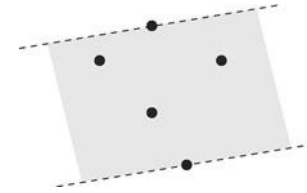
Chapitre 15

Autour des années 1950, Paul Erdős a conjecturé que tout ensemble de plus de 2^d points de $\mathbb{R}^d$ détermine au moins un *angle obtus*, c'est-à-dire un angle strictement supérieur à $\frac{\pi}{2}$. En d'autres termes, tout ensemble de points de $\mathbb{R}^d$ qui ne font que des angles aigus (incluant les angles droits) a un cardinal au plus égal à 2^d. Ce problème, posé comme question primée par la Société Mathématique Néerlandaise, ne recueillit de solutions que pour $d = 2$ et $d = 3$.

Pour $d = 2$, le problème est facile : les cinq points peuvent déterminer un pentagone convexe, qui a toujours un angle obtus (en fait, au moins un angle d'au moins $108°$). Sinon, un point est contenu dans l'enveloppe convexe de trois autres, qui forment un triangle. Ce point voit les trois arêtes du triangle sous trois angles dont la somme est $360°$; l'un des angles vaut donc au moins $120°$ (le deuxième cas concerne aussi les situations dans lesquelles trois points sont alignés et où figure donc un angle de $180°$).



Indépendamment, Victor Klee a demandé quelques années plus tard — question étendue par Erdős — quelle pouvait être la taille d'un ensemble de points de $\mathbb{R}^d$ possédant la propriété d'« antipodalité » suivante : pour *tout* couple de points de l'ensemble, il existe une bande (bornée par deux hyperplans parallèles) contenant l'ensemble des points et telle que les deux points choisis se trouvent sur deux côtés distincts du bord.



Puis, en 1962, Ludwig Danzer et Branko Grünbaum ont résolu les deux problèmes à la fois : ils ont intercalé les deux cardinaux étudiés dans une chaîne d'inégalités, qui commence et finit par 2^d. Ainsi, la réponse 2^d est commune au problème d'Erdős et à celui de Klee.

Dans la suite, nous considérons des ensembles (finis) $S \subseteq \mathbb{R}^d$ de points, leurs enveloppes convexes $\mathrm{conv}(S)$ et des polytopes convexes généraux $Q \subseteq \mathbb{R}^d$ (voir encadré sur les polytopes, page 68 pour un rappel des notions élémentaires). Nous supposons que la dimension de ces ensembles est effectivement d, c'est-à-dire qu'ils ne sont inclus dans aucun hyperplan. De tels ensembles *sont contigus* s'ils ont au moins un point de leur bord en commun, alors que leurs intérieurs ne se coupent pas. Pour tout ensemble $Q \subseteq \mathbb{R}^d$ et tout vecteur $\boldsymbol{s} \in \mathbb{R}^d$, on note $Q + \boldsymbol{s}$ l'image de Q par la translation qui envoie $\mathbf{0}$ sur $\boldsymbol{s}$. De façon similaire, $Q - \boldsymbol{s}$ est le translaté obtenu par l'application qui envoie $\boldsymbol{s}$ sur l'origine.

Que le lecteur ne soit pas intimidé : ce chapitre propose une excursion en géométrie de dimension d. Toutefois, les arguments qui suivent ne demandent pas d'« intuition en grande dimension » puisqu'ils peuvent tous être suivis, visualisés et donc *compris* en dimension trois ou même dans le

plan. Ainsi, des figures vont illustrer la démonstration lorsque $d = 2$ (un « hyperplan » est ici simplement une droite) ; le lecteur pourra imaginer les figures pour $d = 3$ (un « hyperplan » est alors un plan).

Théorème 1. *Pour tout d, on a la chaîne d'égalités et d'inégalités suivante :*

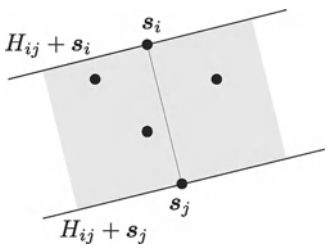
$$
\begin{aligned}
2^d &\overset{(1)}{\leqslant} \max\left\{\#S \,\middle|\, S \subseteq \mathbb{R}^d,\ \sphericalangle(\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{s}_k) \leqslant \tfrac{\pi}{2} \text{ pour tout } \{\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{s}_k\} \subseteq S\right\} \\
&\overset{(2)}{\leqslant} \max\left\{\#S \,\middle|\, \begin{array}{l} S \subseteq \mathbb{R}^d \text{ tel que pour tout couple de points } \{\boldsymbol{s}_i, \boldsymbol{s}_j\} \subseteq \\ S, \text{ il existe une bande } \mathcal{S}(i,j) \text{ qui contient } S, \text{ telle que} \\ \boldsymbol{s}_i \text{ et } \boldsymbol{s}_j \text{ se trouvent dans les hyperplans parallèles formant le bord de } \mathcal{S}(i,j) \end{array}\right\} \\
&\overset{(3)}{=} \max\left\{\#S \,\middle|\, \begin{array}{l} S \subseteq \mathbb{R}^d \text{ tel que les translatés } P - \boldsymbol{s}_i \text{ de } P := \operatorname{conv}(S) \\ \text{se coupent en un même point et sont seulement contigus} \end{array}\right\} \\
&\overset{(4)}{\leqslant} \max\left\{\#S \,\middle|\, \begin{array}{l} S \subseteq \mathbb{R}^d \text{ tel que les translatés } Q + \boldsymbol{s}_i \text{ d'un certain} \\ \text{polytope convexe de dimension } d,\ Q \subseteq \mathbb{R}^d, \text{ sont deux} \\ \text{à deux contigus} \end{array}\right\} \\
&\overset{(5)}{=} \max\left\{\#S \,\middle|\, \begin{array}{l} S \subseteq \mathbb{R}^d \text{ tel que les translatés } Q^* + \boldsymbol{s}_i \text{ d'un certain} \\ \text{polytope convexe à symétrie centrale } Q^* \subseteq \mathbb{R}^d \text{ sont} \\ \text{deux à deux contigus} \end{array}\right\} \\
&\overset{(6)}{\leqslant} 2^d.
\end{aligned}
$$

■ **Preuve.** Nous devons vérifier six affirmations (égalités ou inégalités). Allons-y.

(1) Considérons $S := \{0,1\}^d$ l'ensemble des sommets du cube unité standard de $\mathbb{R}^d$ et choisissons $\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{s}_k \in S$. Par symétrie, nous pouvons supposer que : $\boldsymbol{s}_j = \mathbf{0}$ est le vecteur nul. Ainsi, l'angle peut être calculé à partir de :

$$\cos \sphericalangle(\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{s}_k) = \frac{\langle \boldsymbol{s}_i, \boldsymbol{s}_k \rangle}{|\boldsymbol{s}_i||\boldsymbol{s}_k|}$$

qui est évidemment positif. Ainsi S est un ensemble tel que $|S| = 2^d$ et qui n'a pas d'angles obtus.



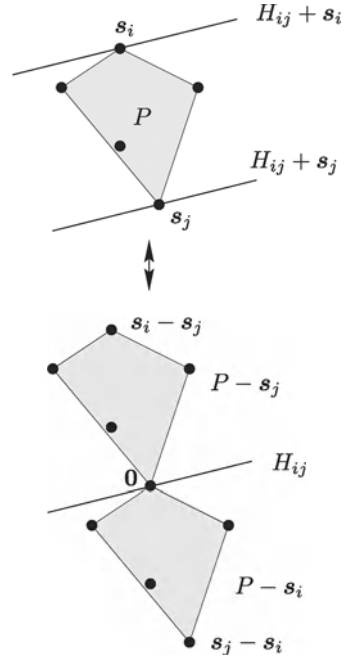
(2) Si S ne contient pas d'angles obtus, alors pour tout $\boldsymbol{s}_i, \boldsymbol{s}_j \in S$ nous pouvons définir les hyperplans parallèles $H_{ij} + \boldsymbol{s}_i$ (respectivement $H_{ij} + \boldsymbol{s}_j$) passant par $\boldsymbol{s}_i$ (respectivement $\boldsymbol{s}_j$) orthogonaux à l'arête $[\boldsymbol{s}_i, \boldsymbol{s}_j]$. Ici, $H_{ij} = \{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{s}_i - \boldsymbol{s}_j \rangle = 0\}$ est l'hyperplan *passant par l'origine* et orthogonal à la droite passant par $\boldsymbol{s}_i$ et $\boldsymbol{s}_j$, $H_{ij} + \boldsymbol{s}_j = \{\boldsymbol{x} + \boldsymbol{s}_j : \boldsymbol{x} \in H_{ij}\}$ est le translaté de H_{ij} qui passe par $\boldsymbol{s}_j$, etc. Ainsi, la bande qui se trouve entre $H_{ij} + \boldsymbol{s}_i$ et $H_{ij} + \boldsymbol{s}_j$ est exactement constituée, en plus de $\boldsymbol{s}_i$ et $\boldsymbol{s}_j$, de tous les points $\boldsymbol{x} \in \mathbb{R}^d$ tels que les angles $\sphericalangle(\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{x})$ et $\sphericalangle(\boldsymbol{s}_j, \boldsymbol{s}_i, \boldsymbol{x})$ soient non-obtus. Par conséquent, la bande contient tout S.

(3) P est contenu dans le demi-espace de $H_{ij} + \boldsymbol{s}_j$ qui contient $\boldsymbol{s}_i$ si et seulement si $P - \boldsymbol{s}_j$ est contenu dans le demi-espace de H_{ij} qui contient

$\boldsymbol{s}_i - \boldsymbol{s}_j$: une propriété du type « un objet est contenu dans un demi-espace » n'est pas détruite si l'on translate à la fois l'objet et le demi-espace de la même quantité (ici par $-\boldsymbol{s}_j$). De la même façon, P est contenu dans le demi-espace de $H_{ij} + \boldsymbol{s}_i$ qui contient $\boldsymbol{s}_j$ si et seulement si $P - \boldsymbol{s}_i$ est contenu dans le demi-espace de H_{ij} qui contient $\boldsymbol{s}_j - \boldsymbol{s}_i$.

En réunissant ces deux assertions, on conclut que le polytope P est contenu dans la bande qui se trouve entre $H_{ij}+\boldsymbol{s}_i$ et $H_{ij}+\boldsymbol{s}_j$ si et seulement si $P-\boldsymbol{s}_i$ et $P - \boldsymbol{s}_j$ appartiennent à différents demi-espaces relatifs à l'hyperplan H_{ij}.

Cette correspondance est illustrée par la figure qui se trouve dans la marge.



En utilisant, en outre, le fait que $\boldsymbol{s}_i \in P = \mathrm{conv}(S)$, on constate que l'origine $\mathbf{0}$ est contenue dans tous les translatés $P - \boldsymbol{s}_i$ ($\boldsymbol{s}_i \in S$). Ainsi, les ensembles $P - \boldsymbol{s}_i$ se coupent tous en $\mathbf{0}$ mais sont seulement contigus : leurs intérieurs sont disjoints deux à deux puisqu'ils se trouvent sur les côtés opposés des hyperplans H_{ij} correspondants.

(4) On obtient gratuitement le résultat suivant : la condition « les translatés sont deux à deux contigus » est plus faible que la condition « ils se coupent en un point commun mais en étant contigus ». De la même façon, on peut affaiblir les conditions en supposant que P est un d-polytope convexe arbitraire de $\mathbb{R}^d$. On peut aussi remplacer S par $-S$.

(5) La partie $\geqslant$ de l'égalité est triviale mais ne constitue pas le sens de l'égalité qui importe pour clore la démonstration. On doit commencer avec une configuration $S \subseteq \mathbb{R}^d$ et un d-polytope $Q \subseteq \mathbb{R}^d$ arbitraire dont les translatés $Q + \boldsymbol{s}_i$ ($\boldsymbol{s}_i \in S$) sont deux à deux contigus. Dans cette situation, on peut utiliser :

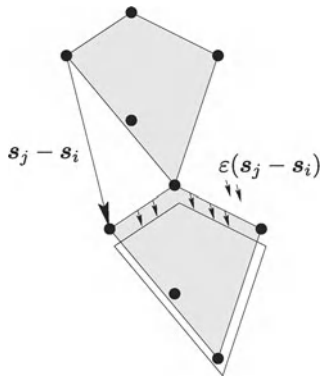
$$Q^* := \left\{ \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{y}) \in \mathbb{R}^d : \boldsymbol{x}, \boldsymbol{y} \in Q \right\}$$

à la place de Q. Il n'est pas difficile de s'en convaincre : Q^* est de dimension d, il est convexe et à symétrie centrale. On peut vérifier que Q^* est un polytope (ses sommets sont de la forme $\frac{1}{2}(\boldsymbol{q}_i - \boldsymbol{q}_j)$, où $\boldsymbol{q}_i, \boldsymbol{q}_j$ sont des sommets de Q) mais ce n'est pas important pour nous.

Nous allons montrer maintenant que $Q + \boldsymbol{s}_i$ et $Q + \boldsymbol{s}_j$ sont contigus *si et seulement si* $Q^* + \boldsymbol{s}_i$ et $Q^* + \boldsymbol{s}_j$ sont contigus. À cet effet, on remarque, comme l'a suggéré Minkowski que :

$$\begin{aligned}
&(Q^*+\boldsymbol{s}_i) \cap (Q^* + \boldsymbol{s}_j) \neq \emptyset \\
&\quad\iff \exists\, \boldsymbol{q}'_i, \boldsymbol{q}''_i, \boldsymbol{q}'_j, \boldsymbol{q}''_j \in Q : \tfrac{1}{2}(\boldsymbol{q}'_i - \boldsymbol{q}''_i) + \boldsymbol{s}_i = \tfrac{1}{2}(\boldsymbol{q}'_j - \boldsymbol{q}''_j) + \boldsymbol{s}_j \\
&\quad\iff \exists\, \boldsymbol{q}'_i, \boldsymbol{q}''_i, \boldsymbol{q}'_j, \boldsymbol{q}''_j \in Q : \tfrac{1}{2}(\boldsymbol{q}'_i + \boldsymbol{q}''_j) + \boldsymbol{s}_i = \tfrac{1}{2}(\boldsymbol{q}'_j + \boldsymbol{q}''_i) + \boldsymbol{s}_j \\
&\quad\iff \exists\, \boldsymbol{q}_i, \boldsymbol{q}_j \in Q : \boldsymbol{q}_i + \boldsymbol{s}_i = \boldsymbol{q}_j + \boldsymbol{s}_j \\
&\quad\iff (Q + \boldsymbol{s}_i) \cap (Q + \boldsymbol{s}_j) \neq \emptyset
\end{aligned}$$

où la troisième équivalence (cruciale) repose sur le fait que chaque $\boldsymbol{q} \in Q$ peut s'écrire comme $\boldsymbol{q} = \frac{1}{2}(\boldsymbol{q} + \boldsymbol{q})$ pour obtenir l'implication réciproque et que Q est convexe, donc que $\frac{1}{2}(\boldsymbol{q}'_i + \boldsymbol{q}''_j)$, $\frac{1}{2}(\boldsymbol{q}'_j + \boldsymbol{q}''_i) \in Q$ pour obtenir l'implication directe.



Ainsi, le passage de Q à Q^* (connu sous le nom de *symétrisation de Minkowski*) conserve la propriété que deux translatés $Q + s_i$ et $Q + s_j$ se coupent. En d'autres termes, nous avons montré que pour tout convexe Q, deux translatés $Q + s_i$ et $Q + s_j$ se coupent si et seulement si les translatés $Q^* + s_i$ et $Q^* + s_j$ se coupent.

La caractérisation suivante montre que la symétrisation de Minkowski conserve la contiguïté :

> *$Q + s_i$ et $Q + s_j$ sont contigus si et seulement s'ils se coupent, alors que $Q + s_i$ et $Q + s_j + \varepsilon(s_j - s_i)$ ne se coupent pour aucun $\varepsilon > 0$.*

(6) Supposons que $Q^* + s_i$ et $Q^* + s_j$ soient contigus. Pour tout point d'intersection :

$$x \in (Q^* + s_i) \cap (Q^* + s_j)$$

nous avons :

$$x - s_i \in Q^* \quad \text{and} \quad x - s_j \in Q^*$$

donc, puisque Q^* est à symétrie centrale :

$$s_i - x = -(x - s_i) \in Q^*$$

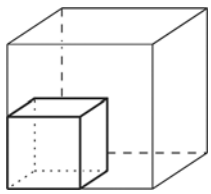
et donc, puisque Q^* est convexe :

$$\tfrac{1}{2}(s_i - s_j) = \tfrac{1}{2}\left((x - s_j) + (s_i - x)\right) \in Q^*$$

$\frac{1}{2}(s_i + s_j)$ est donc contenu dans $Q^* + s_j$ pour tout i. En conséquence, pour $P := \mathrm{conv}(S)$, nous obtenons :

$$P_j := \tfrac{1}{2}(P + s_j) = \mathrm{conv}\left\{\tfrac{1}{2}(s_i + s_j) : s_i \in S\right\} \subseteq Q^* + s_j$$

ce qui implique que les ensembles $P_j = \frac{1}{2}(P + s_j)$ peuvent seulement être contigus.



Facteur de normalisation $\frac{1}{2}$, $\mathrm{vol}(P_j) = \frac{1}{8}\mathrm{vol}(P)$.

Enfin, les ensembles P_j sont contenus dans P, parce que tous les points s_i, s_j et $\frac{1}{2}(s_i + s_j)$ sont dans P car P est convexe. Cependant, les P_j sont seulement plus petits, normalisés, translatés de P et contenus dans P. Le facteur de normalisation est $\frac{1}{2}$, ce qui implique que :

$$\mathrm{vol}(P_j) \;=\; \frac{1}{2^d}\mathrm{vol}(P)$$

puisque nous travaillons avec des ensembles de dimension d. Cela signifie qu'au plus 2^d ensembles P_j s'ajustent dans P et donc que $|S| \leqslant 2^d$.

La chaîne des inégalités est complète et la preuve est terminée. □

L'histoire n'est pas encore terminée ! Danzer et Grünbaum ont posé la question naturelle suivante :

> *Que se passe-t-il si l'on demande que tous les angles soient* aigus *au lieu d'être simplement non-obtus, c'est-à-dire si l'on interdit les angles droits ?*

Ils ont construit des configurations de $2d-1$ points de $\mathbb{R}^d$ ne formant que des angles aigus, conjecturant que cela devait être la meilleure possibilité. Grünbaum a montré que c'est vrai pour $d \leqslant 3$. Cependant, vingt et un an après, en 1983, Paul Erdős et Zoltan Füredi ont montré que la conjecture est fausse, de manière considérable si la dimension est grande ! Leur démonstration constitue une excellente illustration de la puissance des arguments probabilistes (se reporter au chapitre 40 pour une introduction aux méthodes probabilistes). La version de la démonstration que nous proposons tient compte d'une amélioration à la preuve de Erdős et Füredi proposée par David Bevan, un de nos lecteurs.

Théorème 2. *Pour tout $d \geqslant 2$, il existe un ensemble $S \subseteq \{0,1\}^d$ de $2\lfloor \frac{\sqrt{6}}{9}\big(\frac{2}{\sqrt{3}}\big)^d \rfloor$ points de $\mathbb{R}^d$ (sommets du cube unité de dimension d) qui ne déterminent que des angles aigus.*
En particulier, en dimension $d = 34$, il existe un ensemble de $72 > 2\,.\,34-1$ points ne formant que des angles aigus.

■ **Preuve.** Posons $m := \lfloor \frac{\sqrt{6}}{9}\big(\frac{2}{\sqrt{3}}\big)^d \rfloor$ et prenons $3m$ vecteurs

$$\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(3m) \;\in\; \{0,1\}^d$$

en choisissant toutes leurs coordonnées indépendamment et aléatoirement, de façon à ce qu'elles soient égales à 0 ou 1, avec une probabilité $\frac{1}{2}$ pour chaque alternative. On peut lancer une pièce parfaite $3md$ fois à cet effet ; cependant, si d est grand, cela devient rapidement lassant.

Nous avons vu précédemment que les angles définis par des vecteurs de coordonnées 0 ou 1 sont nécessairement aigus. Rappelons que trois vecteurs $\boldsymbol{x}(i), \boldsymbol{x}(j), \boldsymbol{x}(k)$ déterminent un angle droit de sommet $\boldsymbol{x}(j)$ si et seulement si le produit scalaire :

$$\big\langle\, \boldsymbol{x}(i) - \boldsymbol{x}(j), \boldsymbol{x}(k) - \boldsymbol{x}(j) \,\big\rangle$$

s'annule, c'est-à-dire, dans le cas présent, si :

$$x(i)_\ell - x(j)_\ell = 0 \quad \text{ou} \quad x(k)_\ell - x(j)_\ell = 0$$

pour chaque coordonnée ℓ. Nous appelons (i,j,k) un *mauvais triplet* si cela se produit (si $\boldsymbol{x}(i) = \boldsymbol{x}(j)$ ou $\boldsymbol{x}(j) = \boldsymbol{x}(k)$, alors l'angle n'est pas défini mais le triplet (i,j,k) est certainement mauvais).

La probabilité qu'un triplet particulier soit mauvais est exactement $\big(\frac{3}{4}\big)^d$. En effet, il sera bon si et seulement si l'une des d coordonnées ℓ vérifie :

$$\begin{aligned} &\text{soit} \quad x(i)_\ell = x(k)_\ell = 0, \quad x(j)_\ell = 1, \\ &\text{soit} \quad x(i)_\ell = x(k)_\ell = 1, \quad x(j)_\ell = 0. \end{aligned}$$

On se retrouve donc avec six mauvaises options parmi huit équiprobables. Un triplet sera mauvais si et seulement si l'une des mauvaises options (de probabilité $\frac{3}{4}$) se produit pour chacune des d coordonnées.

Le nombre de triplets que nous devons considérer est $3\binom{3m}{3}$ puisqu'il y a $\binom{3m}{3}$ ensemble de trois vecteurs et que, pour chacun d'entre eux, il y a trois choix pour chaque sommet. Bien sûr, les événements définissant des situations dans lesquelles différents triplets sont mauvais ne sont pas indépendants mais la *linéarité de l'espérance* (valeur obtenue en faisant la moyenne sur l'ensemble de toutes les sélections possibles ; voir appendice) implique que le nombre *attendu* de mauvais triplets est exactement $3\binom{3m}{3}\left(\frac{3}{4}\right)^d$. Cela signifie — c'est là que les méthodes probabilistes montrent leur puissance — qu'il existe un *certain* choix des $3m$ vecteurs tel qu'il y ait au plus $3\binom{3m}{3}\left(\frac{3}{4}\right)^d$ mauvais triplets, avec :

$$3\tbinom{3m}{3}\left(\tfrac{3}{4}\right)^d \;<\; 3\tfrac{(3m)^3}{6}\left(\tfrac{3}{4}\right)^d \;=\; m^3\left(\tfrac{9}{\sqrt{6}}\right)^2\left(\tfrac{3}{4}\right)^d \;\leqslant\; m,$$

par le choix de m.

Toutefois s'il n'y a pas plus de m mauvais triplets, alors nous pouvons retirer m des $3m$ vecteurs $\boldsymbol{x}(i)$ de telle sorte que les $2m$ vecteurs restants ne contiennent pas de mauvais triplets, c'est-à-dire qu'ils ne déterminent que des angles aigus. □

La construction « probabiliste » d'un grand ensemble de points de coordonnées 0 ou 1 sans angle droits peut être facilement implémentée à l'aide d'un générateur de nombres aléatoires pour « lancer la pièce ». David Bevan à ainsi construit un ensemble de 31 points déterminant des angles aigus en dimension $d = 15$.

Appendice - Trois outils issus des probabilités

Nous rassemblons ici trois outils fondamentaux issus de la théorie des probabilités discrètes qui apparaîtront plusieurs fois : les variables aléatoires, la linéarité de l'espérance et l'inégalité de Markov.

Soit (Ω, p) un *espace probabilisé* fini, c'est-à-dire que Ω est un ensemble fini et $p = \text{Prob}$ est une application de Ω dans l'intervalle $[0, 1]$ telle que $\sum_{\omega\in\Omega} p(\omega) = 1$. Une *variable aléatoire* X sur Ω est une application $X : \Omega \longrightarrow \mathbb{R}$. Il est d'usage de noter $(X = x)$ l'événement $X^{-1}(\{x\})$ et on a donc un espace probabilisé sur l'ensemble image $X(\Omega)$ en posant $p(X = x) := \sum_{X(\omega)=x} p(\omega)$. Un dé non truqué constitue un exemple simple pour lequel tous les $p(\omega)$ valent $\frac{1}{6}$ et pour lequel $X =$ « le nombre figurant sur la face supérieure du dé lorsqu'il a été lancé ».

L'*espérance* $E(X)$ de X est la moyenne attendue, c'est-à-dire :

$$E(X) \;=\; \sum_{\omega\in\Omega} p(\omega)X(\omega)$$

Supposons à présent que X et Y soient deux variables aléatoires sur Ω.

Leur somme $X+Y$ est encore une variable aléatoire et :

$$\begin{aligned} E(X+Y) &= \sum_{\omega} p(\omega)(X(\omega)+Y(\omega)) \\ &= \sum_{\omega} p(\omega)X(\omega) + \sum_{\omega} p(\omega)Y(\omega) \\ &= E(X)+E(Y) \end{aligned}$$

Il est clair que cela peut s'étendre à une combinaison linéaire finie de variables aléatoires. Cette propriété s'appelle la *linéarité de l'espérance*. Notons que nous n'avons pas besoin d'hypothèse sur « l'indépendance » des variables aléatoires, en quelque sens que ce soit !

Notre troisième outil concerne les variables aléatoires X qui prennent des valeurs non négatives, simplement notées $X \geqslant 0$. Soit :

$$\text{Prob}(X \geqslant a) = \sum_{\omega : X(\omega) \geqslant a} p(\omega)$$

la probabilité que X soit supérieure ou égale à un certain $a > 0$. Alors :

$$E(X) = \sum_{\omega : X(\omega) \geqslant a} p(\omega)X(\omega) + \sum_{\omega : X(\omega) < a} p(\omega)X(\omega) \geqslant a \sum_{\omega : X(\omega) \geqslant a} p(\omega)$$

et nous avons ainsi prouvé l'*inégalité de Markov* :

$$\text{Prob}(X \geqslant a) \leqslant \frac{E(X)}{a} .$$

Bibliographie

[1] L. DANZER & B. GRÜNBAUM : *Über zwei Probleme bezüglich konvexer Körper von P. Erdös und von V. L. Klee,* Math. Zeitschrift **79** (1962), 95-99.

[2] P. ERDŐS & Z. FÜREDI : *The greatest angle among n points in the d-dimensional Euclidean space,* Annals of Discrete Mathematics **17** (1983), 275-283.

[3] H. MINKOWSKI : *Dichteste gitterförmige Lagerung kongruenter Körper,* Nachrichten Ges. Wiss. Göttingen, Math.-Phys. Klasse 1904, 311-355.

La conjecture de Borsuk — Chapitre 16

L'article de Karol Borsuk intitulé « Trois théorèmes sur la sphère euclidienne de dimension n » de 1933 est célèbre parce qu'il contient un résultat important, conjecturé par Stanisław Ulam, connu maintenant sous le nom de théorème de Borsuk-Ulam :

> *Toute application continue $f : S^d \to \mathbb{R}^d$ envoie deux points antipodaux de la sphère S^d sur un même point de $\mathbb{R}^d$.*

Nous pourrons apprécier la puissance de ce résultat dans une application à la théorie des graphes au chapitre 38. Le même article est également célèbre parce qu'à la fin y figure un problème désormais connu sous le nom de conjecture de Borsuk :

Karol Borsuk

> *Tout ensemble $S \subseteq \mathbb{R}^d$ de diamètre borné $diam(S) > 0$ peut-il être partitionné en au plus $d+1$ ensembles de diamètre plus petit ?*

La borne $d+1$ est la meilleure possible : si S est un simplexe régulier de dimension d, ou simplement l'ensemble de ses $d+1$ sommets, alors aucune partie d'une partition réduisant le diamètre ne peut contenir plus d'un des sommets du simplexe. Si $f(d)$ désigne le nombre minimal tel que tout ensemble borné $S \subseteq \mathbb{R}^d$ ait une partition réduisant le diamètre en $f(d)$ parties, alors l'exemple d'un simplexe régulier donne $f(d) \geqslant d+1$.

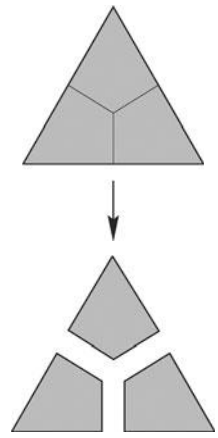
La conjecture de Borsuk a été démontrée par Borsuk lui-même lorsque S est une sphère ainsi que pour les corps lisses en ayant recours au théorème de Borsuk-Ulam pour $d \leqslant 3$ mais la conjecture générale restait ouverte. La meilleure borne supérieure pour $f(d)$ a été établie par Oded Schramm, qui a montré que :

$$f(d) \leqslant (1.23)^d$$

pour tout d suffisamment grand.

Cette borne semble bien faible comparée à la conjecture $f(d) = d+1$ mais elle a subitement semblé raisonnable quand Jeff Kahn et Gil Kalai ont mis en échec de façon spectaculaire la conjecture de Borsuk en 1993. Soixante ans après l'article de Borsuk, Kahn et Kalai ont montré que : $f(d) \geqslant (1.2)^{\sqrt{d}}$ pour d suffisamment grand.

La version de la preuve de Kahn-Kalai pour le Grand Livre a été fournie par A. Nilli : courte et sans références externes, elle donne un contre-exemple explicite à la conjecture de Borsuk en dimension $d = 946$. Nous présentons



Tout d-simplexe *peut* être décomposé en $d+1$ parties, chacune d'entre elles ayant un diamètre plus petit.

ici une version modifiée de cette preuve, due à Andrei M. Raigorodskii et à Bernulf Weißbach, qui réduit la dimension à $d = 561$, et même à $d = 560$. De nouveaux « records » ont été établis par Aicke Hinrichs à $d = 323$ durant l'été 2000, puis par Aicke Hinrichs et Christian Richter à $d = 298$ en 2003.

A. Nilli

Théorème. *Soit $q = p^m$, p étant un entier premier, $n := 4q - 2$, et $d := \binom{n}{2} = (2q-1)(4q-3)$. Alors il existe un ensemble $S \subseteq \{+1, -1\}^d$ de 2^{n-2} points de $\mathbb{R}^d$ tel que : toute partition de S dont les parties ont un diamètre plus petit que celui de S, se compose d'au moins :*

$$\frac{2^{n-2}}{\sum_{i=0}^{q-2} \binom{n-1}{i}}$$

parties. Pour $q = 9$, cela implique que la conjecture de Borsuk est fausse en dimension $d = 561$. En outre, $f(d) > (1.2)^{\sqrt{d}}$ pour d assez grand.

■ **Preuve.** L'ensemble S se construit en quatre étapes.

(1) Soit q un entier premier, soit $n = 4q - 2$ et soit :

$$Q \;:=\; \Big\{ \boldsymbol{x} \in \{+1,-1\}^n : x_1 = 1,\ \text{Card}\{i : x_i = -1\} \text{ est pair} \Big\}$$

Cet ensemble Q se compose de 2^{n-2} vecteurs de $\mathbb{R}^n$. Nous allons voir que l'on a $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \equiv 2 \pmod 4$ pour tous vecteurs $\boldsymbol{x}, \boldsymbol{y} \in Q$. Nous dirons que $\boldsymbol{x}, \boldsymbol{y}$ sont *presque orthogonaux* si $|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| = 2$. Nous allons montrer : que tout sous-ensemble $Q' \subseteq Q$ qui ne contient pas de vecteurs presque orthogonaux doit être « petit » : $|Q'| \leqslant \sum_{i=0}^{q-2} \binom{n-1}{i}$.

$$\boldsymbol{x} \;=\; \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \\ -1 \end{pmatrix} \quad \Longrightarrow$$

$$\boldsymbol{x}^T \;=\; \begin{pmatrix} 1 & -1 & -1 & 1 & -1 \end{pmatrix}$$

$$\boldsymbol{x}\boldsymbol{x}^T = \begin{pmatrix} 1 & -1 & -1 & 1 & -1 \\ -1 & 1 & 1 & -1 & 1 \\ -1 & 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 \\ -1 & 1 & 1 & -1 & 1 \end{pmatrix}$$

Vecteurs, matrices et produits scalaires

Avec les notations retenues, les vecteurs $\boldsymbol{x}, \boldsymbol{y}, \ldots$ sont des vecteurs colonnes ; les vecteurs transposés $\boldsymbol{x}^T, \boldsymbol{y}^T, \ldots$ sont donc des vecteurs lignes. Le produit de matrices $\boldsymbol{x}\boldsymbol{x}^T$ est une matrice de rang 1 et $(\boldsymbol{x}\boldsymbol{x}^T)_{ij} = x_i x_j$.

Si $\boldsymbol{x}, \boldsymbol{y}$ sont des vecteurs colonnes, leur *produit scalaire* est défini par :

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \;=\; \sum_i x_i y_i \;=\; \boldsymbol{x}^T \boldsymbol{y}$$

Nous aurons également besoin de produits scalaires de matrices $X, Y \in \mathbb{R}^{n \times n}$ qui peuvent être interprétées comme des vecteurs de dimension n^2. Leur produit scalaire est défini par :

$$\langle X, Y \rangle \;:=\; \sum_{i,j} x_{ij} y_{ij}$$

(2) À partir de Q, on construit l'ensemble :

$$R \;:=\; \{\boldsymbol{x}\boldsymbol{x}^T : \boldsymbol{x} \in Q\}$$

des 2^{n-2} matrices symétriques $(n \times n)$ de rang 1.Nous les interprétons comme des vecteurs à n^2 composantes, $R \subseteq \mathbb{R}^{n^2}$. Nous allons montrer que ces vecteurs ne forment que des angles aigus : leurs produits scalaires sont positifs et au moins égaux à 4. En outre, si $R' \subseteq R$ ne contient aucun couple de vecteurs dont le produit scalaire est minimal et égal à 4, alors $|R'|$ est « petit » : $|R'| \leqslant \sum_{i=0}^{q-2} \binom{n-1}{i}$.

(3) À partir de R, on construit un ensemble de points de $\mathbb{R}^{\binom{n}{2}}$ dont les coordonnées sont les coefficients qui se trouvent sous les diagonales des matrices correspondantes :

$$S \;:=\; \{(\boldsymbol{x}\boldsymbol{x}^T)_{i>j} : \boldsymbol{x}\boldsymbol{x}^T \in R\}$$

À nouveau, S est un ensemble de 2^{n-2} points. La distance maximale entre ces points est précisément obtenue avec les vecteurs presque orthogonaux $\boldsymbol{x}, \boldsymbol{y} \in Q$. Ainsi, un sous-ensemble $S' \subseteq S$ de diamètre plus petit que celui de S doit être « petit » : $|S'| \leqslant \sum_{i=0}^{q-2} \binom{n-1}{i}$.

(4) Estimations : on déduit de **(3)** qu'il faut au moins :

$$g(q) \;:=\; \frac{2^{4q-4}}{\sum_{i=0}^{q-2} \binom{4q-3}{i}}$$

parties dans chaque partition de S réduisant le diamètre. Ainsi :

$$f(d) \;\geqslant\; \max\{g(q), d+1\} \qquad \text{pour } d \;=\; (2q-1)(4q-3).$$

Par conséquent, chaque fois que $g(q) > (2q-1)(4q-3)+1$, on a un contre-exemple de la conjecture de Borsuk en dimension $d = (2q-1)(4q-3)$.
Nous allons montrer plus loin que $g(9) > 562$, ce qui donne un contre-exemple de dimension $d = 561$, et que :

$$g(q) \;>\; \frac{e}{64\,q^2}\left(\frac{27}{16}\right)^q$$

ce qui conduit à la borne asymptotique $f(d) > (1.2)^{\sqrt{d}}$ pour d suffisamment grand.

Détails de (1) : commençons avec quelques considérations inoffensives de divisibilité.

Lemme. *La fonction $P(z) := \binom{z-2}{q-2}$ est polynomiale de degré $q-2$. Elle prend des valeurs entières pour tous les entiers z. L'entier $P(z)$ est divisible par p si et seulement si z n'est pas congru à 0 ou 1 modulo q.*

Proposition. *Si $a \equiv b \not\equiv 0 \pmod q$, alors a et b ont le même nombre de p-facteurs.*

■ **Preuve.** On a $a = b + sp^m$, où b n'est pas divisible par $p^m = q$. Ainsi, chaque puissance p^k qui divise b vérifie $k < m$; elle divise donc également a. L'énoncé est symétrique en a et b. □

■ **Preuve.** Écrivons le coefficient binomial comme suit :

$$P(z) \;=\; \binom{z-2}{q-2} \;=\; \frac{(z-2)(z-3)\cdot\ldots\cdot(z-q+1)}{(q-2)(q-3)\cdot\;\ldots\;\ldots\cdot 2\cdot 1} \qquad (*)$$

et comparons le nombre de p-facteurs au dénominateur et au numérateur. Le dénominateur a autant de p-facteurs que $(q-2)!$ ou que $(q-1)!$, puisque $q-1$ n'est pas divisible par p. En effet, d'après la proposition figurant dans la marge, on obtient un entier ayant le même nombre de p-facteurs en prenant *n'importe quel* produit de $q-1$ entiers, un dans chaque classe résiduelle non nulle modulo q.

Si $z \equiv 0$ ou $1 \pmod q$, alors le numérateur est aussi de ce type : tous les facteurs du produit proviennent de différentes classes résiduelles ; les seules classes qui n'apparaissent pas sont la classe nulle (les multiples de q) et soit la classe de -1, soit la classe de $+1$. Cependant, ni $+1$ ni -1 ne sont divisibles par p. Donc le dénominateur et le numérateur ont le même nombre de p-facteurs, le quotient n'est donc pas divisible par p.

D'autre part, si $z \not\equiv 0, 1 \pmod q$, alors le numérateur de $(*)$ contient un facteur qui est divisible par $q = p^m$. En même temps, le produit n'a pas de facteurs provenant de deux classes résiduelles adjacentes non nulles : l'une d'elles représente des nombres qui n'ont pas de p-facteurs du tout, l'autre a moins de p-facteurs que $q = p^m$. Ainsi, il y a davantage de p-facteurs au numérateur qu'au dénominateur donc le quotient est divisible par p. □

Considérons maintenant un sous-ensemble arbitraire $Q' \subseteq Q$ qui ne contient pas de vecteurs presque orthogonaux. Nous voulons établir que Q' doit être « petit ».

Proposition 1. *Si $\boldsymbol{x}$ et $\boldsymbol{y}$ sont des vecteurs distincts de Q, alors $\frac14(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + 2)$ est un entier vérifiant l'encadrement :*

$$-(q-2) \;\leqslant\; \tfrac14(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + 2) \;\leqslant\; q-1$$

$\boldsymbol{x}$ et $\boldsymbol{y}$ ont tous deux un nombre pair de composantes égales à (-1), donc le nombre de composantes où $\boldsymbol{x}$ et $\boldsymbol{y}$ diffèrent est également pair. Ainsi,

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle \;=\; (4q-2) \;-\; 2\#\{i : x_i \neq y_i\} \;\equiv\; -2 \pmod 4$$

pour tout $\boldsymbol{x}, \boldsymbol{y} \in Q$, c'est-à-dire que : $\frac14(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + 2)$ est un entier.

Comme $\boldsymbol{x}, \boldsymbol{y} \in \{+1, -1\}^{4q-2}$ nous voyons que : $-(4q-2) \leqslant \langle \boldsymbol{x}, \boldsymbol{y}\rangle \leqslant 4q-2$, c'est-à-dire que : $-(q-1) \leqslant \frac14(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + 2) \leqslant q$. La borne inférieure n'est jamais atteinte, puisque : $x_1 = y_1 = 1$ implique que $\boldsymbol{x} \neq -\boldsymbol{y}$. La borne supérieure est atteinte seulement si $\boldsymbol{x} = \boldsymbol{y}$.

Proposition 2. *Pour tout $\boldsymbol{y} \in Q'$, le polynôme à n variables $x_1, \ldots, x_n$ de degré $q-2$ défini par :*

$$F_{\boldsymbol{y}}(\boldsymbol{x}) \;:=\; P\big(\tfrac14(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + 2)\big) \;=\; \binom{\frac14(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + 2) - 2}{q-2}$$

est tel que $F_{\boldsymbol{y}}(\boldsymbol{x})$ est divisible par p pour tout $\boldsymbol{x} \in Q' \backslash \{\boldsymbol{y}\}$ mais ne l'est pas pour $\boldsymbol{x} = \boldsymbol{y}$.

L'écriture utilisant un coefficient binomial montre que $F_{\boldsymbol{y}}(\boldsymbol{x})$ est un polynôme à valeurs entières. Si $\boldsymbol{x} = \boldsymbol{y}$, on obtient la valeur $F_{\boldsymbol{y}}(\boldsymbol{y}) = 1$. Si $\boldsymbol{x} \neq \boldsymbol{y}$, le lemme implique que $F_{\boldsymbol{y}}(\boldsymbol{x})$ n'est pas divisible par p si et seulement si $\frac{1}{4}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle + 2)$ est congru à 0 ou 1 (mod q). En utilisant le résultat 1, on voit que cela a lieu seulement si $\frac{1}{4}(\langle \boldsymbol{x}, \boldsymbol{y} \rangle + 2)$ est égal à 0 ou 1, c'est-à-dire si $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \in \{-2, +2\}$. $\boldsymbol{x}$ et $\boldsymbol{y}$ doivent donc être presque orthogonaux, ce qui contredit la définition de Q'.

Proposition 3. *Le même résultat est vrai pour les polynômes $\overline{F}_{\boldsymbol{y}}(\boldsymbol{x})$ à $n-1$ variables $x_2, \ldots, x_n$ obtenus comme suit : on décompose $F_{\boldsymbol{y}}(\boldsymbol{x})$ en monômes, on retire la variable x_1, enfin on réduit toutes les puissances plus élevées des autres variables en substituant $x_1 = 1$ et $x_i^2 = 1$ si $i > 1$. Les polynômes $\overline{F}_{\boldsymbol{y}}(\boldsymbol{x})$ ont un degré au plus égal à $q-2$.*

Les vecteurs $\boldsymbol{x} \in Q \subseteq \{+1, -1\}^n$ satisfont tous $x_1 = 1$ et $x_i^2 = 1$. Ainsi, les substitutions ne modifient pas les valeurs des polynômes sur l'ensemble Q. En outre, elles n'augmentent pas le degré donc $\overline{F}_{\boldsymbol{y}}(\boldsymbol{x})$ a un degré au plus égal à $q-2$.

Proposition 4. *Il n'y a pas de relation linéaire (à coefficients rationnels) entre les polynômes $\overline{F}_{\boldsymbol{y}}(\boldsymbol{x})$, c'est-à-dire que les polynômes $\overline{F}_{\boldsymbol{y}}(\boldsymbol{x})$, $\boldsymbol{y} \in Q'$, sont linéairement indépendants sur $\mathbb{Q}$. En particulier, ils sont distincts.*

Supposons qu'il existe une relation de la forme $\sum_{\boldsymbol{y} \in Q'} \alpha_{\boldsymbol{y}} \overline{F}_{\boldsymbol{y}}(\boldsymbol{x}) = 0$ telle que les coefficients $\alpha_{\boldsymbol{y}}$ ne soient pas tous nuls. Après multiplication par un scalaire convenable, nous pouvons supposer que tous les coefficients sont entiers mais que tous ne sont pas divisibles par p. Dans ce cas, pour tout $\boldsymbol{y} \in Q'$ l'évaluation en $\boldsymbol{x} := \boldsymbol{y}$ implique que $\alpha_{\boldsymbol{y}} \overline{F}_{\boldsymbol{y}}(\boldsymbol{y})$ est divisible par p, donc $\alpha_{\boldsymbol{y}}$ l'est puisque $\overline{F}_{\boldsymbol{y}}(\boldsymbol{y})$ ne l'est pas.

Proposition 5. *$|Q'|$ est borné par le nombre de monômes sans carrés de degré au plus $q-2$ en $n-1$ variables, c'est-à-dire par $\sum_{i=0}^{q-2} \binom{n-1}{i}$.*

Par construction, les polynômes $\overline{F}_{\boldsymbol{y}}$ sont sans carré : aucun de leur monôme ne contient de variable de degré supérieur à 1. Ainsi, chaque $\overline{F}_{\boldsymbol{y}}(\boldsymbol{x})$ est une combinaison linéaire de monômes sans carré de degré au plus égal à $q-2$ en $n-1$ variables $x_2, \ldots, x_n$. Puisque les polynômes $\overline{F}_{\boldsymbol{y}}(\boldsymbol{x})$ sont linéairement indépendants, leur nombre (qui est $|Q'|$) ne peut pas être supérieur au nombre de monômes en question.

Détails de (2) : la première colonne de $\boldsymbol{x}\boldsymbol{x}^T$ est $\boldsymbol{x}$. Ainsi, pour des $\boldsymbol{x} \in Q$ distincts, nous obtenons des matrices distinctes $M(\boldsymbol{x}) := \boldsymbol{x}\boldsymbol{x}^T$. Interprétons ces matrices comme des vecteurs de longueurs n^2 de composantes $x_i x_j$.

Un calcul simple :

$$\begin{aligned}\big\langle M(\boldsymbol{x}), M(\boldsymbol{y})\big\rangle &= \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i x_j)(y_i y_j)\\ &= \Big(\sum_{i=1}^{n} x_i y_i\Big)\Big(\sum_{j=1}^{n} x_j y_j\Big) \;=\; \langle \boldsymbol{x}, \boldsymbol{y}\rangle^2 \;\geqslant\; 4\end{aligned}$$

montre que le produit scalaire de $M(\boldsymbol{x})$ et $M(\boldsymbol{y})$ est minimal si et seulement si $\boldsymbol{x}, \boldsymbol{y} \in Q$ sont presque orthogonaux.

Détails de (3) : notons $U(\boldsymbol{x}) \in \{+1, -1\}^d$ le vecteur composé de tous les coefficients qui se trouvent sous la diagonale de $M(\boldsymbol{x})$. Puisque $M(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{x}^T$ est symétrique et puisque les termes de sa diagonale sont $+1$, on voit que : $M(\boldsymbol{x}) \neq M(\boldsymbol{y})$ implique : $U(\boldsymbol{x}) \neq U(\boldsymbol{y})$. En outre :

$$M(\boldsymbol{x}) \;=\; \begin{pmatrix} 1 & & U(\boldsymbol{x}) \\ & 1 & \\ & \ddots & \\ U(\boldsymbol{x}) & & 1 \end{pmatrix}$$

$$4 \;\leqslant\; \langle M(\boldsymbol{x}), M(\boldsymbol{y})\rangle \;=\; 2\langle U(\boldsymbol{x}), U(\boldsymbol{y})\rangle + n$$

c'est-à-dire :

$$\langle U(\boldsymbol{x}), U(\boldsymbol{y})\rangle \;\geqslant\; -\frac{n}{2} + 2$$

l'égalité ayant lieu si et seulement si $\boldsymbol{x}$ et $\boldsymbol{y}$ sont presque orthogonaux. Puisque tous les vecteurs $U(\boldsymbol{x}) \in S$ ont même longueur $\sqrt{\langle U(\boldsymbol{x}), U(\boldsymbol{x})\rangle} = \sqrt{\binom{n}{2}}$, cela signifie que la distance maximale entre les points $U(\boldsymbol{x}), U(\boldsymbol{y}) \in S$ est atteinte exactement lorsque $\boldsymbol{x}$ et $\boldsymbol{y}$ sont presque orthogonaux.

Détails de (4) : Pour $q = 9$ on a $g(9) \approx 758.31$ qui est plus grand que $d + 1 = \binom{34}{2} + 1 = 562$.
Pour obtenir une borne générale pour de grandes valeurs de d, nous utilisons la monotonie des coefficients binomiaux, ainsi que les estimations $n! > e(\frac{n}{e})^n$ et $n! < en(\frac{n}{e})^n$ (se reporter à l'appendice du chapitre 2) et on trouve que :

$$\sum_{i=0}^{q-2}\binom{4q-3}{i} < q\binom{4q}{q} = q\,\frac{(4q)!}{q!(3q)!} < q\,\frac{e\,4q\left(\frac{4q}{e}\right)^{4q}}{e\left(\frac{q}{e}\right)^{q} e\left(\frac{3q}{e}\right)^{3q}} = \frac{4q^2}{e}\Big(\frac{256}{27}\Big)^q$$

Ainsi :

$$f(d) \;\geqslant\; g(q) \;=\; \frac{2^{4q-4}}{\sum_{i=0}^{q-2}\binom{4q-3}{i}} \;>\; \frac{e}{64q^2}\Big(\frac{27}{16}\Big)^q$$

À partir de ce résultat, avec :

$$d \;=\; (2q-1)(4q-3) \;=\; 5q^2 + (q-3)(3q-1) \;\geqslant\; 5q^2 \quad \text{pour } q \geqslant 3,$$

$$q = \tfrac{5}{8} + \sqrt{\tfrac{d}{8} + \tfrac{1}{64}} > \sqrt{\tfrac{d}{8}}$$

et :

$$\left(\tfrac{27}{16}\right)^{\sqrt{\frac{1}{8}}} > 1.2032,$$

on obtient :

$$f(d) > \frac{e}{13d}(1.2032)^{\sqrt{d}} > (1.2)^{\sqrt{d}}$$

pour tout d suffisamment grand. □

On peut obtenir un contre-exemple de dimension 560 en remarquant que pour $q = 9$ le quotient $g(q) \approx 758$ est *beaucoup* plus grand que la dimension $d(q) = 561$. Grâce à cela, on obtient encore un contre-exemple de dimension 560 en prenant uniquement les « trois quarts » de l'ensemble S, définis par les points de Q qui satisfont $(x_1, x_2, x_3) \neq (1, 1, 1)$.

On sait que la conjecture de Borsuk est vraie pour $d \leqslant 3$ mais elle n'a pas été vérifiée pour des dimensions plus grandes. En revanche, elle *est* vraie jusqu'à $d = 8$ si l'on se restreint aux sous-ensembles $S \subseteq \{1, -1\}^d$ construits comme ci-dessus (voir [8]). Il est fort possible que l'on puisse trouver des contre-exemples dans des dimensions raisonnablement petites.

Bibliographie

[1] K. BORSUK : *Drei Sätze über die n-dimensionale euklidische Sphäre,* Fundamenta Math. **20** (1933), 177-190.

[2] A. HINRICHS & C. RICHTER : *New sets with large Borsuk numbers,* Discrete Math. **270** (2003), 137-147.

[3] J. KAHN & G. KALAI : *A counterexample to Borsuk's conjecture,* Bulletin Amer. Math. Soc. **29** (1993), 60-62.

[4] A. NILLI : *On Borsuk's problem,* in : "Jerusalem Combinatorics '93" (H. Barcelo and G. Kalai, eds.), Contemporary Mathematics **178**, Amer. Math. Soc. 1994, 209-210.

[5] A. M. RAIGORODSKII : *On the dimension in Borsuk's problem,* Russian Math. Surveys (6) **52** (1997), 1324-1325.

[6] O. SCHRAMM : *Illuminating sets of constant width,* Mathematika **35** (1988), 180-199.

[7] B. WEISSBACH : *Sets with large Borsuk number,* Beiträge zur Algebra und Geometrie/Contributions to Algebra and Geometry **41** (2000), 417-423.

[8] G. M. ZIEGLER : *Coloring Hamming graphs, optimal binary codes, and the 0/1-Borsuk problem in low dimensions,* Lecture Notes in Computer Science **2122**, Springer-Verlag 2001, 164-175.

Analyse

« L'hôtel de la plage de Hilbert ».

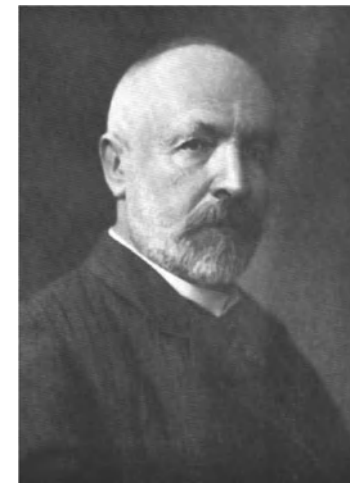
Ensembles, fonctions et hypothèse du continu

Chapitre 17

La théorie des ensembles, élaborée par Georg Cantor dans la seconde moitié du 19^e siècle, a profondément transformé les mathématiques. Les mathématiques modernes sont inconcevables sans le concept d'ensemble et, comme l'a affirmé David Hilbert : « Personne ne nous chassera du paradis (de la théorie des ensembles) que Cantor a créé pour nous ».

L'un des concepts fondamentaux introduits par Cantor est la notion de *taille* ou de *cardinal* d'un ensemble M, noté $|M|$. Pour les ensembles finis, la notion ne présente pas de difficulté : on compte simplement le nombre d'éléments et l'on dit que M est un n-ensemble ou qu'il a n pour cardinal si M contient précisément n éléments. Deux ensembles finis M et N ont donc le même cardinal $|M| = |N|$ s'ils contiennent le même nombre d'éléments.

Pour étendre cette notion d'*égalité* de taille aux ensembles infinis, recourons à l'expérience pratique suivante pour les ensembles finis : supposons qu'un certain nombre de personnes montent dans un bus. Quand peut-on dire que le nombre de personnes est le même que le nombre de sièges libres ? Une façon simple de procéder consiste à laisser tout le monde s'asseoir. Si chacun trouve un siège et si aucun siège ne reste libre, alors le nombre d'éléments de l'ensemble des personnes est le même que le nombre d'éléments de l'ensemble des sièges. En d'autres termes, les deux ensembles ont le même cardinal s'il existe une *bijection* de l'un des ensembles sur l'autre.



Georg Cantor

Nous prendrons donc cette définition : deux ensembles arbitraires M et N (finis ou infinis) ont même *taille* (ou *cardinal*) si et seulement s'il existe une bijection de M sur N. Il est clair que cette notion d'égalité de taille définit une (sorte de) relation d'équivalence sur les ensembles et l'on peut ainsi associer un nombre, appelé *nombre cardinal*, à chaque classe d'ensembles de même cardinal. Pour les ensembles finis, par exemple, on obtient les nombres cardinaux $0, 1, 2, \ldots, n, \ldots$; l'entier k est le cardinal de la classe des k-ensembles et, en particulier, 0 est celui de l'*ensemble vide* $\varnothing$. En outre, on observe le fait évident qu'un sous-ensemble propre d'un ensemble fini M a toujours un cardinal plus petit que celui de M.

La théorie devient très intéressante et nettement moins intuitive lorsqu'on se tourne vers les ensembles infinis. Considérons à cet effet l'ensemble $\mathbb{N}^* = \{1, 2, 3, \ldots\}$ des nombres entiers non nuls. Un ensemble M est dit *dénombrable* s'il peut être mis en bijection [1] avec $\mathbb{N}^*$. En d'autres termes,

1. N.d.T. : nous conservons les définitions du texte original, d'où la mise en bijection avec $\mathbb{N}^*$ et non $\mathbb{N}$ comme le veut l'usage français. L'usage anglo-saxon veut en effet que $\mathbb{N}$ soit défini comme $\{1, 2, 3, \ldots\}$ et non comme $\{0, 1, 2, 3, \ldots\}$. Nous avons choisi de garder la notation française $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ par respect pour les habitudes du lecteur. Il en résulte quelques occurrences de $\mathbb{N}^*$ qui peuvent sembler superflues au premier abord...

M est dénombrable si l'on peut numéroter la liste des éléments de M de la manière suivante : $m_1, m_2, m_3, \ldots$ Toutefois un étrange phénomène se produit alors. Supposons qu'on ajoute à $\mathbb{N}^*$ un nouvel élément x : l'ensemble $\mathbb{N}^* \cup \{x\}$ est toujours dénombrable et son cardinal est donc encore égal à celui de $\mathbb{N}^*$!

Ce résultat est agréablement illustré par l'« hôtel de Hilbert ». Supposons qu'un hôtel dispose d'un ensemble infini dénombrable de chambres numérotées $1, 2, 3, \ldots$ et que chaque chambre i est occupée par un hôte g_i si bien que l'hôtel est complet. À son arrivée, un nouvel hôte x qui demande une chambre se voit répondre : « désolé, toutes les chambres sont réservées ». « Pas de problème », dit le nouvel arrivant, « déplacez simplement l'hôte g_1 dans la chambre 2, l'hôte g_2 dans la chambre 3 et ainsi de suite, puis je prendrai la chambre 1 ». À la surprise du gérant (il n'est pas mathématicien), ce système fonctionne : il parvient à loger tous ses hôtes ainsi que le nouvel arrivant x !

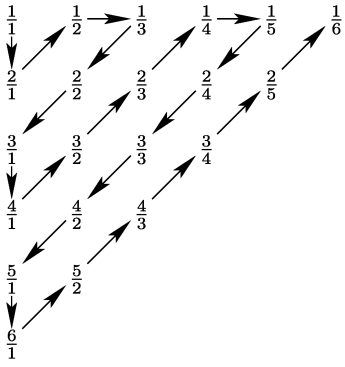
Il est maintenant clair qu'il peut encore loger un nouvel arrivant y, puis encore un autre z, etc. Ainsi, contrairement au cas fini, il peut très bien arriver qu'un sous-ensemble propre d'un ensemble *infini* M ait le même cardinal que M. En fait, comme nous allons le voir, c'est une caractérisation de l'infinité : un ensemble est infini si et seulement s'il a le même cardinal que l'un de ses sous-ensembles propres.

Laissons là l'hôtel de Hilbert et examinons les ensembles de nombres usuels. L'ensemble $\mathbb{Z}$ des entiers est encore dénombrable puisqu'on peut énumérer $\mathbb{Z}$ de la manière suivante $\mathbb{Z} = \{0, 1, -1, 2, -2, 3, -3, \ldots\}$. Il est peut-être plus surprenant que l'ensemble $\mathbb{Q}$ des nombres rationnels soit aussi dénombrable.

Théorème 1. *L'ensemble $\mathbb{Q}$ des nombres rationnels est dénombrable.*

■ **Preuve.**

En listant l'ensemble $\mathbb{Q}^+$ des rationnels positifs comme indiqué sur la figure ci-contre, tout en éliminant les nombres déjà rencontrés, on observe que $\mathbb{Q}^+$ est dénombrable ; $\mathbb{Q}$ l'est donc aussi puisqu'il suffit d'en faire la liste en commençant par 0 et en plaçant $-\frac{p}{q}$ juste après $\frac{p}{q}$. Avec ce procédé d'énumération :



$$\mathbb{Q} = \{0, 1, -1, 2, -2, \tfrac{1}{2}, -\tfrac{1}{2}, \tfrac{1}{3}, -\tfrac{1}{3}, 3, -3, 4, -4, \tfrac{3}{2}, -\tfrac{3}{2}, \ldots\}$$

□

On peut aussi interpréter la figure autrement et dire que :

> *Une réunion dénombrable d'ensembles dénombrables M_n est dénombrable.*

En effet, il suffit de poser $M_n = \{a_{n1}, a_{n2}, a_{n3}, \ldots\}$ et de lister :

$$\bigcup_{n=1}^{\infty} M_n = \{a_{11}, a_{21}, a_{12}, a_{13}, a_{22}, a_{31}, a_{41}, a_{32}, a_{23}, a_{14}, \ldots\}$$

exactement comme précédemment.

Observons l'énumération des rationnels positifs selon le schéma de Cantor un peu plus attentivement. En considérant la figure, nous avions obtenu la suite :

$$\tfrac{1}{1}, \tfrac{2}{1}, \tfrac{1}{2}, \tfrac{1}{3}, \tfrac{2}{2}, \tfrac{3}{1}, \tfrac{4}{1}, \tfrac{3}{2}, \tfrac{2}{3}, \tfrac{1}{4}, \tfrac{1}{5}, \tfrac{2}{4}, \tfrac{3}{3}, \tfrac{4}{2}, \tfrac{5}{1}, \ldots$$

de laquelle il convenait de retirer les doublons comme $\frac{2}{2} = \frac{1}{1}$ ou $\frac{2}{4} = \frac{1}{2}$.
En fait il existe une manière encore plus élégante et systématique de lister les éléments en question qui évite les doublons. Elle a été proposée très récemment par Neil Calkin et Herbert Wilf. Leur liste commence par :

$$\tfrac{1}{1}, \tfrac{1}{2}, \tfrac{2}{1}, \tfrac{1}{3}, \tfrac{3}{2}, \tfrac{2}{3}, \tfrac{3}{1}, \tfrac{1}{4}, \tfrac{4}{3}, \tfrac{3}{5}, \tfrac{5}{2}, \tfrac{2}{5}, \tfrac{5}{3}, \tfrac{3}{4}, \tfrac{4}{1}, \ldots .$$

Le dénominateur du n-ième rationnel de la suite est égal au numérateur du $(n+1)$-ième. En d'autres termes, la n-ième fraction est de la forme $b(n)/b(n+1)$ où $\big(b(n)\big)_{n\geqslant 0}$ est une suite qui commence par :

$$(1, 1, 2, 1, 3, 2, 3, 1, 4, 3, 5, 2, 5, 3, 4, 1, 5, \ldots).$$

Dans un article de 1858, le mathématicien allemand Moritz Abraham Stern est le premier à s'être intéressé à cette suite, désormais connue sous le nom de *suite diatomique de Stern*.

Comment obtenir les éléments de cette suite et ainsi une manière efficace de lister les rationnels positifs ? Considérons l'arbre binaire infini représenté dans la marge. Il est facile de constater la nature récursive de sa construction :

– $\frac{1}{1}$ figure à la racine de l'arbre ;
– chaque nœud $\frac{i}{j}$ a deux fils : le fils gauche est $\frac{i}{i+j}$ et le fils droit est $\frac{i+j}{j}$.

On vérifie facilement les propriétés suivantes :

(1) Toutes les fractions qui apparaissent dans l'arbre sont réduites, c'est-à-dire que lorsqu'une fraction $\frac{r}{s}$ apparaît dans l'arbre, r et s sont premiers entre eux.

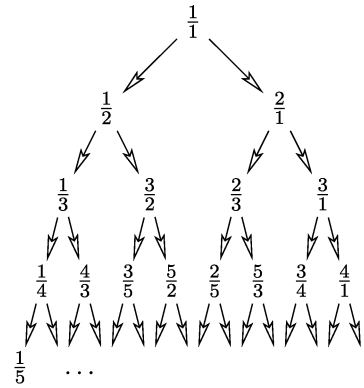
La propriété est vraie pour la racine de l'arbre $\frac{1}{1}$ et l'on va procéder par récurrence vers le bas de l'arbre. Si r et s sont premiers entre eux, alors il en va de même pour r et $r+s$, comme pour s et $r+s$.

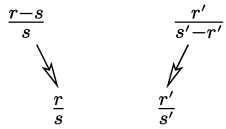
(2) Toute fraction réduite $\frac{r}{s} > 0$ apparaît dans l'arbre.

On procède par récurrence sur la somme $r+s$. La plus petite valeur possible pour cette somme est $r+s=2$, correspondant à $\frac{r}{s} = \frac{1}{1}$, valeur qui apparaît à la racine de l'arbre. Si $r > s$, alors, d'après l'hypothèse de récurrence, $\frac{r-s}{s}$ apparaît déjà dans l'arbre et l'on obtient $\frac{r}{s}$ comme son fils droit. De manière analogue, si $r < s$, $\frac{r}{s-r}$ apparaît déjà dans l'arbre et $\frac{r}{s}$ apparaît comme son fils gauche.

(3) Toute fraction réduite apparaît exactement une fois.

Si $\frac{r}{s}$ apparaissait plus d'une fois dans l'arbre, alors $r \neq s$ puisque tout nœud de l'arbre, à l'exception de la racine, est de la forme $\frac{i}{i+j} < 1$ ou $\frac{i+j}{j} > 1$. Comme $r > s$ ou $s > r$, on procède par récurrence, comme on l'a fait dans le point précédent.





Tout rationnel positif apparaît donc exactement une fois dans l'arbre et on peut écrire la suite de ces nombres de gauche à droite, niveau par niveau, en descendant dans l'arbre[2]. On obtient ainsi les premiers termes listés précédemment.

(4) Le dénominateur de la n-ième fraction de la suite est égal au numérateur de la $(n+1)$-ième.

Le résultat est évident pour $n = 0$ ou lorsque la n-ième fraction est un fils gauche. On raisonne par récurrence. Supposons, que la n-ième fraction $\frac{r}{s}$ est un fils droit. Si $\frac{r}{s}$ est l'élément le plus à droite d'un niveau, alors $s = 1$ si bien que la fraction qui lui succède se trouve la plus à gauche dans le niveau suivant et présente un numérateur qui est 1. Si, en revanche, $\frac{r}{s}$ n'est pas au bord de son niveau et, si $\frac{r'}{s'}$ désigne la fraction qui la suit immédiatement, alors $\frac{r}{s}$ est le fils droit de $\frac{r-s}{s}$ et $\frac{r'}{s'}$ est le fils gauche de $\frac{r'}{s'-r'}$ et alors, d'après l'hypothèse de récurrence, le dénominateur de $\frac{r-s}{s}$ est le numérateur de $\frac{r'}{s'-r'}$ si bien que $s = r'$.

Tout cela est très bien mais il y a encore mieux. Deux questions se présentent de manière naturelle :

- Est-ce que la suite $\big(b(n)\big)_{n\geqslant 0}$ a une « signification » ? En d'autres termes, est-ce que $b(n)$ dénombre quelque chose de simple ?
- Étant donné $\frac{r}{s}$, dispose-t-on d'un moyen simple pour calculer son successeur dans la suite ?

Pour répondre à la première question, nous allons montrer que le nœud $b(n)/b(n+1)$ admet les fils $b(2n+1)/b(2n+2)$ et $b(2n+2)/b(2n+3)$. La procédé de construction de l'arbre conduit à la récurrence :

$$b(2n+1) = b(n) \quad \text{et} \quad b(2n+2) = b(n) + b(n+1). \tag{1}$$

Posant $b(0) = 1$, la suite $(b(n))_{n\geqslant 0}$ est parfaitement définie par les relations (1).

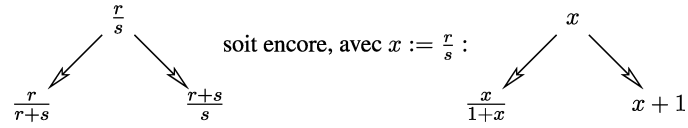
Ainsi la question devient : y a-t-il une « jolie » suite « connue » satisfaisant à ces relations de récurrence ? La réponse est affirmative. Nous savons que tout nombre n peut être écrit de manière unique comme la somme de puissances de 2 distinctes : c'est l'habituelle représentation binaire de n. Une représentation *hyperbinaire* de n est une décomposition de n en somme de puissances de 2 dans laquelle chaque puissance peut apparaître au plus *deux fois*. Soit $h(n)$ le nombre de représentations hyperbinaires possibles pour n. Le lecteur est invité à constater que la suite de terme général $h(n)$ satisfait aux relations de récurrence (1), ce qui suffit à prouver que $b(n) = h(n)$ pour tout n.

Par exemple, $h(6) = 3$, avec les représentations hyperbinaires possibles suivantes :
$6 = 4 + 2$
$6 = 4 + 1 + 1$
$6 = 2 + 2 + 1 + 1$.

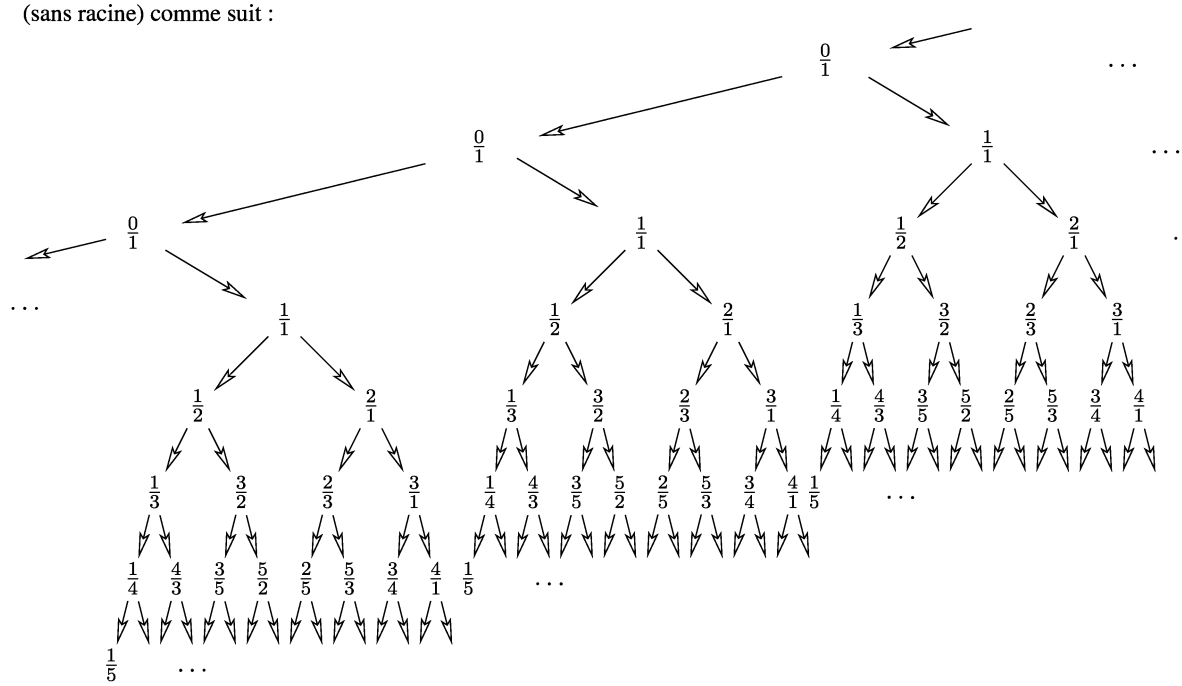
Incidemment nous avons prouvé un résultat surprenant : soit $\frac{r}{s}$ une fraction réduite, il existe un unique entier n tel que $r = h(n)$ et $s = h(n+1)$.

2. N.d.T. : l'informaticien dirait : « en effectuant un parcours en largeur » de l'arbre.

Tournons nous à présent vers la deuxième question. Dans l'arbre figure :



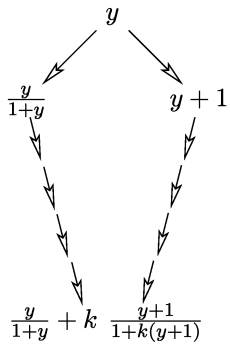
Nous utilisons ce schéma pour générer un arbre infini encore plus grand (sans racine) comme suit :



Dans cet arbre, toutes les lignes sont identiques et elles fournissent toutes le procédé de dénombrement des rationnels positifs de Calkin-Wilf (en partant de $\frac{0}{1}$).

Comment fait-on alors pour passer d'un rationnel à son successeur ? Pour répondre à cette question, remarquons d'abord que pour tout rationnel x, son fils droit est $x+1$, son petit-fils droit est $x+2$ et son descendant droit d'ordre k est $x+k$. De manière analogue, le fils gauche de x est $\frac{x}{1+x}$ dont le propre fils gauche est $\frac{x}{1+2x}$ et ainsi de suite. Le descendant gauche d'ordre k de x est $\frac{x}{1+kx}$.

À présent, pour trouver comment on passe de $\frac{r}{s} = x$ à son successeur $f(x)$ dans la liste, il nous faut analyser la situation représentée dans la marge. En fait, si l'on considère un rationnel x strictement positif quelconque dans l'arbre binaire infini, alors on voit que c'est le descendant droit d'ordre k du fils gauche d'un certain rationnel $y \geqslant 0$ (pour un certain $k \geqslant 0$) tandis que $f(x)$ s'obtient comme le descendant gauche d'ordre k du fils droit de



ce même y. Ainsi, à partir de la formule donnant la valeur des descendants d'ordre k on trouve :

$$x \;=\; \frac{y}{1+y} + k,$$

comme indiqué sur la figure ci-contre. Ici, $k = \lfloor x \rfloor$ est la partie entière de x tandis que $\frac{y}{1+y} = \{x\}$ est sa partie fractionnaire. On en déduit :

$$f(x) = \frac{y+1}{1+k(y+1)} = \frac{1}{\frac{1}{y+1}+k} = \frac{1}{k+1-\frac{y}{y+1}} = \frac{1}{\lfloor x \rfloor + 1 - \{x\}}\,.$$

Nous avons donc obtenu une magnifique formule pour le successeur $f(x)$ de x, récemment établie par Moshe Newman :

La fonction :

$$x \;\longmapsto\; f(x) \;=\; \frac{1}{\lfloor x \rfloor + 1 - \{x\}}$$

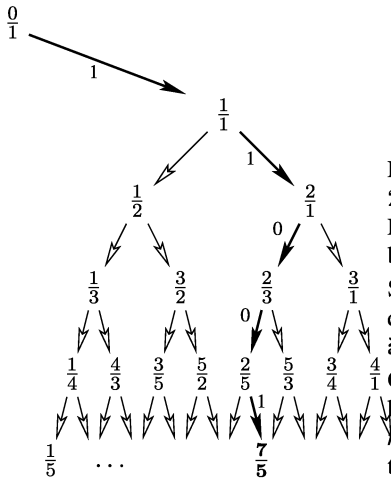
engendre la suite de Calkin-Wilf.

$$\tfrac{1}{1} \mapsto \tfrac{1}{2} \mapsto \tfrac{2}{1} \mapsto \tfrac{1}{3} \mapsto \tfrac{3}{2} \mapsto \tfrac{2}{3} \mapsto \tfrac{3}{1} \mapsto \tfrac{1}{4} \mapsto \tfrac{4}{3} \mapsto \ldots$$

dans laquelle figure exactement une fois tout rationnel positif.

Le procédé d'énumération de Calkin-Wilf-Newman présente d'autres propriétés remarquables. Par exemple, on peut se demander si l'on dispose d'un moyen rapide de calculer la n-ième fraction de la suite, disons, par exemple, pour $n = 10^6$. Le voici :

> Pour trouver la n-ième fraction de la suite de Calkin-Wilf, il faut exprimer n sous sa forme binaire $n = (b_k b_{k-1} \ldots b_1 b_0)_2$ et suivre dans l'arbre de Calkin-Wilf le chemin défini par les chiffres b_i, en partant de $\frac{s}{t} = \frac{0}{1}$. L'égalité $b_i = 1$ signifie « choisir le fils droit », c'est-à-dire « ajouter le dénominateur au numérateur », tandis que $b_i = 0$ signifie « choisir le fils gauche », c'est-à-dire « ajouter le numérateur au dénominateur ».



La figure représentée ci-contre dans la marge montre le chemin pour $n = 25 = (11001)_2$. Ainsi le 25-ième terme de la suite de Calkin-Wilf est $\frac{7}{5}$. Le lecteur pourra aisément établir un procédé qui calcule la (représentation binaire de la) position n d'une fraction donnée $\frac{s}{t}$ dans la suite.

Signalons, pour en finir avec $\mathbb{Q}$, une autre façon, très élégante, de montrer que $\mathbb{Q}^+$ est dénombrable : elle consiste à observer que l'application de $\mathbb{Q}^+$ à valeurs dans $\mathbb{N}$ qui à p/q irréductible associe $2^p 3^q$ est injective.

Qu'en est-il de l'ensemble des nombres réels $\mathbb{R}$? Est-il lui-aussi dénombrable ? Il n'en est rien. La méthode employée pour le démontrer — la *méthode diagonale* de Cantor — n'a pas seulement une importance capitale pour la théorie des ensembles toute entière, elle appartient aussi à coup sûr au Grand Livre pour le trait de génie qu'elle constitue.

Théorème 2. *L'ensemble $\mathbb{R}$ des nombres réels n'est pas dénombrable.*

■ **Preuve.** Tout sous-ensemble N d'un ensemble dénombrable $M = \{m_1, m_2, m_3, \ldots\}$ est *au plus dénombrable* (c'est-à-dire fini ou dénombrable). En effet, il suffit de lister simplement les éléments de N tels qu'ils apparaissent dans M. Si donc on peut trouver un sous-ensemble de $\mathbb{R}$ qui n'est pas dénombrable, alors *a fortiori* $\mathbb{R}$ ne peut pas être dénombrable. Le sous-ensemble M de $\mathbb{R}$ que nous allons examiner est l'intervalle $]0, 1]$ de tous les nombres réels r tels que $0 < r \leqslant 1$. Supposons, au contraire, que M soit dénombrable et soit $M = \{r_1, r_2, r_3, \ldots\}$ une énumération des éléments de M. On écrit r_n sous la forme de son unique développement décimal *infini* (sans suite infinie de zéros à la fin [3]) :

$$r_n = 0.a_{n1}a_{n2}a_{n3}\ldots$$

où $a_{ni} \in \{0, 1, \ldots, 9\}$ pour tout n et pour tout i. Par exemple, $0.7 = 0.6999\ldots$ Considérons maintenant le tableau doublement infini :

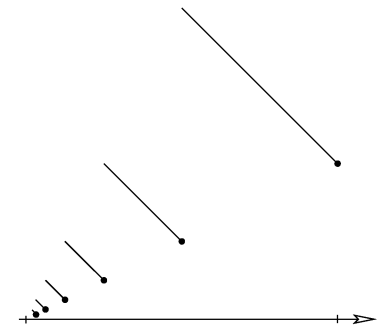
$$\begin{array}{rcl} r_1 & = & 0.a_{11}a_{12}a_{13}... \\ r_2 & = & 0.a_{21}a_{22}a_{23}... \\ & \vdots & \\ r_n & = & 0.a_{n1}a_{n2}a_{n3}... \\ & \vdots & \end{array}$$

Pour tout n, soit b_n le plus petit élément de $\{1, 2\}$ différent de a_{nn}. Alors $b = 0.b_1b_2b_3...b_n...$ est un nombre réel appartenant à l'ensemble M qui doit donc être indexé sous la forme $b = r_k$. Cela est impossible puisque b_k est différent de a_{kk}. □

Attardons-nous un moment sur les nombres réels et remarquons que les quatre types d'intervalles $]0, 1[,]0, 1], [0, 1[$ et $[0, 1]$ ont le même cardinal. Nous pouvons notamment vérifier que $]0, 1]$ et $]0, 1[$ ont même cardinal. L'application f définie par $f :]0, 1] \longrightarrow]0, 1[, x \longmapsto y$ par :

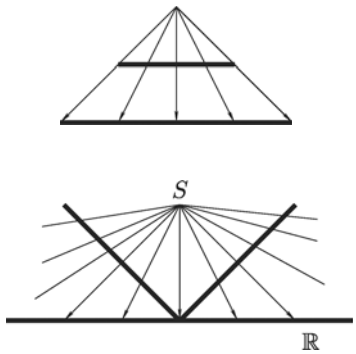
$$y := \begin{cases} \frac{3}{2} - x & \text{si} \quad \frac{1}{2} < x \leqslant 1, \\ \frac{3}{4} - x & \text{si} \quad \frac{1}{4} < x \leqslant \frac{1}{2}, \\ \frac{3}{8} - x & \text{si} \quad \frac{1}{8} < x \leqslant \frac{1}{4}, \\ \quad \vdots & \end{cases}$$

convient. En effet, cette application est bijective, puisque y décrit l'intervalle $\frac{1}{2} \leqslant y < 1$ à la première ligne, $\frac{1}{4} \leqslant y < \frac{1}{2}$ à la deuxième ligne, $\frac{1}{8} \leqslant y < \frac{1}{4}$ à la troisième ligne et ainsi de suite.



$f :]0, 1] \longrightarrow]0, 1[$

3. N.d.T. : c'est-à-dire en faisant le choix (justifié par le raisonnement qui suit) de représenter, par exemple, $3/10$ par le développement $0,29999...$ plutôt que par le développement $0,30000...$



Nous constatons ensuite que *tous* les intervalles (de longueur finie strictement positive) ont même cardinal en considérant la projection centrale représentée sur la figure. Il y a même mieux : chaque intervalle (de longueur strictement positive) a le même cardinal que la droite réelle $\mathbb{R}$ entière. Pour s'en convaincre, il suffit de regarder l'intervalle ouvert plié $]0,1[$ et de le projeter sur $\mathbb{R}$ à partir du centre S.

En conclusion, tous les intervalles ouverts, semi-ouverts, fermés (finis ou infinis) de longueur strictement positive ont le même cardinal. On note c ce cardinal, c pour *continu*. On dit que ces ensembles de nombres ont la puissance du continu.

Le fait que des intervalles finis ou infinis aient le même cardinal peut apparaître comme intuitif après réflexion ; voici néanmoins un résultat qui va complètement à l'encontre de l'intuition.

Théorème 3. *L'ensemble $\mathbb{R}^2$ de tous les couples de nombres réels, (c'est-à-dire le plan réel) a le même cardinal que $\mathbb{R}$.*

Cantor proposa cet énoncé en 1878 avec l'idée de fusionner le développement décimal de deux réels pour en faire un seul. La variante de la méthode de Cantor présentée ici relève du Grand Livre. Abraham Fraenkel attribue à Julius König l'astuce qui fournit directement la bijection adéquate.

■ **Preuve.** Il suffit de montrer que l'ensemble de tous les couples (x,y), $0 < x,y \leqslant 1$, peut être envoyé de façon bijective sur $]0,1]$. Ici encore, la démonstration mérite de figurer dans le Grand Livre. Considérons le couple (x,y) et écrivons l'unique développement décimal infini de x et y comme dans l'exemple suivant :

$$\begin{array}{rcccccccc} x & = & 0.3 & & 01 & 2 & 007 & 08 & \ldots \\ y & = & 0.009 & 2 & 05 & 1 & 0008 & & \ldots \end{array}$$

Nous avons séparé les chiffres de x et y en groupes qui s'arrêtent dès qu'un chiffre non nul apparaît dans le développement. Nous associons ensuite à (x,y) le nombre $z \in]0,1]$ en écrivant le premier groupe de chiffres apparaissant dans l'écriture de x, puis le premier groupe issu de y, ensuite le deuxième groupe issu de x et ainsi de suite. Ainsi, avec l'exemple précédent, on obtient :

$$z = 0.3\ 009\ 01\ 2\ 2\ 05\ 007\ 1\ 08\ 0008\ \ldots$$

Comme ni x ni y ne sont composés que de zéros à partir d'un certain rang, l'expression de z est un développement décimal infini. Réciproquement, à partir du développement de z, nous pouvons immédiatement reconstituer son image réciproque (x,y), l'application est donc bijective. □

Comme $(x,y) \longmapsto x + iy$ est une bijection de $\mathbb{R}^2$ sur l'ensemble des nombres complexes $\mathbb{C}$, on obtient : $|\mathbb{C}| = |\mathbb{R}| = c$. Pourquoi le résultat $|\mathbb{R}^2| = |\mathbb{R}|$ est-il aussi inattendu ? Parce qu'il va à l'encontre de l'intuition qu'on peut avoir de la notion de *dimension*. Le résultat précédent dit que le plan $\mathbb{R}^2$ de dimension 2 (et plus généralement, par récurrence, l'espace

$\mathbb{R}^n$ de dimension n) peut être envoyé bijectivement sur la droite $\mathbb{R}$ qui est de dimension 1. Ainsi, la dimension n'est pas conservée en général par les applications bijectives. Cependant, si nous demandons à l'application et à son inverse d'être continues, alors la dimension est conservée, propriété que Luitzen Brouwer fut le premier à montrer.

Allons un peu plus loin. Jusqu'à présent, nous avons la notion d'égalité de cardinal. Quand pourra-t-on dire que M est au plus aussi grand que N ? Ce sont encore les applications qui apportent la clé. Nous dirons que le nombre cardinal $\mathfrak{m}$ est *inférieur ou égal à* $\mathfrak{n}$ si, étant donnés des ensembles M et N tels que $|M| = \mathfrak{m}$ et $|N| = \mathfrak{n}$, il existe une *injection* de M dans N. Il est clair que la relation $\mathfrak{m} \leqslant \mathfrak{n}$ est indépendante des représentants M et N choisis. Si les ensembles sont finis, cela correspond encore à l'intuition. Un m-ensemble est au plus aussi grand qu'un n-ensemble si et seulement si $m \leqslant n$.

On est maintenant confronté à un problème fondamental. Il serait évidemment souhaitable que les lois usuelles concernant les inégalités soient également vérifiées pour les nombres cardinaux (notamment infinis). En particulier, est-il vrai que $\mathfrak{m} \leqslant \mathfrak{n}$ et $\mathfrak{n} \leqslant \mathfrak{m}$ implique : $\mathfrak{m} = \mathfrak{n}$?

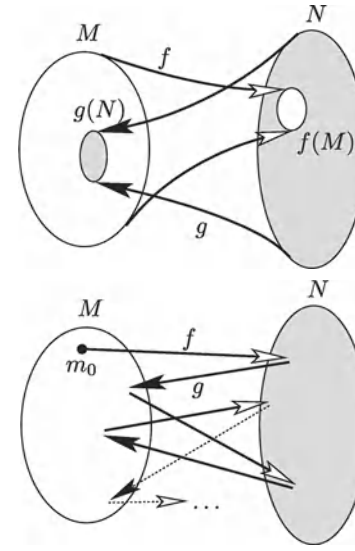
La réponse affirmative est fournie par le célèbrethéorème de Cantor-Bernstein que Cantor énonça en 1883. La première démonstration complète de ce résultat fut établie par Felix Bernstein bien des années plus tard dans le séminaire de Cantor. De nouvelles preuves ont été établies par Richard Dedekind, Ernst Zermelo et bien d'autres. La démonstration que nous en donnons ici est due à Julius König (1906).

Théorème 4. *Si chacun des deux ensembles M et N s'injecte dans l'autre, alors il existe une bijection de M sur N, c'est-à-dire $|M| = |N|$.*

■ **Preuve.** Nous pouvons supposer que M et N sont disjoints — si tel n'est pas le cas, on remplace N par une copie de N.

Soit f et g les applications qui transportent respectivement les éléments de M vers N et ceux de N vers M. Une manière d'éclaircir la situation consiste à aligner les éléments de $M \cup N$ en une chaîne. On va prendre un élément arbitraire m_0 de M ; on forme une chaîne à partir de cet élément en lui appliquant f, puis en appliquant g à son image, puis à nouveau f et ainsi de suite. La chaîne peut se refermer (cas 1) si l'on retombe sur m_0 au cours de l'itération de ce procédé ou, au contraire, être constituée d'une suite infinie d'éléments distincts (le premier élément qui se répéterait dans la chaîne ne pourrait être que m_0 par injectivité des applications considérées).

Si la chaîne est infinie, alors on va essayer de la parcourir en marche arrière, en passant de m_0 à $g^{-1}(m_0)$ si m_0 appartient à l'image de g, puis à $f^{-1}(g^{-1}(m_0))$ si $g^{-1}(m_0)$ appartient à l'image de f et ainsi de suite. Trois nouveaux cas peuvent alors se présenter. Soit le parcours rétrograde de la chaîne est infini (cas 2), soit le parcours s'interrompt en un élément de M ne figurant pas dans l'image de g (cas 3), soit le parcours s'interrompt en un élément de N ne figurant pas dans l'image de f (cas 4).



Ainsi, quatre cas peuvent se présenter pour les chaînes de $M \cup N$ et nous allons voir que dans chaque cas il est possible d'indexer les éléments de la chaîne de telle sorte que $F : m_i \longmapsto n_i$ soit une bijection.

Cas 1 : cycle fini constitué de $2k + 2$ éléments distincts ($k \geqslant 0$).

$$m_0 \xrightarrow{f} n_0 \xrightarrow{g} m_1 \xrightarrow{f} \cdots \quad m_k \xrightarrow{f} n_k \xrightarrow{g} m_0$$

Cas 2 : chaîne infinie des deux côtés constituée d'éléments distincts.

$$\cdots \longrightarrow m_0 \xrightarrow{f} n_0 \xrightarrow{g} m_1 \xrightarrow{f} n_1 \xrightarrow{g} m_2 \xrightarrow{f} \cdots$$

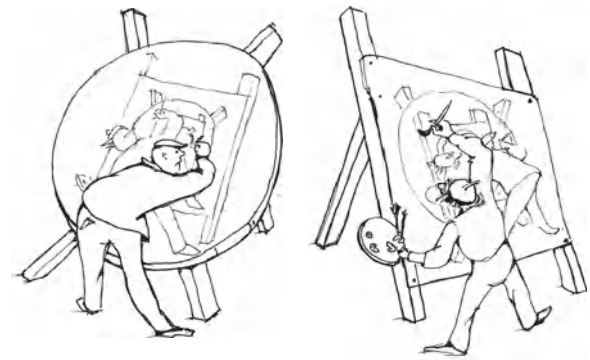
Cas 3 : chaîne infinie constituée d'éléments distincts commençant en $m_0 \in M \backslash g(N)$.

$$m_0 \xrightarrow{f} n_0 \xrightarrow{g} m_1 \xrightarrow{f} n_1 \xrightarrow{g} m_2 \xrightarrow{f} \cdots$$

Cas 4 : chaîne infinie constituée d'éléments distincts commençant en $n_0 \in N \backslash f(M)$.

$$n_0 \xrightarrow{g} m_0 \xrightarrow{f} n_1 \xrightarrow{g} m_1 \xrightarrow{f} \cdots$$

□



« Cantor et Bernstein en train de peindre ».

Qu'en est-il des autres relations gouvernant les inégalités ? Comme d'habitude, nous posons $\mathfrak{m} < \mathfrak{n}$ si $\mathfrak{m} \leqslant \mathfrak{n}$ et $\mathfrak{m} \neq \mathfrak{n}$. Nous venons de voir qu'étant donnés deux cardinaux $\mathfrak{m}$ et $\mathfrak{n}$ l'une au plus des trois possibilités :

$$\mathfrak{m} < \mathfrak{n},\ \mathfrak{m} = \mathfrak{n},\ \mathfrak{m} > \mathfrak{n}$$

peut se produire. Il découle de la théorie des cardinaux qu'en fait, une relation exactement est vraie (se reporter à l'appendice de ce chapitre, proposition 2).

En outre, le théorème de Schroeder-Bernstein nous dit que la relation $<$ est transitive, c'est-à-dire que $\mathfrak{m} < \mathfrak{n}$ et $\mathfrak{n} < \mathfrak{p}$ impliquent $\mathfrak{m} < \mathfrak{p}$. Ainsi,

les cardinaux sont placés dans un ordre linéaire qui commence avec les cardinaux finis $0, 1, 2, 3, \ldots$ En faisant appel au système axiomatique de Zermelo-Fraenkel (notamment à l'axiome du choix), on établit facilement qu'un ensemble infini M contient un sous-ensemble dénombrable. En effet, M contient un élément, soit m_1. L'ensemble $M \setminus \{m_1\}$ n'est pas vide (puisque M est infini) et il contient donc un élément m_2. De la même façon, $M \setminus \{m_1, m_2\}$ contient un élément m_3 et ainsi de suite. Ainsi, le cardinal d'un ensemble infini dénombrable est *le plus petit cardinal infini* habituellement noté $\aleph_0$ (prononcer « aleph zéro »).

Comme $\aleph_0 \leqslant \mathfrak{m}$ pour tout cardinal infini $\mathfrak{m}$, on obtient immédiatement le résultat de l'« hôtel de Hilbert » pour tout nombre cardinal infini $\mathfrak{m}$, c'est-à-dire qu'on a $|M \cup \{x\}| = |M|$ pour tout ensemble infini M. En effet, M contient un sous-ensemble $N = \{m_1, m_2, m_3, \ldots\}$. Maintenant, envoyons x sur m_1, m_1 sur m_2 et ainsi de suite, en laissant fixes les éléments de $M \setminus N$. Ce procédé fournit la bijection souhaitée.

On a montré par la même occasion un résultat annoncé un peu plus tôt : *Tout ensemble infini a le même cardinal que l'un de ses sous-ensembles propres.*

Une autre application du théorème de Cantor-Bernstein consiste à prouver que l'ensemble $\mathcal{P}(\mathbb{N})$ de tous les sous-ensembles de $\mathbb{N}$ est de cardinal c. Comme on l'a dit précédemment, il suffit de montrer que $|\mathcal{P}(\mathbb{N}) \setminus \{\emptyset\}| = |]0, 1]|$. L'application f et g suivantes sont injectives et suffisent à prouver cette égalité :

$$\begin{aligned} f : \mathcal{P}(\mathbb{N}) \setminus \{\emptyset\} &\longrightarrow]0, 1], \\ A &\longmapsto \sum_{i \in A} 10^{-i}, \end{aligned}$$

$$\begin{aligned} g :]0, 1] &\longrightarrow \mathcal{P}(\mathbb{N}) \setminus \{\emptyset\}, \\ 0.b_1 b_2 b_3 \ldots &\longmapsto \{b_i 10^i : i \in \mathbb{N}\} \end{aligned}$$

Jusqu'à présent, nous avons rencontré les cardinaux $0, 1, 2, \ldots, \aleph_0$; nous savons aussi que le cardinal c de $\mathbb{R}$ est plus grand que $\aleph_0$. Le passage de $\mathbb{Q}$ de cardinal $\aleph_0$ à $\mathbb{R}$ de cardinal c suggère immédiatement la question suivante :

Le cardinal infini $c = |\mathbb{R}|$ est-il celui qui suit immédiatement $\aleph_0$?

À présent nous sommes confrontés à la question de savoir s'il *existe* un nombre cardinal plus grand suivant $\aleph_0$ ou, en d'autres termes, s'il est pertinent d'introduire un nouveau cardinal $\aleph_1$. On peut répondre par l'affirmative et la preuve est esquissée dans l'appendice de ce chapitre.

L'énoncé $c = \aleph_1$ est connu sous le nom d'*hypothèse du continu*. La question de savoir si l'hypothèse du continu est vraie a constitué pendant plusieurs décennies l'un des plus grands défis de toutes les mathématiques. La réponse, finalement donnée par Kurt Gödel et Paul Cohen, nous mène aux

« Le plus petit cardinal infini ».

limites de la pensée logique. Ils ont montré que l'énoncé $c = \aleph_1$ est *indépendant* du système d'axiomes de Zermelo-Fraenkel, tout comme l'axiome de l'unique parallèle est indépendant des autres axiomes de la géométrie euclidienne. Il y a des modèles pour lesquels $c = \aleph_1$ et d'autres modèles de la théorie des ensembles pour lesquels $c \neq \aleph_1$.

À la lumière de ces remarques, il est intéressant de se demander s'il existe d'autres conditions (provenant par exemple de l'analyse) qui sont équivalentes à l'hypothèse du continu. Dans ce qui suit, nous voulons présenter un exemple de ce type ainsi que sa solution, très élégante et très simple, due à Paul Erdős. En 1962, Wetzel a posé la question suivante :

> *Soit $\{f_\alpha\}$ une famille de fonctions analytiques sur $\mathbb{C}$, deux à deux distinctes, telles que pour tout $z \in \mathbb{C}$ l'ensemble des valeurs $\{f_\alpha(z)\}$ soit dénombrable ; appelons (P_0) cette propriété.*
> *Peut-on en déduire que cette famille est elle-même dénombrable ?*

Très peu de temps après, Erdős a montré de façon surprenante que la réponse dépend de l'hypothèse du continu.

Théorème 5. *Si $c > \aleph_1$, alors toute famille $\{f_\alpha\}$ vérifiant (P_0) est dénombrable. Si, en revanche, $c = \aleph_1$, alors il existe une famille $\{f_\alpha\}$ de cardinal c vérifiant la propriété (P_0).*

La démonstration de ce théorème fait appel à certains résultats élémentaires sur les nombres cardinaux et ordinaux. Le lecteur qui ne serait pas familier avec ces notions est invité à se reporter à l'appendice où se trouvent réunis les résultats requis.

■ **Preuve du théorème 5.** Supposons d'abord $c > \aleph_1$. Nous allons montrer que pour toute famille $\{f_\alpha\}$ de fonctions analytiques de cardinal $\aleph_1$, il existe un nombre complexe z_0 tel que *toutes les* $\aleph_1$ valeurs $f_\alpha(z_0)$ soient distinctes. Par conséquent, si une famille de fonctions satisfait (P_0), alors elle doit être au plus dénombrable.

Pour s'en convaincre, il faut faire appel aux nombres ordinaux. On commence par bien ordonner la famille $\{f_\alpha\}$ relativement au nombre ordinal initial ω_1 de $\aleph_1$. Cela signifie, d'après la proposition 1 de l'appendice, que l'ensemble des indices parcourt tous les nombres ordinaux α qui sont plus petits que ω_1. Ensuite, on montre que l'ensemble des paires (α, β), $\alpha < \beta < \omega_1$, a $\aleph_1$ pour cardinal. Puisque tout $\beta < \omega_1$ est un ordinal dénombrable, l'ensemble des paires (α, β), $\alpha < \beta$, est au plus dénombrable pour tout β fixé. En prenant la réunion sur tous les β (au nombre de $\aleph_1$), on déduit de la proposition 6 de l'appendice que l'ensemble de toutes les paires (α, β), $\alpha < \beta$, a pour cardinal $\aleph_1$.

À présent, pour toute paire $\alpha < \beta$, considérons l'ensemble :

$$S(\alpha, \beta) \; = \; \{z \in \mathbb{C} : f_\alpha(z) = f_\beta(z)\}$$

Nous affirmons que chaque ensemble $S(\alpha, \beta)$ est au plus dénombrable. Pour le vérifier, considérons les disques C_k de rayon $k = 1, 2, 3, \ldots$ autour de l'origine du plan complexe. Si f_α et f_β coïncident en une infinité de points sur un certain C_k, alors f_α et f_β sont identiques selon un résultat bien connu sur les fonctions analytiques. Ainsi, f_α et f_β coïncident seulement sur un nombre fini de points sur chaque C_k, donc au plus sur un nombre dénombrable de points. Maintenant on pose $S = \bigcup_{\alpha<\beta} S(\alpha, \beta)$. Encore une fois, grâce à la proposition 6, on établit que S a pour cardinal $\aleph_1$ puisque chaque ensemble $S(\alpha, \beta)$ est au plus dénombrable. Le point clé apparaît ici : comme nous le savons, le cardinal de $\mathbb{C}$ est c et c est plus grand que $\aleph_1$ par hypothèse. Il existe donc un nombre complexe z_0 qui n'appartient pas à S ; pour un tel z_0 toutes les $\aleph_1$ valeurs $f_\alpha(z_0)$ sont distinctes.

Supposons maintenant que $c = \aleph_1$ et considérons l'ensemble $D \subseteq \mathbb{C}$ des nombres complexes $p + iq$ de parties réelle et imaginaire rationnelles. Puisque pour tout p l'ensemble $\{p + iq : q \in \mathbb{Q}\}$ est dénombrable, D est dénombrable. En outre, D est un ensemble *dense* dans $\mathbb{C}$ donc tout disque ouvert du plan complexe contient un point de D. Soit $\{z_\alpha : 0 \leqslant \alpha < \omega_1\}$ un bon ordre sur $\mathbb{C}$. Nous allons construire une famille $\{f_\beta : 0 \leqslant \beta < \omega_1\}$ de $\aleph_1$ fonctions analytiques telles que :

$$f_\beta(z_\alpha) \in D \text{ quand } \alpha < \beta \tag{1}$$

Une telle famille vérifie la condition (P_0). En effet, chaque point $z \in \mathbb{C}$ a un indice, disons $z = z_\alpha$. Pour tout $\beta > \alpha$, les valeurs $\{f_\beta(z_\alpha)\}$ appartiennent à l'ensemble *dénombrable* D. Puisque α est un nombre ordinal dénombrable, les fonctions f_β telles que $\beta \leqslant \alpha$ vont ajouter une quantité au plus dénombrable de valeurs supplémentaires $f_\beta(z_\alpha)$. Ainsi, l'ensemble de toutes les valeurs $\{f_\beta(z)\}$ est encore au plus dénombrable. Donc, si nous pouvons construire une famille $\{f_\beta\}$ satisfaisant (1), alors la seconde partie du théorème est démontrée.

La construction de $\{f_\beta\}$ utilise une récurrence transfinie. Pour f_0, nous pouvons prendre n'importe quelle fonction analytique, par exemple une fonction constante. Supposons que f_β ait déjà été construite pour tout $\beta < \gamma$. Puisque γ est un ordinal dénombrable, nous pouvons réordonner $\{f_\beta : 0 \leqslant \beta < \gamma\}$ en une suite $g_1, g_2, g_3, \ldots$ Le même ordonnancement de $\{z_\alpha : 0 \leqslant \alpha < \gamma\}$ donne une suite $w_1, w_2, w_3, \ldots$ Nous allons maintenant construire une fonction f_γ satisfaisant pour tout n les conditions :

$$f_\gamma(w_n) \in D \qquad \text{et} \qquad f_\gamma(w_n) \neq g_n(w_n) \tag{2}$$

La deuxième condition va permettre de s'assurer que toutes les fonctions f_γ $(0 \leqslant \gamma < \omega_1)$ sont distinctes ; la première condition est simplement la condition (1), ce qui implique (P_0) grâce à notre argument précédent. Notons que la condition $f_\gamma(w_n) \neq g_n(w_n)$ est une fois de plus un argument diagonal.

Pour construire f_γ, on écrit :

$$\begin{aligned} f_\gamma(z) \quad := \quad & \varepsilon_0 + \varepsilon_1(z - w_1) + \varepsilon_2(z - w_1)(z - w_2) \\ & + \varepsilon_3(z - w_1)(z - w_2)(z - w_3) + \ldots . \end{aligned}$$

Si γ est un ordinal fini, alors f_γ est un polynôme ; elle est donc analytique et nous pouvons certainement choisir des nombres ε_i tels que (2) soit satisfaite. Supposons maintenant que γ soit un ordinal dénombrable ; alors :

$$f_\gamma(z) \;=\; \varepsilon_0 + \sum_{k=1}^{\infty} \varepsilon_k (z - w_1) \cdots (z - w_k) \tag{3}$$

Remarquons que les valeurs de ε_m $(m \geqslant n)$ n'ont pas d'influence sur la valeur $f_\gamma(w_n)$. Nous pouvons donc choisir les ε_n pas à pas. Si la suite (ε_n) converge vers 0 suffisamment rapidement, alors (3) définit une fonction analytique. En définitive, puisque D est un ensemble dense, nous pouvons choisir cette suite (ε_n) de sorte que f_γ satisfasse (2), ce qui termine la démonstration. □

Appendice - À propos des nombres cardinaux et ordinaux

Interrogeons-nous, tout d'abord, sur la question de l'existence d'un nombre cardinal supérieur à un nombre cardinal donné. Pour commencer, nous montrons que pour tout nombre cardinal $\mathfrak{m}$, il existe toujours un nombre cardinal $\mathfrak{n}$ plus grand que $\mathfrak{m}$. Pour ce faire, nous utilisons encore une fois une version de la méthode diagonale de Cantor.

Soit M un ensemble. Nous affirmons que l'ensemble $\mathcal{P}(M)$ *de tous les sous-ensembles* de M a un cardinal plus grand que M. En associant à $m \in M$ le sous-ensemble $\{m\} \in \mathcal{P}(M)$, on voit que M peut être envoyé bijectivement sur un sous-ensemble de $\mathcal{P}(M)$, ce qui implique $|M| \leqslant |\mathcal{P}(M)|$ par définition. Il reste à montrer que $\mathcal{P}(M)$ *ne peut pas* être envoyé bijectivement sur un sous-ensemble de M. Supposons qu'au contraire

$$\varphi : N \longrightarrow \mathcal{P}(M)$$

soit une bijection de $N \subseteq M$ sur $\mathcal{P}(M)$. Considérons le sous-ensemble $U \subseteq N$ constitué des éléments de N qui *ne sont pas* contenus dans leur image par φ, c'est-à-dire que $U = \{m \in N : m \notin \varphi(m)\}$. Puisque φ est une bijection, il existe $u \in N$ tel que $\varphi(u) = U$. On a donc soit $u \in U$, soit $u \notin U$, mais les deux sont impossibles ! En effet, si $u \in U$, alors $u \notin \varphi(u) = U$ par définition de U, et si $u \notin U = \varphi(u)$, alors $u \in U$, ce qui est contradictoire.

« Une légende raconte que saint Augustin, marchant le long de la plage et contemplant l'infini, vit un enfant qui essayait de vider l'océan avec un petit coquillage… »

Le lecteur connaît probablement déjà cet argument. C'est le vieux paradoxe du barbier : « un barbier est un homme qui rase les personnes qui ne se rasent pas elles-mêmes. Est-ce que le barbier se rase lui-même ? »

Pour poursuivre cette théorie, nous introduisons un autre concept important dû à Cantor : les *ensembles ordonnés* et les *nombres ordinaux*. Un ensemble M est *ordonné* par $<$ si la relation $<$ est transitive et si, pour tout couple d'éléments distincts a et b de M, on a soit $a < b$ soit $b < a$. On peut, par exemple, ordonner $\mathbb{N}^*$ de manière naturelle $\mathbb{N}^* = \{1, 2, 3, 4, \ldots\}$. On peut

aussi évidemment l'ordonner à l'envers $\mathbb{N} = \{\ldots, 4, 3, 2, 1\}$, ou encore en énumérant d'abord les nombres impairs, ensuite les nombres pairs $\mathbb{N}^* = \{1, 3, 5, \ldots, 2, 4, 6, \ldots\}$.

Voici maintenant le concept fondamental. Un ensemble ordonné M est dit *bien ordonné* si chaque sous-ensemble non vide de M a un premier élément (ou plus petit élément). Ainsi, le premier et le troisième ordres associés à $\mathbb{N}^*$ ci-dessus sont de bons ordres mais ce n'est pas le cas du deuxième. Le *théorème fondamental du bon ordre*, déduit des axiomes, (incluant l'axiome du choix), affirme que *tout* ensemble M admet un bon ordre. À partir de maintenant, nous ne considérerons que des ensembles munis d'un bon ordre.

Les ensembles bien ordonnés :

$$\mathbb{N}^* = \{1, 2, 3, \ldots\}$$

et :

$$\mathbb{N}^* = \{1, 3, 5, \ldots, 2, 4, 6, \ldots\}$$

ne sont pas similaires : le premier ordre n'a qu'un élément sans prédécesseur immédiat, tandis que le second en a deux.

On dit que deux ensembles bien ordonnés M et N sont *similaires* (ou *isomorphes*, ou encore de *même type d'ordre*) s'il existe une bijection φ de M sur N qui respecte l'ordre, c'est-à-dire telle que $m <_M n$ implique $\varphi(m) <_N \varphi(n)$. Remarquons que tout ensemble ordonné similaire à un ensemble bien ordonné est lui-même bien ordonné.

Il est clair que la similarité est une relation d'équivalence. Nous pouvons donc parler d'un *nombre ordinal* α appartenant à une classe d'ensembles similaires. En ce qui concerne les ensembles finis, deux ensembles ordonnés quelconques ont des bons ordres similaires ; nous utilisons à nouveau le nombre ordinal n pour désigner la classe des n-ensembles. Remarquons que, par définition, deux ensembles similaires ont le même cardinal. Ainsi, on peut donner un sens au *cardinal* $|\alpha|$ d'un nombre ordinal α. Notons encore que tout sous-ensemble d'un ensemble bien ordonné est aussi bien ordonné par l'ordre induit.

Comme nous l'avons fait pour les nombres cardinaux, nous pouvons comparer deux nombres ordinaux. Soit M un ensemble bien ordonné, $m \in M$, alors $M_m = \{x \in M : x < m\}$ est appelé le *segment* (*initial*) de M déterminé par m ; N est un segment de M si $N = M_m$ pour un certain m. Ainsi, en particulier, M_m est l'ensemble vide lorsque m est le premier élément de M. Soient, maintenant, μ et ν les nombres ordinaux des ensembles bien ordonnés M et N. Nous disons que μ est *plus petit* que ν, $\mu < \nu$, si M est similaire à un segment de N. Cette relation est à nouveau transitive, c'est-à-dire que $\mu < \nu$, $\nu < \pi$ impliquent $\mu < \pi$, puisqu'une application de similarité envoie un segment sur un segment.

Le nombre ordinal de $\{1, 2, 3, \ldots\}$ est plus petit que le nombre ordinal de $\{1, 3, 5, \ldots, 2, 4, 6, \ldots\}$.

Pour les ensembles finis, $m < n$ a le sens habituel. Notons ω le nombre ordinal de $\mathbb{N}^* = \{1, 2, 3, 4, \ldots\}$ naturellement ordonné. En considérant le segment $\mathbb{N}_{n+1}$, on trouve que $n < \omega$ pour tout n fini. Ensuite, on voit que $\omega \leqslant \alpha$ pour tout nombre ordinal infini α. Bien sûr, si l'ensemble infini et bien ordonné M a pour nombre ordinal α, alors M contient un premier élément m_1, l'ensemble $M \backslash \{m_1\}$ contient un premier élément m_2, $M \backslash \{m_1, m_2\}$ contient un premier élément m_3. En continuant de cette façon, on construit une suite $m_1 < m_2 < m_3 < \ldots$ de M. Si $M = \{m_1, m_2, m_3, \ldots\}$, alors M est similaire à $\mathbb{N}^*$, et ainsi $\alpha = \omega$. D'autre part, si $M \backslash \{m_1, m_2, \ldots\}$ est non vide, alors il contient un premier élément m donc $\mathbb{N}^*$ est semblable au segment M_m, c'est-à-dire $\omega < \alpha$ par définition.

Nous énonçons maintenant (sans développer les démonstrations qui ne présentent pas de difficulté) trois résultats fondamentaux sur les nombres ordinaux. Le premier affirme que tout nombre ordinal μ a un ensemble bien ordonné « standard » W_μ qui le représente.

Proposition 1. *Soit μ un nombre ordinal. Notons W_μ l'ensemble des nombres ordinaux plus petits que μ. Alors*
(i) *les éléments de W_μ sont deux à deux comparables.*
(ii) *si l'on ordonne W_μ naturellement, alors W_μ est bien ordonné et a pour nombre ordinal μ.*

Proposition 2. *Deux nombres ordinaux μ et ν vérifient exactement l'une des relations suivantes : $\mu < \nu$, $\mu = \nu$ ou $\mu > \nu$.*

Proposition 3. *Tout ensemble de nombres ordinaux (naturellement ordonnés) est bien ordonné.*

Après cette présentation des nombres ordinaux, revenons aux nombres cardinaux. Soit $\mathfrak{m}$ un nombre cardinal. Notons $O_\mathfrak{m}$ l'ensemble de tous les nombres ordinaux μ tels que $|\mu| = \mathfrak{m}$. En utilisant la proposition 3, nous savons qu'il existe un *plus petit* nombre ordinal $\omega_\mathfrak{m}$ dans $O_\mathfrak{m}$, qu'on appelle le *nombre ordinal initial* de $O_\mathfrak{m}$. Par exemple, ω est le nombre ordinal initial de $\aleph_0$.
Après ces préliminaires, nous pouvons démontrer un résultat fondamental de ce chapitre.

Proposition 4. *Tout nombre cardinal $\mathfrak{m}$ a un successeur.*

■ **Preuve.** Nous savons déjà qu'il existe au moins un nombre cardinal $\mathfrak{n}$ supérieur à $\mathfrak{m}$. Considérons maintenant l'ensemble $\mathcal{K}$ de tous les nombres cardinaux plus grands que $\mathfrak{m}$ et inférieurs ou égaux à $\mathfrak{n}$. Associons à chaque $\mathfrak{p} \in \mathcal{K}$ son nombre ordinal initial $\omega_\mathfrak{p}$. Parmi ces nombres initiaux, il en existe un plus petit (proposition 3) ; le nombre cardinal correspondant est alors le plus petit élément de $\mathcal{K}$. C'est donc le nombre cardinal successeur de $\mathfrak{m}$ que l'on cherchait. □

Proposition 5. *Soit M un ensemble infini de cardinal $\mathfrak{m}$. Supposons M bien ordonné relativement au nombre ordinal initial $\omega_\mathfrak{m}$. Alors M n'a pas de dernier élément.*

■ **Preuve.** En effet, si M avait un dernier élément m, alors le segment M_m aurait un nombre ordinal $\mu < \omega_\mathfrak{m}$ tel que $|\mu| = \mathfrak{m}$, contredisant la définition de $\omega_\mathfrak{m}$. □

Nous avons besoin d'une amélioration considérable du résultat qui affirme que la réunion d'une famille dénombrable d'ensembles dénombrables est encore dénombrable. Dans le résultat suivant, nous considérons des familles *arbitraires* d'ensembles dénombrables.

Proposition 6. *Supposons que $\{A_\alpha\}$ soit une famille de cardinal $\mathfrak{m}$ d'ensembles au plus dénombrables A_α, où $\mathfrak{m}$ est un cardinal infini. Alors la réunion $\bigcup_\alpha A_\alpha$ a un cardinal au plus égale à $\mathfrak{m}$.*

■ **Preuve.** On peut supposer que les ensembles A_α sont deux à deux disjoints et dénombrables, puisque cette hypothèse ne peut qu'augmenter la taille de la réunion. Soit M tel que $|M| = \mathfrak{m}$ soit l'ensemble des indices, bien ordonné selon le nombre ordinal initial $\omega_{\mathfrak{m}}$. Remplaçons chaque $\alpha \in M$ par un ensemble au plus dénombrable $B_\alpha = \{b_{\alpha 1} = \alpha, b_{\alpha 2}, b_{\alpha 3}, \ldots\}$, ordonné par ω, et appelons le nouvel ensemble $\widetilde{M}$. Alors $\widetilde{M}$ est encore bien ordonné en posant $b_{\alpha i} < b_{\beta j}$ si $\alpha < \beta$ et $b_{\alpha i} < b_{\alpha j}$ pour $i < j$. Soit $\widetilde{\mu}$ le nombre ordinal de $\widetilde{M}$. Puisque M est un sous-ensemble de $\widetilde{M}$, on a $\mu \leqslant \widetilde{\mu}$ par un argument précédemment développé. Si $\mu = \widetilde{\mu}$, alors M est similaire à $\widetilde{M}$, et si $\mu < \widetilde{\mu}$, alors M est similaire à un segment de $\widetilde{M}$. Maintenant, puisque l'ordre $\omega_{\mathfrak{m}}$ sur M n'a pas de dernier élément (proposition 5), on voit que, dans les deux cas, M est similaire à la réunion d'ensembles dénombrables B_β donc de même cardinalité.

Le reste est facile. Soit $\varphi : \bigcup B_\beta \longrightarrow M$ une bijection. Supposons que $\varphi(B_\beta) = \{\alpha_1, \alpha_2, \alpha_3, \ldots\}$. Remplaçons chaque α_i par A_{α_i} et considérons la réunion $\bigcup A_{\alpha_i}$. Puisque $\bigcup A_{\alpha_i}$ est la réunion *dénombrable* d'une famille d'ensembles dénombrables (il est donc dénombrable), on voit que B_β a le même cardinal que $\bigcup A_{\alpha_i}$. En d'autres termes, il existe une bijection de B_β sur $\bigcup A_{\alpha_i}$ pour tout β donc une bijection ψ de $\bigcup B_\beta$ sur $\bigcup A_\alpha$. Par conséquent, $\psi \circ \varphi^{-1}$ fournit la bijection désirée de M sur $\bigcup A_\alpha$. Ainsi, $|\bigcup A_\alpha| = \mathfrak{m}$. □

Bibliographie

[1] L. E. J. BROUWER : *Beweis der Invarianz der Dimensionszahl,* Math. Annalen **70** (1911), 161-165.

[2] N. CALKIN & H. WILF : *Recounting the rationals,* Amer. Math. Monthly **107** (2000), 360-363.

[3] G. CANTOR : *Ein Beitrag zur Mannigfaltigkeitslehre,* Journal für die reine und angewandte Mathematik **84** (1878), 242-258.

[4] P. COHEN : *Set Theory and the Continuum Hypothesis,* W. A. Benjamin, New York 1966.

[5] P. ERDŐS : *An interpolation problem associated with the continuum hypothesis,* Michigan Math. J. **11** (1964), 9-10.

[6] E. KAMKE : *Theory of Sets,* Dover Books 1950.

[7] M. A. STERN : *Ueber eine zahlentheoretische Funktion,* Journal für die reine und angewandte Mathematik **55** (1858), 193-220.

« Infiniment plus de cardinaux ».

À la gloire des inégalités

Chapitre 18

L'analyse fourmille d'inégalités, comme en témoigne, par exemple, le célèbre ouvrage *Inequalities* de Hardy, Littlewood et Pólya. Voici deux inégalités parmi les plus fondamentales, deux applications pour chacune d'elles, et les démonstrations retenues par George Pólya, lui-même champion du Grand Livre.

Notre première inégalité est suivant les cas attribuée à Cauchy, à Schwarz et/ou à Buniakowski :

Théorème I. (Inégalité de Cauchy-Schwarz)
Soit $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ un produit scalaire sur l'espace vectoriel réel V muni de la norme $|\boldsymbol{a}|^2 := \langle \boldsymbol{a}, \boldsymbol{a} \rangle$. Alors :

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle^2 \ \leqslant \ |\boldsymbol{a}|^2 |\boldsymbol{b}|^2$$

pour tous vecteurs $\boldsymbol{a}, \boldsymbol{b} \in V$, l'égalité ayant lieu si et seulement si $\boldsymbol{a}$ et $\boldsymbol{b}$ sont linéairement dépendants.

■ **Preuve.** La démonstration suivante est probablement la plus courte. Considérons la fonction quadratique :

$$|x\boldsymbol{a} + \boldsymbol{b}|^2 \ = \ x^2 |\boldsymbol{a}|^2 + 2x \langle \boldsymbol{a}, \boldsymbol{b} \rangle + |\boldsymbol{b}|^2$$

de la variable x. Nous pouvons supposer $\boldsymbol{a} \neq \boldsymbol{0}$. Si $\boldsymbol{b} = \lambda \boldsymbol{a}$, il est clair que : $\langle \boldsymbol{a}, \boldsymbol{b} \rangle^2 = |\boldsymbol{a}|^2 |\boldsymbol{b}|^2$. D'autre part, si $\boldsymbol{a}$ et $\boldsymbol{b}$ sont linéairement indépendants, alors $|x\boldsymbol{a} + \boldsymbol{b}|^2 > 0$, pour tout x, et le discriminant $\langle \boldsymbol{a}, \boldsymbol{b} \rangle^2 - |\boldsymbol{a}|^2 |\boldsymbol{b}|^2$ est inférieur à 0. □

Notre second exemple présente une relation entre *les moyennes harmonique, géométrique et arithmétique* :

Théorème II. (Moyennes harmonique, géométrique et arithmétique)
Soient $a_1, \ldots, a_n$ des nombres réels positifs, alors :

$$\frac{n}{\frac{1}{a_1} + \ldots + \frac{1}{a_n}} \ \leqslant \ \sqrt[n]{a_1 a_2 \ldots a_n} \ \leqslant \ \frac{a_1 + \ldots + a_n}{n}$$

l'égalité ayant lieu, dans chaque cas, si et seulement si tous les a_i sont égaux.

■ **Preuve.** La belle démonstration qui suit repose sur un raisonnement par récurrence très original, attribué à Cauchy (voir [7]). Soit $P(n)$ l'énoncé concernant la deuxième inégalité, écrite sous la forme :

$$a_1 a_2 \ldots a_n \ \leqslant \ \Big(\frac{a_1 + \ldots + a_n}{n} \Big)^n$$

Pour $n = 2$, il faut établir que $a_1 a_2 \leqslant (\frac{a_1+a_2}{2})^2$ c'est-à-dire $(a_1 - a_2)^2 \geqslant 0$, ce qui est vrai. Nous allons maintenant procéder en deux étapes :
(A) $P(n) \Longrightarrow P(n-1)$
(B) $P(n)$ et $P(2) \Longrightarrow P(2n)$
qui manifestement impliquent le résultat pour tout n.

Pour montrer **(A)**, posons $A := \sum_{k=1}^{n-1} \frac{a_k}{n-1}$. Alors :

$$\Big(\prod_{k=1}^{n-1} a_k\Big) A \overset{P(n)}{\leqslant} \left(\frac{\sum_{k=1}^{n-1} a_k \; + \; A}{n}\right)^n = \Big(\frac{(n-1)A + A}{n}\Big)^n = A^n$$

et donc : $\displaystyle\prod_{k=1}^{n-1} a_k \leqslant A^{n-1} = \left(\frac{\sum_{k=1}^{n-1} a_k}{n-1}\right)^{n-1}$

Pour **(B)**, on voit que :

$$\prod_{k=1}^{2n} a_k = \Big(\prod_{k=1}^{n} a_k\Big)\Big(\prod_{k=n+1}^{2n} a_k\Big) \overset{P(n)}{\leqslant} \Big(\sum_{k=1}^{n} \frac{a_k}{n}\Big)^n \Big(\sum_{k=n+1}^{2n} \frac{a_k}{n}\Big)^n$$

$$\overset{P(2)}{\leqslant} \left(\frac{\sum_{k=1}^{2n} \frac{a_k}{n}}{2}\right)^{2n} = \left(\frac{\sum_{k=1}^{2n} a_k}{2n}\right)^{2n}$$

La condition d'égalité s'en déduit facilement. L'inégalité de gauche, entre la moyenne harmonique et géométrique, se déduit de la précédente en considérant $\frac{1}{a_1}, \ldots, \frac{1}{a_n}$. □

■ **Une autre Preuve.** Parmi bien d'autres preuves de l'inégalité entre les moyennes arithmétique et géométrique (la monographie [2] en dénombre plus de 50), on peut en retenir une récente, particulièrement étonnante et due à Alzer. En fait, cette preuve conduit à l'inégalité plus forte :

$$a_1^{p_1} a_2^{p_2} \ldots a_n^{p_n} \;\leqslant\; p_1 a_1 + p_2 a_2 + \ldots + p_n a_n$$

pour tous nombres positifs $a_1, \ldots, a_n, p_1, \ldots, p_n$ tels que : $\sum_{i=1}^n p_i = 1$. Notons G l'expression de gauche et A celle de droite. On peut supposer que : $a_1 \leqslant \ldots \leqslant a_n$. Il est clair que : $a_1 \leqslant G \leqslant a_n$. Il doit donc exister un k tel que : $a_k \leqslant G \leqslant a_{k+1}$. Par suite :

$$\sum_{i=1}^{k} p_i \int_{a_i}^{G} \Big(\frac{1}{t} - \frac{1}{G}\Big)\, dt \;+ \sum_{i=k+1}^{n} p_i \int_{G}^{a_i} \Big(\frac{1}{G} - \frac{1}{t}\Big)\, dt \;\geqslant\; 0 \qquad (1)$$

puisque tous les intégrandes sont positifs. En réécrivant (1), on obtient :

$$\sum_{i=1}^{n} p_i \int_{G}^{a_i} \frac{1}{G}\, dt \;\geqslant\; \sum_{i=1}^{n} p_i \int_{G}^{a_i} \frac{1}{t}\, dt$$

où l'expression de gauche est égale à :

$$\sum_{i=1}^{n} p_i \frac{a_i - G}{G} = \frac{A}{G} - 1$$

tandis que celle de droite est égale à :

$$\sum_{i=1}^{n} p_i(\ln a_i - \ln G) = \ln \prod_{i=1}^{n} a_i^{p_i} - \ln G = 0$$

On en conclut que : $\frac{A}{G} - 1 \geqslant 0$, donc que $A \geqslant G$. Dans le cas d'égalité, toutes les intégrales de (1) doivent être égales à 0, ce qui implique : $a_1 = \ldots = a_n = G$. □

Notre première application est un beau résultat de Laguerre (voir [7]) concernant la localisation des racines des polynômes.

Théorème 1. *Supposons que toutes les racines du polynôme* $x^n + a_{n-1}x^{n-1} + \ldots + a_0$ *sont réelles. Alors elles sont contenues dans l'intervalle d'extrémités :*

$$-\frac{a_{n-1}}{n} \pm \frac{n-1}{n}\sqrt{a_{n-1}^2 - \frac{2n}{n-1}a_{n-2}}$$

■ **Preuve.** Soit y l'une des racines et soient $y_1, \ldots, y_{n-1}$ les autres. Alors le polynôme s'écrit $(x-y)(x-y_1)\cdots(x-y_{n-1})$. Ainsi :

$$\begin{aligned} -a_{n-1} &= y + y_1 + \ldots + y_{n-1}, \\ a_{n-2} &= y(y_1 + \ldots + y_{n-1}) + \sum_{i<j} y_i y_j \end{aligned}$$

et donc :

$$a_{n-1}^2 - 2a_{n-2} - y^2 = \sum_{i=1}^{n-1} y_i^2$$

L'inégalité de Cauchy-Schwarz appliquée à $(y_1, \ldots, y_{n-1})$ et $(1, \ldots, 1)$ implique :

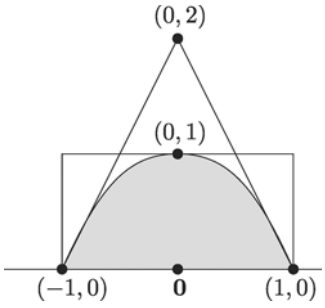
$$\begin{aligned} (a_{n-1} + y)^2 &= (y_1 + y_2 + \ldots + y_{n-1})^2 \\ &\leqslant (n-1)\sum_{i=1}^{n-1} y_i^2 = (n-1)(a_{n-1}^2 - 2a_{n-2} - y^2) \end{aligned}$$

soit encore :

$$y^2 + \frac{2a_{n-1}}{n}y + \frac{2(n-1)}{n}a_{n-2} - \frac{n-2}{n}a_{n-1}^2 \leqslant 0$$

Ainsi, y (et donc tous les y_i) se trouve entre les deux racines du trinôme du second degré. Ces racines sont les bornes recherchées. □

Pour notre deuxième application, nous commençons par une propriété élémentaire bien connue de la parabole. Considérons l'arc de parabole décrit

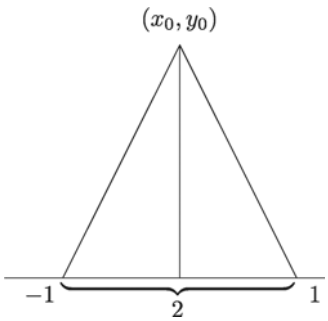


par $f(x) = 1 - x^2$ entre $x = -1$ et $x = 1$. Nous associons à $f(x)$ le *triangle tangent* et le *rectangle tangent* comme indiqué sur la figure ci-contre. Nous constatons que l'aire ombrée $A = \int_{-1}^{1}(1 - x^2)dx$ est égale à $\frac{4}{3}$; les aires T et R du triangle comme du rectangle sont égales à 2. Ainsi $\frac{T}{A} = \frac{3}{2}$ et $\frac{R}{A} = \frac{3}{2}$.

Dans un bel article, Paul Erdős et Tibor Gallai se sont demandé ce qui arrive lorsque $f(x)$ est un polynôme réel arbitraire de degré n tel que $f(x) > 0$ si $-1 < x < 1$, et $f(-1) = f(1) = 0$. L'aire A est donc égale à $\int_{-1}^{1} f(x)dx$. Supposons que $f(x)$ atteigne son maximum sur $]-1, 1[$ en b. Alors $R = 2f(b)$. En calculant les tangentes en -1 et en 1, il est facile de se convaincre (voir encadré) que :

$$T = \frac{2f'(1)f'(-1)}{f'(1) - f'(-1)} \tag{2}$$

et $T = 0$ si $f'(1) = f'(-1) = 0$.



Le triangle tangent

L'aire T du triangle tangent est précisément y_0, où (x_0, y_0) est le point d'intersection des deux tangentes. Les équations de ces tangentes sont $y = f'(-1)(x + 1)$ et $y = f'(1)(x - 1)$, par suite :

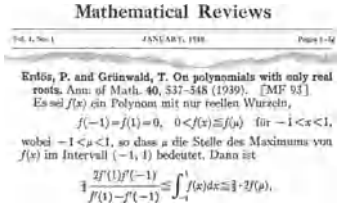
$$x_0 = \frac{f'(1) + f'(-1)}{f'(1) - f'(-1)}$$

Ainsi :

$$y_0 = f'(1)\left(\frac{f'(1) + f'(-1)}{f'(1) - f'(-1)} - 1\right) = 2\frac{f'(1)f'(-1)}{f'(1) - f'(-1)}$$

En général, il n'y a pas de bornes non triviales de $\frac{T}{A}$ et $\frac{R}{A}$. Pour s'en convaincre, prenons $f(x) = 1 - x^{2n}$. Alors $T = 2n$, $A = \frac{4n}{2n+1}$ et donc $\frac{T}{A} > n$. De même, $R = 2$ donc $\frac{R}{A} = \frac{2n+1}{2n}$, qui tend vers 1 lorsque n tend vers l'infini.

Cependant, comme Erdős et Gallai l'ont montré, les polynômes qui n'ont que des racines réelles, admettent effectivement de telles bornes.


Mathematical Reviews

JANUARY, 1940

Erdös, P. and Grünwald, T. On polynomials with only real roots. Ann. of Math. **40**, 537–548 (1939). [MF 93].

Es sei $f(x)$ ein Polynom mit nur reellen Wurzeln,

$$f(-1)=f(1)=0, \quad 0<f(x)\leqq f(\mu) \quad \text{für } -1<x<1,$$

wobei $-1<\mu<1$, so dass μ die Stelle des Maximums von $f(x)$ im Intervall $(-1, 1)$ bedeutet. Dann ist

$$\tfrac{2}{3}\frac{f'(1)f'(-1)}{f'(1)-f'(-1)} \leqq \int_{-1}^{1} f(x)dx \leqq \tfrac{2}{3}\cdot 2f(\mu).$$

Théorème 2. *Soit $f(x)$ un polynôme réel de degré $n \geqslant 2$ qui n'a que des racines réelles, tel que : $f(x) > 0$ si $-1 < x < 1$ et $f(-1) = f(1) = 0$. Alors :*

$$\frac{2}{3}T \leqslant A \leqslant \frac{2}{3}R,$$

l'égalité n'ayant lieu, dans chaque cas, que si $n = 2$.

Erdős et Gallai ont démontré ce résultat à l'aide d'une récurrence compliquée. Dans la critique de leur article, qui fut publiée en première page du

premier volume du Mathematical Reviews en 1940, George Pólya expliqua comment la première inégalité peut aussi être démontrée à l'aide de l'inégalité entre les moyennes arithmétique et géométrique : cela constitue à la fois un bel exemple de critique consciencieuse et une Preuve du Grand Livre.

■ **Preuve de $\frac{2}{3}T \leqslant A$.** Puisque $f(x)$ n'a que des racines réelles et qu'aucune d'elles n'est dans l'intervalle ouvert $]-1,1[$, elle peut s'écrire (à un facteur constant positif près qui disparaît à la fin) sous la forme :

$$f(x) \;=\; (1-x^2)\prod_i(\alpha_i - x)\prod_j(\beta_j + x) \tag{3}$$

avec $\alpha_i \geqslant 1$, $\beta_j \geqslant 1$. Ainsi :

$$A \;=\; \int_{-1}^{1}(1-x^2)\prod_i(\alpha_i - x)\prod_j(\beta_j + x)dx$$

En faisant la substitution $x \longmapsto -x$, on trouve également que :

$$A \;=\; \int_{-1}^{1}(1-x^2)\prod_i(\alpha_i + x)\prod_j(\beta_j - x)dx$$

et donc, d'après l'inégalité entre les moyennes géométrique et arithmétique (remarquons que tous les facteurs sont positifs) :

$$\begin{aligned}
A \;&=\; \int_{-1}^{1}\frac{1}{2}\Big[\;(1-x^2)\prod_i(\alpha_i - x)\prod_j(\beta_j + x) \;+ \\
&\qquad\qquad (1-x^2)\prod_i(\alpha_i + x)\prod_j(\beta_j - x)\;\Big]dx \\
&\geqslant\; \int_{-1}^{1}(1-x^2)\Big(\prod_i(\alpha_i^2 - x^2)\prod_j(\beta_j^2 - x^2)\Big)^{1/2}dx \\
&\geqslant\; \int_{-1}^{1}(1-x^2)\Big(\prod_i(\alpha_i^2 - 1)\prod_j(\beta_j^2 - 1)\Big)^{1/2}dx \\
&=\; \frac{4}{3}\Big(\prod_i(\alpha_i^2 - 1)\prod_j(\beta_j^2 - 1)\Big)^{1/2}
\end{aligned}$$

Calculons $f'(1)$ et $f'(-1)$ (on peut supposer $f'(-1), f'(1) \neq 0$, sinon $T = 0$ et l'inégalité $\frac{2}{3}T \leqslant A$ devient triviale). De (3) on déduit :

$$f'(1) \;=\; -2\prod_i(\alpha_i - 1)\prod_j(\beta_j + 1)$$

et de la même façon :

$$f'(-1) \;=\; 2\prod_i(\alpha_i + 1)\prod_j(\beta_j - 1)\,.$$

Ainsi :

$$A \geqslant \frac{2}{3}(-f'(1)f'(-1))^{1/2}$$

En appliquant maintenant l'inégalité des moyennes harmonique et géométrique à $-f'(1)$ et $f'(1)$, puis en utilisant (2) on arrive à la conclusion :

$$A \geqslant \frac{2}{3}\frac{2}{\frac{1}{-f'(1)}+\frac{1}{f'(-1)}} = \frac{4}{3}\frac{f'(1)f'(-1)}{f'(1)-f'(-1)} = \frac{2}{3}T$$

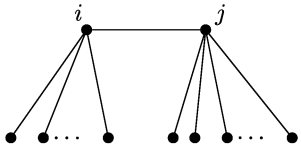
ce qu'il fallait démontrer. En analysant le cas d'égalité dans toutes les inégalités, le lecteur peut facilement déduire la dernière partie du théorème. □

Le lecteur est invité à chercher une preuve aussi inspirée de la seconde inégalité du Théorème 2.

Après tout, l'analyse peut se résumer à des inégalités ! Voici maintenant un exemple issu de la théorie des graphes où l'on utilise des inégalités de façon inattendue. Au chapitre 36 nous parlerons du théorème de Turán. Dans le cas le plus simple, il prend la forme suivante :

Théorème 3. *Soit G un graphe à n sommets sans triangles. Alors G a au plus $\frac{n^2}{4}$ arêtes, l'égalité ayant lieu seulement si n est pair et si G est le graphe complet biparti $K_{n/2,n/2}$.*

■ **Première preuve.** Cette preuve, qui repose sur l'inégalité de Cauchy-Schwarz, est due à Mantel. Soit $V = \{1, \dots, n\}$ l'ensemble des sommets et E l'ensemble des arêtes de G. Nous notons d_i le degré de i, donc : $\sum_{i \in V} d_i = 2|E|$ (voir page 188 du chapitre 25 sur le double dénombrement). Soit ij une arête. Comme G n'a pas de triangles, on a $d_i + d_j \leqslant n$ puisqu'aucun sommet n'est voisin à la fois de i et de j.



Par suite :

$$\sum_{ij \in E} (d_i + d_j) \leqslant n|E|$$

Comme d_i apparaît exactement d_i fois dans la somme, on obtient :

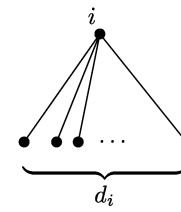
$$n|E| \geqslant \sum_{ij \in E} (d_i + d_j) = \sum_{i \in V} d_i^2$$

et donc, en appliquant l'inégalité de Cauchy-Schwarz aux vecteurs $(d_1, \dots, d_n)$ et $(1, \dots, 1)$:

$$n|E| \geqslant \sum_{i \in V} d_i^2 \geqslant \frac{(\sum d_i)^2}{n} = \frac{4|E|^2}{n}$$

on trouve immédiatement le résultat. En cas d'égalité, nous trouvons $d_i = d_j$ pour tous les i, j, et, en outre, $d_i = \frac{n}{2}$ (puisque $d_i + d_j = n$). Comme G est sans triangle, on en conclut immédiatement que $G = K_{n/2,n/2}$. □

■ **Deuxième preuve.** La preuve suivante du théorème 3, qui utilise l'inégalité des moyennes arithmétique et géométrique, est une preuve folklorique du Grand Livre. Soit α la taille d'un ensemble indépendant maximal A ; posons $\beta = n - \alpha$. Puisque G n'a pas de triangle, les voisins d'un sommet i forment un ensemble indépendant, et donc $d_i \leqslant \alpha$ pour tout i.



L'ensemble $B = V \backslash A$ de taille β rencontre chaque arête de G. En comptant les arêtes de G selon leurs extrémités dans B, on trouve : $|E| \leqslant \sum_{i \in B} d_i$. L'inégalité des moyennes arithmétique et géométrique implique alors :

$$|E| \leqslant \sum_{i \in B} d_i \leqslant \alpha\beta \leqslant \Big(\frac{\alpha+\beta}{2}\Big)^2 = \frac{n^2}{4}$$

Le cas d'égalité se traite, encore une fois, sans difficulté. □

Bibliographie

[1] H. ALZER : *A proof of the arithmetic mean-geometric mean inequality,* Amer. Math. Monthly **103** (1996), 585.

[2] P. S. BULLEN, D. S. MITRINOVICS & P. M. VASIĆ : *Means and their Inequalities,* Reidel, Dordrecht 1988.

[3] P. ERDŐS & T. GRÜNWALD : *On polynomials with only real roots,* Annals Math. **40** (1939), 537-548.

[4] G. H. HARDY, J. E. LITTLEWOOD & G. PÓLYA : *Inequalities,* Cambridge University Press, Cambridge 1952.

[5] W. MANTEL : *Problem 28,* Wiskundige Opgaven **10** (1906), 60-61.

[6] G. PÓLYA : *Review of* [3], Mathematical Reviews **1** (1940), 1.

[7] G. PÓLYA & G. SZEGŐ : *Problems and Theorems in Analysis, Vol. I,* Springer-Verlag, Berlin Heidelberg New York 1972/78 ; Reprint 1998.

Le théorème fondamental de l'algèbre

Chapitre 19

> *Tout polynôme à coefficients complexes non constant admet au moins une racine dans le corps des complexes.*

Jean le Rond dit d'Alembert

Gauss, qui proposa quatre démonstrations différentes de ce résultat, donna à ce théorème le nom de « théorème fondamental des équations algébriques ». Ce résultat constitue sans aucun doute une étape importante de l'histoire des mathématiques. Comme le suggère Reinhold Remmert dans son pertinent article : « Plus que toute autre chose, c'est la possibilité d'établir ce résultat dans le corps des complexes qui a ouvert la voie à une reconnaissance des nombres complexes. »

Les plus grands noms ont contribué à la question : Gauss, Cauchy, Liouville, Laplace... Un article de Netto et Le Vavasseur donne une liste d'une centaine de preuves différentes. La démonstration que nous présentons ici est l'une des plus élégantes et c'est certainement la plus courte. Elle s'appuie sur une idée originale de d'Alembert et Argand ; elle ne fait appel qu'à des propriétés élémentaires des polynômes et des nombres complexes. Nous devons à France Dacar une version améliorée de cette démonstration. La même idée fondamentale apparaît aussi dans les articles de Redheffer [4] et de Wolfenstein [6] ; il est probable qu'elle figure dans d'autres démonstrations encore.

Nous utiliserons les trois résultats suivants, qui relèvent d'un cours d'analyse élémentaire.

(A) Les fonctions polynomiales sont des fonctions continues.
(B) Tout nombre complexe admet une racine m-ième pour tout $m \geqslant 1$.
(C) Le principe du minimum de Cauchy : toute fonction numérique continue sur un compact atteint son minimum sur ce compact.

Considérons le polynôme p à coefficients complexes de degré $n \geqslant 1$ défini par $p(z) = \sum_{k=0}^{n} c_k z^k$. La première étape — décisive — consiste à prouver le résultat suivant, connu sous le nom de lemme de d'Alembert ou d'inégalité d'Argand.

Lemme. *Si $p(a) \neq 0$, alors tout disque ouvert D centré en a contient un point b tel que $|p(b)| < |p(a)|$.*

■ **Preuve.** Soit R le rayon du disque D. Alors les points de D sont de la forme $a + w$ avec $|w| < R$. À l'aide d'une manipulation algébrique simple,

nous allons commencer par montrer que :

$$p(a+w) \;=\; p(a) + cw^m(1+r(w)), \tag{1}$$

où c est un nombre complexe non nul, où $1 \leqslant m \leqslant n$ et où $r(w)$ est une expression polynomiale de degré $n-m$ telle que $r(0)=0$.
En effet, on constate que :

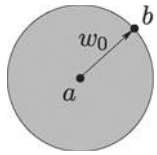
$$\begin{aligned} p(a+w) &= \sum_{k=0}^{n} c_k (a+w)^k \\ &= \sum_{k=0}^{n} c_k \sum_{i=0}^{k} \binom{k}{i} a^{k-i} w^i \;=\; \sum_{i=0}^{n} \Big(\sum_{k=i}^{n} \binom{k}{i} c_k a^{k-i} \Big) w^i \\ &= p(a) + \sum_{i=1}^{n} \Big(\sum_{k=i}^{n} \binom{k}{i} c_k a^{k-i} \Big) w^i \;=\; p(a) + \sum_{i=1}^{n} d_i w^i \,. \end{aligned}$$

Désignons par $m \geqslant 1$ le plus petit indice i pour lequel d_i est non nul, posons $c = d_m$ et factorisons cw^m pour obtenir :

$$p(a+w) \;=\; p(a) + cw^m(1+r(w)).$$

À présent, nous voulons majorer $|cw^m|$ et $|r(w)|$. Si $|w|$ est inférieur à $\rho_1 := \sqrt[m]{|p(a)/c|}$, alors $|cw^m| < |p(a)|$. En outre, comme r est une fonction continue et que $r(0)=0$, il existe un réel strictement positif ρ_2 tel que $|r(w)|<1$ pour $|w|<\rho_2$. Ainsi, pour $|w|$ inférieur à $\rho := \min(\rho_1,\rho_2)$ on obtient :

$$|cw^m| \;<\; |p(a)| \qquad \text{et} \qquad |r(w)| \;<\; 1. \tag{2}$$



On en vient à utiliser le deuxième ingrédient, les racines m-ièmes de l'unité. Soit ζ une racine m-ième de $-\frac{p(a)/c}{|p(a)/c|}$, qui est un nombre complexe de module 1. Soit ε un nombre réel vérifiant $0 < \varepsilon < \min(\rho, R)$. Posant $w_0 = \varepsilon\zeta$, on va montrer que le point $b = a + w_0$ est un point de D tel que $|p(b)| < |p(a)|$. Tout d'abord, b appartient à D puisque $|w_0| = \varepsilon < R$ et que d'après (1) :

$$|p(b)| = |p(a+w_0)| = |p(a) + cw_0^m(1+r(w_0))|. \tag{3}$$

On définit à présent le facteur δ par :

$$c\,w_0^m \;=\; c\,\varepsilon^m\zeta^m \;=\; -\frac{\varepsilon^m}{|p(a)/c|}\,p(a) \;=\; -\delta\,p(a)\,.$$

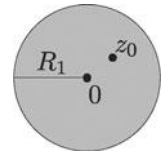
D'après (2), δ vérifie :

$$0 \;<\; \delta \;=\; \varepsilon^m \frac{|c|}{|p(a)|} \;<\; 1.$$

En appliquant l'inégalité triangulaire au membre de droite de l'égalité (3), on obtient :

$$\begin{aligned} |p(a) + cw_0^m\big(1+r(w_0)\big)| &= |p(a) - \delta p(a)(1+r(w_0))| \\ &= |(1-\delta)p(a) - \delta p(a) r(w_0)| \\ &\leqslant (1-\delta)|p(a)| \;+\; \delta|p(a)||r(w_0)| \\ &< (1-\delta)|p(a)| \;+\; \delta|p(a)| \;=\; |p(a)|, \end{aligned}$$

La suite est facile. Il est manifeste que $p(z)z^{-n}$ tend vers le coefficient dominant c_n de $p(z)$ lorsque $|z|$ tend vers l'infini. Donc $|p(z)|$ tend vers l'infini lorsque $|z| \to \infty$. Il existe donc $R_1 > 0$ tel que $|p(z)| > |p(0)|$ pour tous les z situés sur le cercle $\{z/|z| = R_1\}$. On utilise à présent le résultat (C). La fonction numérique $z \mapsto |p(z)|$ est continue sur le compact $D_1 = \{z/|z| \leqslant R_1\}$ si bien qu'elle atteint son minimum sur D_1 en un certain point z_0. Comme $|p(z)| > |p(0)|$ pour les z situés sur le bord de D_1, z_0 se trouve nécessairement à l'intérieur de D_1. D'après le lemme de d'Alembert, $|p(z_0)|$ est nécessairement égal à 0, ce qui termine la preuve.



Bibliographie

[1] D'ALEMBERT : *Recherches sur le calcul intégral,* Histoire de l'Académie royale des sciences et belles lettres (1746), 182-224.

[2] R. ARGAND : *Réflexions sur la nouvelle théorie d'analyse,* Annales de Mathématiques **5** (1814), 197-209.

[3] E. NETTO & R. LE VAVASSEUR : *Les fonctions rationelles,* Enc. Sciences Math. Pures Appl. I **2** (1907), 1-232.

[4] R. M. REDHEFFER : *What ! Another note just on the fundamental theorem of algebra ?* Amer. Math. Monthly **71** (1964), 180-185.

[5] R. REMMERT : *The fundamental theorem of algebra,* chap. 4 in : "Numbers" (H. D. Ebbinghaus *et al.*, eds.), Graduate Texts in Mathematics **123**, Springer, New York 1991.

[6] S. WOLFENSTEIN : *Proof of the fundamental theorem of algebra,* Amer. Math. Monthly **74** (1967), 853-854.

– Que se passe-t-il cette fois ?

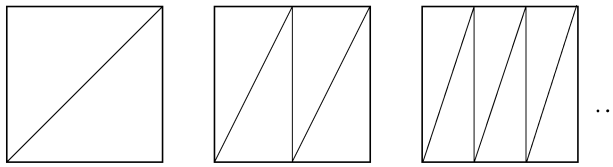
– Eh bien, je transporte 100 démonstrations du théorème fondamental de l'algèbre !

– Preuves divines : une preuve pour le théorème fondamental, une pour la loi de réciprocité quadratique !

Un carré et un nombre impair de triangles

Chapitre 20

Supposons que l'on souhaite découper un carré en n triangles d'aires égales. Lorsque n est pair, un tel découpage est facile à réaliser. Par exemple, on peut diviser les côtés horizontaux du carré en $\frac{n}{2}$ segments de longueurs égales et tracer une diagonale dans chacun des $\frac{n}{2}$ rectangles ainsi délimités :



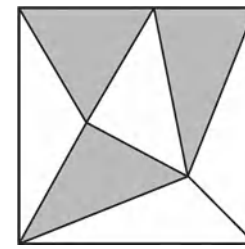
Considérons maintenant le cas où n est impair. Même pour $n = 3$, cela pose problème. Après quelques tentatives, on peut penser qu'il n'y a pas de solution. Nous posons donc le problème général suivant :

> *Est-il possible de découper un carré en un nombre n impair de triangles d'aires égales ?*

Ainsi posé, ce problème semble constituer une question classique de géométrie euclidienne et l'on pourrait penser que la réponse est connue depuis longtemps, peut-être même depuis l'Antiquité. Alors qu'ils vulgarisaient le sujet dans les années 1960, Fred Richman et John Thomas ont eu la surprise de constater que personne ne connaissait la réponse et qu'il n'y avait aucune référence bibliographique sur le sujet.

La réponse est en fait négative, non seulement pour $n = 3$ mais aussi pour tout entier n impair.

Comment peut-on démontrer un tel résultat ? Par homothétie, on se ramène à l'étude du carré unité de sommets $(0,0), (1,0), (0,1), (1,1)$. Toute démonstration doit utiliser le fait que l'aire des triangles d'un tel découpage serait $\frac{1}{n}$ où n est impair. La preuve qui suit, due à Paul Monsky ainsi qu'à des travaux préliminaires de John Thomas, constitue un trait de génie très surprenant. Elle utilise un outil algébrique, les valeurs absolues, pour composer une étonnante coloration du plan et conclure à l'aide d'un raisonnement combinatoire étonnamment simple. Qui plus est, aucune autre démonstration de ce résultat n'a été proposée à ce jour.



Il existe des découpages d'un carré en un nombre de triangles d'aires *à peu près* égales.

Avant d'établir le théorème, préparons le terrain en présentant rapidement les valeurs absolues.

Tout le monde connaît la fonction valeur absolue $x \mapsto |x|$ définie sur l'ensemble des nombres rationnels $\mathbb{Q}$ (ou sur l'ensemble des nombres réels $\mathbb{R}$). Elle envoie $\mathbb{Q}$ sur l'ensemble des réels positifs de sorte que pour tout x et tout y :

(i) $|x| = 0$ si et seulement si $x = 0$,
(ii) $|xy| = |x||y|$ et
(iii) $|x + y| \leqslant |x| + |y|$ (inégalité triangulaire).

L'inégalité triangulaire fait de $\mathbb{R}$ un espace métrique et permet d'introduire la notion de convergence. Dans les années 1900, une grande découverte a été de constater qu'au-delà de la valeur absolue usuelle, il y a d'autres « valeurs absolues » sur $\mathbb{Q}$ satisfaisant aux conditions (i) à (iii).

Soit p un nombre premier. Tout nombre rationnel $r \neq 0$ s'écrit de manière unique sous la forme :

$$r = p^k \frac{a}{b}, \quad k \in \mathbb{Z}, \tag{1}$$

où a et $b > 0$ sont premiers avec p. On définit la *valeur absolue p-adique* $|r|_p$ de r par :

$$|r|_p := p^{-k}, \quad \text{si } r \neq 0 \quad \text{et} \quad |0|_p = 0. \tag{2}$$

Exemple : $|\frac{3}{4}|_2 = 4$, $|\frac{6}{7}|_2 = |2|_2 = \frac{1}{2}$ et $|\frac{3}{4} + \frac{6}{7}|_2 = |\frac{45}{28}|_2 = |\frac{1}{4} \cdot \frac{45}{7}|_2 = 4 = \max\{|\frac{3}{4}|_2, |\frac{6}{7}|_2\}$.

Les conditions (i) et (ii) sont manifestement satisfaites. Quant à la condition (iii), on obtient l'inégalité encore plus forte suivante :

(iii′) $|x + y|_p \leqslant \max\{|x|_p, |y|_p\}$ (inégalité ultramétrique).

En effet, soit $r = p^k \frac{a}{b}$ et $s = p^\ell \frac{c}{d}$, sous l'hypothèse $k \geqslant \ell$, c'est-à-dire $|r|_p = p^{-k} \leqslant p^{-\ell} = |s|_p$. On trouve alors :

$$\begin{aligned} |r + s|_p &= \left|p^k \frac{a}{b} + p^\ell \frac{c}{d}\right|_p = \left|p^\ell (p^{k-\ell} \frac{a}{b} + \frac{c}{d})\right|_p \\ &= p^{-\ell} \left|\frac{p^{k-\ell} ad + bc}{bd}\right|_p \leqslant p^{-\ell} = \max\{|r|_p, |s|_p\}, \end{aligned}$$

puisque le dénominateur bd est premier avec p. Le raisonnement qui précède conduit aussi à :

(iv) $|x + y|_p = \max\{|x|_p, |y|_p\}$ dès que $|x|_p \neq |y|_p$,

mais nous allons montrer plus loin que cette propriété est la conséquence de (iii′).

Toute fonction $v : K \to \mathbb{R}_+$ définie sur un corps K vérifiant :

(i) $v(x) = 0$ si et seulement si $x = 0$,
(ii) $v(xy) = v(x)v(y)$ et
(iii′) $v(x + y) \leqslant \max\{v(x), v(y)\}$ (inégalité ultramétrique)

pour tout $x, y \in K$ est appelée une *valeur absolue ultramétrique* (ou *non archimédienne*) sur K.

Pour toute valeur absolue v ainsi définie, on observe que $v(1) = v(1)v(1)$, donc que $v(1) = 1$. Puis $1 = v(1) = v((-1)(-1)) = [v(-1)]^2$, donc $v(-1) = 1$. En utilisant, (ii) on obtient ainsi $v(-x) = v(x)$ pour tout x et $v(x^{-1}) = v(x)^{-1}$ pour $x \neq 0$.

Tout corps est muni d'une valeur absolue *triviale* qui envoie tout élément non nul sur 1. Par ailleurs, si v est une valeur absolue, alors v^t en est une aussi pour tout nombre réel positif t. Pour $\mathbb{Q}$, on dispose donc de la valeur absolue usuelle (archimédienne) et des valeurs absolues p-adiques (ultramétriques), ainsi que de leurs puissances. Un célèbre théorème dû à Ostrowski établit que toutes les valeurs absolues non triviales de $\mathbb{Q}$ sont de cette forme.

Comme promis, montrons que la propriété essentielle suivante :

(iv) $v(x+y) = \max\{v(x), v(y)\}$ si $v(x) \neq v(y)$

est vérifiée par toute valeur absolue ultramétrique. Pour fixer les idées, supposons que $v(x) < v(y)$. Alors :

$$\begin{aligned} v(y) &= v((x+y)-x) &&\leqslant \max\{v(x+y), v(x)\} = v(x+y) \\ &&&\leqslant \max\{v(x), v(y)\} = v(y)\,. \end{aligned}$$

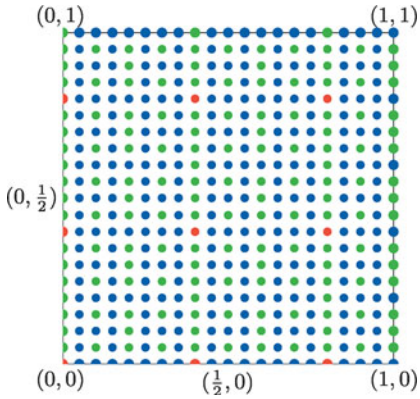
Les inégalités résultent toutes de l'application de (iii′). La première égalité est évidente. Les deux autres égalités sont la conséquence de l'hypothèse $v(x) < v(y)$. Ainsi $v(x+y) = v(y) = \max\{v(x), v(y)\}$.

La propriété (iv) conjuguée avec $v(-x) = v(x)$ implique aussi que $v(a \pm b_1 \pm b_2 \pm \cdots \pm b_\ell) = v(a)$ si $v(a) > v(b_i)$ pour tout i.

La belle idée de Monsky pour traiter la question du découpage d'un carré en triangles utilise le prolongement de la valeur absolue 2-adique $|x|_2$ en une valeur absolue v sur $\mathbb{R}$, où « prolongement » signifie que l'on requiert que $v(x) = |x|_2$ pour tout x de $\mathbb{Q}$. Un tel prolongement ultramétrique existe mais cela ne relève pas du domaine de l'algèbre usuelle. Dans la suite, nous présentons la démonstration de Monsky dans une version rédigée par Hendrik Lenstra qui utilise des arguments nettement moins puissants. La démonstration ne requiert qu'une valeur absolue v à valeurs dans un « groupe ordonné » arbitraire — et non pas nécessairement dans $(\mathbb{R}_+^*, \cdot, <)$ — et telle que $v(\frac{1}{2}) > 1$. La définition et l'existence d'une telle valeur absolue sont établies dans l'appendice présent à la fin de ce chapitre.

Ici nous nous bornons à constater que toute valeur absolue satisfaisant à la relation $v(\frac{1}{2}) > 1$ satisfait à $v(\frac{1}{n}) = 1$ pour les entiers n impairs. En effet, $v(\frac{1}{2}) > 1$ signifie que $v(2) < 1$ et donc que $v(2k) < 1$ en appliquant (iii′) et en raisonnant par récurrence sur k. Il s'ensuit que $v(2k+1) = 1$ d'après (iv) et donc encore que $v(\frac{1}{2k+1}) = 1$ d'après (ii).

Théorème de Monsky. *Il n'est pas possible de découper un carré en un nombre impair de triangles d'aires égales.*



■ **Preuve.** Dans la suite, on construit une 3-coloration particulière du plan qui présente des propriétés étonnantes. L'une de ces propriétés est que l'aire de tout triangle dont les sommets sont affectés de trois couleurs distinctes — appelé *triangle tricolore* dans la suite — présente une valeur absolue (selon v) plus grande que 1, si bien que son aire ne peut être $\frac{1}{n}$ pour n impair. On vérifie ensuite que tout découpage du carré en triangles doit comporter au moins un tel triangle tricolore.

On affecte une couleur aux points (x, y) du plan affine réel en en considérant la coordonnée du triplet $(x, y, 1)$ qui a la plus grande valeur absolue selon v. Ce maximum peut se produire pour une seule coordonnée, pour deux coordonnées ou même pour les trois coordonnées. Les couleurs (bleu, vert ou rouge) vont être significatives de la coordonnée des triplets $(x, y, 1)$ pour laquelle le maximum de v se produit en premier :

$$(x, y) \text{ est colorié en } \begin{cases} \text{bleu} & \text{si } v(x) \geqslant v(y),\ v(x) \geqslant v(1), \\ \text{vert} & \text{si } v(x) < v(y),\ v(y) \geqslant v(1), \\ \text{rouge} & \text{si } v(x) < v(1),\ v(y) < v(1). \end{cases}$$

On affecte ainsi une couleur unique à chaque point du plan. La figure représentée dans la marge montre la couleur de chaque point du carré unité dont les coordonnées sont des fractions de la forme $\frac{k}{20}$.

Le résultat suivant constitue la première étape de la démonstration.

Lemme 1. *Pour tout point bleu $p_b = (x_b, y_b)$, tout point vert $p_g = (x_g, y_g)$ et tout point rouge $p_r = (x_r, y_r)$, la valeur absolue (selon v) du déterminant*

$$\begin{vmatrix} x_b & y_b & 1 \\ x_g & y_g & 1 \\ x_r & y_r & 1 \end{vmatrix}$$

est supérieure ou égale à 1.

■ **Preuve.** Le déterminant considéré est une somme de six termes. Parmi eux figure $x_b y_g 1$ qui est le produit des éléments diagonaux. Par définition de la règle de coloriage appliquée, chacun des termes diagonaux est l'élément de valeur absolue maximale de la ligne à laquelle il appartient. Ainsi en comparant au dernier élément de chaque ligne (qui est 1), on obtient :

$$v(x_b y_g 1) \;=\; v(x_b)v(y_g)v(1) \;\geqslant\; v(1)v(1)v(1) \;=\; 1.$$

Chacun des cinq autre termes du déterminant est un produit de trois éléments de la matrice, un élément de chaque ligne, affecté d'un signe qui ne joue pas de rôle pour la valeur absolue. Chacun de ces termes comporte au moins un terme situé en dessous de la diagonale dont la valeur absolue est inférieure à la valeur absolue de l'élément diagonal situé sur la même ligne. Ainsi chacun des cinq autres termes du déterminant a une valeur absolue qui est strictement inférieure au terme obtenu comme produit des éléments diagonaux. D'après la propriété (iv) des valeurs absolues ultramétriques,

on trouve que la valeur absolue du déterminant s'obtient comme la valeur absolue du terme issu de la diagonale principale :

$$v\Big(\begin{vmatrix} x_b & y_b & 1 \\ x_g & y_g & 1 \\ x_r & y_r & 1 \end{vmatrix}\Big) \;=\; v(x_b y_g 1) \;\geqslant\; 1.$$

□

Corollaire. *Toute droite du plan contient des points d'au plus deux couleurs différentes.*
L'aire d'un triangle tricolore ne peut être 0 *et ne peut être* $\frac{1}{n}$ *pour* n *entier impair.*

■ **Preuve.** L'aire d'un triangle dont les sommets sont respectivement un point bleu p_b, un point vert p_g et un point rouge p_r est la valeur absolue de :

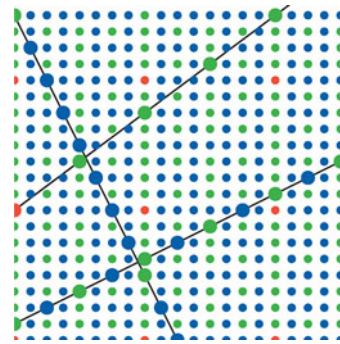
$$\tfrac{1}{2}((x_b - x_r)(y_g - y_r) - (x_g - x_r)(y_b - y_r)),$$

qui, au signe près, est la moitié du déterminant du lemme 1.

Les trois sommets ne peuvent être alignés car le déterminant ne peut pas être nul puisque $v(0) = 0$. L'aire du triangle considéré ne peut être $\frac{1}{n}$ puisque, si tel était le cas, on obtiendrait $\pm\frac{2}{n}$ pour le déterminant et donc :

$$v(\pm\tfrac{2}{n}) \;=\; v(\tfrac{1}{2})^{-1} v(\tfrac{1}{n}) \;<\; 1$$

à cause de $v(\frac{1}{2}) > 1$ et $v(\frac{1}{n}) = 1$, ce qui contredirait le lemme 1. □



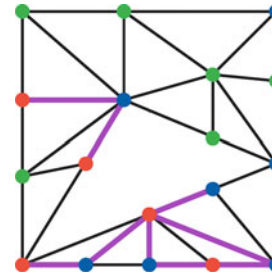
Pourquoi avoir proposé une telle coloration du plan ? Parce que nous allons à présent montrer que dans *tout* découpage du carré unité $S = [0,1]^2$ en triangles (de même aire ou non !) doit nécessairement figurer un triangle tricolore, lequel, selon le corollaire précédent, ne peut être d'aire $\frac{1}{n}$ pour n impair. C'est pourquoi le lemme suivant permet de terminer la démonstration du théorème de Monsky.

Lemme 2. *Tout découpage du carré unité* $S = [0,1]^2$ *en un nombre fini de triangles comporte un nombre impair de triangles tricolores — et donc comporte au moins un triangle tricolore.*

■ **Preuve.** Le calcul de dénombrement qui suit est vraiment brillant. L'idée est due à Emanuel Sperner et va apparaître à nouveau à l'occasion du « lemme de Sperner » au chapitre 25.

Un découpage étant donné, considérons les segments joignant des sommets voisins. Un segment est dit *segment rouge-bleu* si l'une de ses extrémités est rouge et que l'autre est bleue. Dans l'exemple de la figure représentée ci-dessous, les segments rouge-bleu ont été dessinés en violet.

Nous faisons maintenant les deux observations suivantes, en utilisant de manière répétée le corollaire qui indique que l'on trouve au plus deux couleurs sur une droite.



(A) Le côté inférieur (la base) du carré contient un nombre *impair* de segments rouge-bleu puisque $(0,0)$ est rouge et que $(1,0)$ est bleu et que tous les sommets situés entre eux sont soit rouges, soit bleus. Ainsi, sur le trajet entre l'extrémité rouge et l'extrémité bleue de la base du carré, il doit se produire un nombre impair de changements entre rouge et bleu. Les autres côtés du carré ne contiennent aucun segment rouge-bleu.

(B) Si un triangle T présente au plus deux couleurs parmi ses sommets, alors il contient un nombre *pair* de segments rouge-bleu sur son périmètre. Toutefois, tout triangle tricolore présente un nombre *impair* de segments rouge-bleu sur son périmètre.

En effet, il y a un nombre impair de segments rouge-bleu entre un sommet rouge et un sommet bleu d'un triangle mais, le cas échéant, un nombre pair de segments rouge-bleu entre des sommets présentant une autre combinaison de couleurs. Ainsi, un triangle tricolore présente un nombre impair de segments rouge-bleu dans son périmètre alors qu'un triangle non tricolore présente un nombre pair (2 ou 0) de couples de sommets dont la combinaison de couleur associe le rouge et le bleu.

À présent, considérons tous les triangles du découpage et pour chacun d'eux comptons le nombre de côtés rouge-bleu qu'il présente et additionnons ces nombres sur l'ensemble des triangles présents dans le découpage. Comme tout segment rouge-bleu situé à l'intérieur du carré est compté deux fois et qu'il y a un nombre impair de segments bleu-rouge sur le périmètre du carré, le nombre total de ces segments rouge-bleu est *impair*. On déduit alors de **(B)** qu'il doit y avoir un nombre impair de triangles tricolores dans le découpage. □

Appendice - Prolongements d'une valeur absolue

Il n'est pas du tout évident de voir que le prolongement d'une valeur absolue ultramétrique (ou non archimédienne) d'un corps à une extension de ce corps est toujours possible. Tel est pourtant le cas. En outre, ce type de prolongement peut être fait non seulement de $\mathbb{Q}$ à $\mathbb{R}$ mais plus généralement d'un corps K quelconque à un corps L contenant K. Ce résultat est connu sous le nom de « théorème de Chevalley » (voir, par exemple, l'ouvrage de Jacobson [1] sur la question).

Dans la suite, nous démontrons un résultat plus faible — mais suffisant pour l'application qui nous intéresse au cas des découpages du carré en triangles.

En effet, dans la preuve du théorème de Monsky proposée ici, nous n'avons pas utilisé l'addition pour $v : \mathbb{R} \to \mathbb{R}_+$ mais seulement la multiplication et l'ordre dans $\mathbb{R}_+$.

Ainsi, pour notre propos il est suffisant que les images non nulles de v soient éléments d'un groupe abélien ordonné (noté multiplicativement) $(G, \cdot, <)$ tel que cet ordre soit total et compatible avec la loi produit (c'est-à-dire que $a < b$ implique dans G que $ac < bc$ pour tous $a, b, c \in G$). Ayant adopté une notation multiplicative, l'élément neutre de G est naturellement noté 1. Pour définir une valeur absolue, nous considérons un élément particulier, noté 0, tel que $0 \notin G$ et que $0a = 0$ ainsi que $0 < a$ est vérifié pour tout $a \in G$. Bien sûr, l'exemple le plus naturel de groupe abélien totalement ordonné est $(\mathbb{R}_+^*, \cdot, \leqslant)$ muni de l'ordre usuel, et l'exemple le plus simple pour $\{0\} \cup G$ est $(\mathbb{R}_+, \cdot)$.

Définition. Soit K un corps. Une *valeur absolue ultramétrique* (ou *valeur absolue non archimédienne*) v à valeurs dans un groupe abélien totalement ordonné G est une application $v : K \to \{0\} \cup G$ telle que :

(i) $v(x) = 0 \iff x = 0$,
(ii) $v(xy) = v(x)v(y)$,
(iii′) $v(x + y) \leqslant \max\{v(x), v(y)\}$ et
(iv) $v(x + y) = \max\{v(x), v(y)\}$ si $v(x) \neq v(y)$

pour tout $x, y \in K$.

La quatrième condition dans cette énumération se déduit, comme précédemment, des trois premières. Par ailleurs, parmi les conséquences immédiates de cette définition, retenons que si $v(x) < 1$ et $x \neq 0$, alors $v(x^{-1}) = v(x)^{-1} > 1$.

Nous pouvons désormais établir le résultat suivant.

Théorème. *Il existe au moins une valeur absolue ultramétrique définie sur le corps des nombres réels $\mathbb{R}$ à valeurs dans un groupe abélien totalement ordonné G donné*

$$v : \mathbb{R} \to \{0\} \cup G$$

telle que $v(\frac{1}{2}) > 1$.

■ **Preuve.** Nous commençons par mettre en relation une valeur absolue quelconque sur un corps avec un sous-anneau de ce corps (tous les sous-anneaux considérés ici contiennent l'élément neutre 1).
Soit $v : K \to \{0\} \cup G$ est une valeur absolue et soient les ensembles :

$$R := \{x \in K : v(x) \leqslant 1\}, \qquad U := \{x \in K : v(x) = 1\}.$$

R est manifestement un sous-anneau de K, appelé *anneau de valuation* associé à v. En outre, $v(xx^{-1}) = v(1) = 1$ implique que $v(x) = 1$ si et seulement si $v(x^{-1}) = 1$. Ainsi U est-il l'ensemble des éléments inversibles de R. En particulier, U est un sous-groupe de $K^\times$ où $K^\times := K \setminus \{0\}$ est le groupe multiplicatif de K. En posant $R^{-1} := \{x^{-1} : x \in R^*\}$, on trouve $K = R \cup R^{-1}$. En effet, si $x \notin R$ alors $v(x) > 1$ et donc $v(x^{-1}) < 1$, si bien que $x^{-1} \in R$. La condition $K = R \cup R^{-1}$ suffit à caractériser tous les anneaux de valuation d'un corps K donné. □

Lemme. *Un sous-anneau propre $R \subseteq K$ est un anneau de valuation attaché à une valeur absolue v à valeurs dans un groupe ordonné G, si et seulement si $K = R \cup R^{-1}$.*

■ **Preuve.** On vient de voir la preuve de l'implication directe. Afin de montrer l'implication réciproque, supposons à présent que $K = R \cup R^{-1}$. Comment construire le groupe G ? Si $v : K \to \{0\} \cup G$ désigne une valeur absolue correspondant à R, alors $v(x) < v(y)$ est vérifié si et seulement si $v(xy^{-1}) < 1$, c'est-à-dire si et seulement si $xy^{-1} \in R \setminus U$. De même, $v(x) = v(y)$ si et seulement si $xy^{-1} \in U$, soit encore si $xU = yU$ en tant que classes (éléments) du groupe quotient $K^\times/U$.
Voici donc que se dessine devant nous le chemin naturel pour la démonstration. Considérons le groupe quotient $G := K^\times/U$ et la relation d'ordre définie sur G par :

$$xU < yU \quad :\Longleftrightarrow \quad xy^{-1} \in R \setminus U.$$

Nous laissons au lecteur l'exercice intéressant qui consiste à vérifier que cette relation fait bien de G un groupe ordonné.
L'application $v : K \to \{0\} \cup G$ se définit alors naturellement par :

$$v(0) = 0, \quad \text{et} \quad v(x) := xU \ \text{ pour } x \neq 0.$$

On montre facilement que v satisfait aux axiomes (i) à (iii′) et que R est l'anneau de valuation associé à v. □

Afin de démontrer le théorème, il est désormais suffisant de trouver un anneau de valuation $B \subset \mathbb{R}$ tel que $\frac{1}{2} \notin B$.

Proposition. *Tout sous-anneau $B \subset \mathbb{R}$ maximal (au sens de l'inclusion) vérifiant $\frac{1}{2} \notin B$ est un anneau de valuation.*

$\mathbb{Z} \subset \mathbb{R}$ est un tel sous-anneau vérifiant $\frac{1}{2} \notin \mathbb{Z}$ mais il n'est pas maximal.

Commençons par constater qu'il existe bien au moins un sous-anneau *maximal* $B \subset \mathbb{R}$ vérifiant $\frac{1}{2} \notin B$. Ce résultat n'est pas vraiment évident ; cela

résulte d'une application classique du Lemme de Zorn (voir l'encadré qui suit). En effet, si l'on dispose d'une chaîne croissante de sous-anneaux $B_i \subset \mathbb{R}$ qui ne contiennent pas $\frac{1}{2}$, alors cette chaîne admet une borne supérieure qui n'est autre que l'union de tous les sous-anneaux B_i. Cette borne supérieure est elle-même un sous-anneau qui ne contient pas $\frac{1}{2}$.

Le lemme de Zorn

Le lemme de Zorn est de première importance en algèbre et dans d'autres domaines des mathématiques lorsqu'on cherche à construire des structures maximales. Il joue aussi un rôle essentiel pour les fondements logiques des mathématiques.

Lemme. *Soit $(P, \leqslant)$ un ensemble non vide partiellement ordonné tel que toute chaîne croissante (a_i) présente une borne supérieure b (c'est-à-dire que b vérifie $a_i \leqslant b$ pour tout i). Alors P admet un élément maximal M (c'est-à-dire tel qu'il n'existe pas de $c \in P$ tel que $M < c$).*

Afin de démontrer la proposition, supposons que $B \subset \mathbb{R}$ est un sous-anneau maximal ne contenant pas $\frac{1}{2}$. Si B n'est pas un anneau de valuation, alors il existe au moins un élément α dans $\mathbb{R}\backslash(B \cup B^{-1})$. Désignons par $B[\alpha]$ le sous-anneau engendré par $B \cup \{\alpha\}$, c'est-à-dire l'ensemble de tous les nombres réels pouvant être écrits comme des polynômes en α à coefficients dans B. Soit $2B \subseteq B$ l'ensemble des éléments de la forme $2b$, $b \in B$. L'ensemble $2B$ est inclus dans B, donc $2B[\alpha] \subseteq B[\alpha]$ et $2B[\alpha^{-1}] \subseteq B[\alpha^{-1}]$. Comme $1 \in B$, si l'on avait $2B[\alpha] \neq B[\alpha]$ ou $2B[\alpha^{-1}] \neq B[\alpha^{-1}]$, alors cela entraînerait que $\frac{1}{2} \notin B[\alpha]$ ou respectivement que $\frac{1}{2} \notin B[\alpha^{-1}]$, contredisant la maximalité de $B \subset \mathbb{R}$ comme sous-anneau ne contenant pas $\frac{1}{2}$.
On se retrouve donc avec $2B[\alpha] = B[\alpha]$ et $2B[\alpha^{-1}] = B[\alpha^{-1}]$. Cela implique que $1 \in B$ peut s'écrire sous la forme :

$$1 = 2u_0 + 2u_1\alpha + \cdots + 2u_m\alpha^m \quad \text{avec } u_i \in B, \quad (1)$$

et aussi sous la forme :

$$1 = 2v_0 + 2v_1\alpha^{-1} + \cdots + 2v_n\alpha^{-n} \text{ avec } v_i \in B, \quad (2)$$

et après multiplication par α^n et soustraction de $2v_0\alpha^n$ aux deux membres de cette dernière égalité, cela nous conduit à :

$$(1 - 2v_0)\alpha^n = 2v_1\alpha^{n-1} + \cdots + 2v_{n-1}\alpha + 2v_n. \quad (3)$$

Supposons que ces représentations soient choisies de sorte que m et n soient aussi petits que possibles. et que $m \geqslant n$ (sinon, on échange α avec α^{-1} et (1) avec (2)).
À présent multiplions (1) par $1 - 2v_0$ et ajoutons $2v_0$ aux deux membres de l'égalité ; on trouve :

$$1 = 2(u_0(1 - 2v_0) + v_0) + 2u_1(1 - 2v_0)\alpha + \cdots + 2u_m(1 - 2v_0)\alpha^m.$$

Si dans cette dernière égalité on remplace $(1 - 2v_0)\alpha^m$ par l'expression obtenue en multipliant les deux membres de (3) par α^{m-n}, alors on trouve une égalité qui donne l'expression de $1 \in B$ comme un polynôme de $2B[\alpha]$ de degré inférieur ou égal à $m - 1$. Ceci contredit la minimalité de m et termine la démonstration de l'assertion. □

Bibliographie

[1] N. JACOBSON : *Lectures in Abstract Algebra, Part III : Theory of Fields and Galois Theory,* Graduate Texts in Mathematics 32. Springer, New York 1975.

[2] P. MONSKY : *On dividing a square into triangles,* Amer. Math. Monthly **77** (1970), 161-164.

[3] F. RICHMAN & J. THOMAS : *Problem 5471,* Amer. Math. Monthly **74** (1967), 329.

[4] S. K. STEIN & S. SZABÓ *:* Algebra and Tiling : Homomorphisms in the Service of Geometry, Carus Math. Monographs **25**, MAA, Washington DC 1994.

[5] J. THOMAS : *A dissection problem,* Math. Magazine **41** (1968), 187-190.

Un théorème de Pólya sur les polynômes

Chapitre 21

Parmi les nombreuses contributions de Pólya à l'analyse, le résultat suivant a toujours été celui que préférait Erdős, à la fois pour son contenu surprenant et pour la beauté de sa preuve. Soit :

$$f(z) \;=\; z^n + b_{n-1}z^{n-1} + \ldots + b_0$$

un polynôme complexe de degré $n \geqslant 1$ et de coefficient dominant 1. Associons à $f(z)$ l'ensemble :

$$\mathcal{C} \;:=\; \{z \in \mathbb{C} : |f(z)| \leqslant 2\}$$

c'est-à-dire que $\mathcal{C}$ est l'ensemble des points qui sont envoyés par f dans le disque de rayon 2 et de centre l'origine du plan complexe. Si $n = 1$, le domaine $\mathcal{C}$ est tout simplement un disque de diamètre 4.

Par un argument étonnamment simple, Pólya a découvert que cet ensemble $\mathcal{C}$ possède la belle propriété suivante :

> *Prenons une droite quelconque L du plan complexe et considérons la projection orthogonale $\mathcal{C}_L$ de l'ensemble $\mathcal{C}$ sur L. Alors, la longueur totale d'une telle projection n'excède jamais 4.*

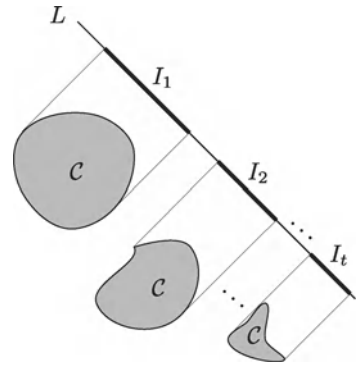
Que signifie le fait que la longueur totale de la projection $\mathcal{C}_L$ est au plus 4 ? Nous allons voir que $\mathcal{C}_L$ *est une réunion finie d'intervalles disjoints* $I_1, \ldots, I_t$; la condition signifie donc que $\ell(I_1) + \ldots + \ell(I_t) \leqslant 4$, où $\ell(I_j)$ est la longueur habituelle d'un intervalle.

Quitte à faire une rotation du plan, il suffit de considérer le cas où L est l'axe réel du plan complexe. Avec ces remarques à l'esprit, énonçons le résultat de Pólya.

George Pólya

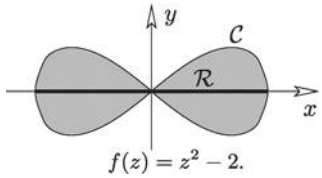
Théorème 1. *Soit $f(z)$ un polynôme complexe de degré au moins égal à 1 et de coefficient dominant 1. Soit $\mathcal{C} = \{z \in \mathbb{C} : |f(z)| \leqslant 2\}$ et $\mathcal{R}$ la projection orthogonale de $\mathcal{C}$ sur l'axe réel. Alors, il existe des intervalles $I_1, \ldots, I_t$ de la droite réelle dont la réunion recouvre $\mathcal{R}$ et vérifiant :*

$$\ell(I_1) + \ldots + \ell(I_t) \;\leqslant\; 4$$

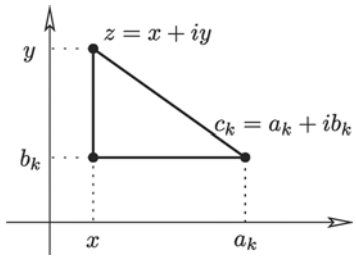


Il est clair que la borne 4 du théorème est atteinte pour $n = 1$. Pour mieux nous imprégner du problème, examinons le polynôme $f(z) = z^2 - 2$, qui atteint lui aussi la borne 4. Si $z = x + iy$ est un nombre complexe, alors x est sa projection orthogonale sur la droite réelle. Ainsi :

$$\mathcal{R} = \{x \in \mathbb{R} : x + iy \in \mathcal{C} \text{ pour un certain } y\}$$



$f(z) = z^2 - 2.$



Le lecteur peut facilement montrer que si $f(z) = z^2 - 2$, on a : $x + iy \in \mathcal{C}$ si et seulement si :

$$(x^2 + y^2)^2 \leqslant 4(x^2 - y^2)$$

Par suite, $x^4 \leqslant (x^2 + y^2)^2 \leqslant 4x^2$ donc : $x^2 \leqslant 4$, c'est-à-dire $|x| \leqslant 2$. D'autre part, tout $z = x \in \mathbb{R}$ tel que $|x| \leqslant 2$ vérifie $|z^2 - 2| \leqslant 2$. On voit que $\mathcal{R}$ est précisément l'intervalle $[-2, 2]$ de longueur 4.

Une première étape vers la preuve consiste à écrire $f(z)$ sous la forme $f(z) = (z - c_1) \cdots (z - c_n)$, avec $c_k = a_k + ib_k$, et à considérer le polynôme *réel* $p(x) = (x - a_1) \cdots (x - a_n)$. Soit $z = x + iy \in \mathcal{C}$. Alors le théorème de Pythagore implique que :

$$|x - a_k|^2 + |y - b_k|^2 = |z - c_k|^2$$

et donc $|x - a_k| \leqslant |z - c_k|$ pour tout k, c'est-à-dire :

$$|p(x)| = |x - a_1| \cdots |x - a_n| \leqslant |z - c_1| \cdots |z - c_n| = |f(z)| \leqslant 2$$

Ainsi, $\mathcal{R}$ est contenu dans l'ensemble $\mathcal{P} = \{x \in \mathbb{R} : |p(x)| \leqslant 2\}$. Si nous pouvons montrer que ce dernier ensemble est recouvert par des intervalles de longueur totale au plus égale à 4, nous aurons terminé. Notre principal théorème 1 sera donc une conséquence du résultat suivant.

Théorème 2. *Soit $p(x)$ un polynôme réel de degré $n \geqslant 1$ et de coefficient dominant 1, dont toutes les racines sont réelles.*
Alors l'ensemble $\mathcal{P} = \{x \in \mathbb{R} : |p(x)| \leqslant 2\}$ peut être recouvert par des intervalles de longueur totale au plus égale à 4.

Comme Pólya le montre dans son article [2], le théorème 2 est à son tour une conséquence du célèbre résultat suivant dû à Chebyshev. Pour que ce chapitre soit autonome, nous en proposons une preuve en appendice (en nous inspirant du bel exposé de Pólya et Szegő).

Timbre soviétique de 1946 à l'effigie de Pavnuty Chebyshev.

Théorème de Chebyshev.
Soit $p(x)$ un polynôme réel de degré $n \geqslant 1$ et de coefficient dominant 1. Alors :

$$\max_{-1 \leqslant x \leqslant 1} |p(x)| \geqslant \frac{1}{2^{n-1}}$$

Notons tout d'abord la conséquence immédiate suivante :

Corollaire. *Soit $p(x)$ un polynôme réel de degré $n \geqslant 1$ et de coefficient dominant 1. Supposons que $|p(x)| \leqslant 2$ pour tout x de l'intervalle $[a, b]$. Alors : $b - a \leqslant 4$.*

■ **Preuve.** Considérons la transformation $y = \frac{2}{b-a}(x - a) - 1$. Elle envoie l'intervalle $[a, b]$ de l'axe des x sur l'intervalle $[-1, 1]$ de l'axe des y. Le polynôme correspondant :

$$q(y) = p(\tfrac{b-a}{2}(y + 1) + a)$$

a pour coefficient dominant $(\frac{b-a}{2})^n$ et vérifie :

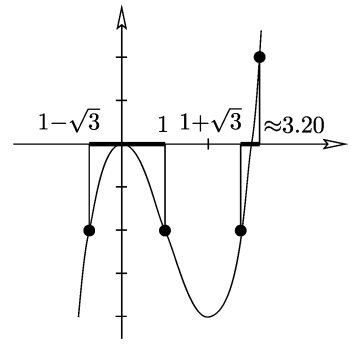
$$\max_{-1\leqslant y\leqslant 1} |q(y)| \;=\; \max_{a\leqslant x\leqslant b} |p(x)|$$

Le théorème de Chebyshev implique :

$$2 \;\geqslant\; \max_{a\leqslant x\leqslant b} |p(x)| \;\geqslant\; (\tfrac{b-a}{2})^n \tfrac{1}{2^{n-1}} \;=\; 2(\tfrac{b-a}{4})^n$$

et donc : $b-a \leqslant 4$, comme nous le souhaitions. □

Ce corollaire nous rapproche beaucoup de l'énoncé du théorème 2. Si l'ensemble $\mathcal{P} = \{x : |p(x)| \leqslant 2\}$ est un *intervalle*, alors la longueur de $\mathcal{P}$ est au plus 4. Cependant, l'ensemble $\mathcal{P}$ peut ne pas être un intervalle, comme dans l'exemple représenté sur la figure ci-contre, où $\mathcal{P}$ est constitué de deux intervalles.



Pour le polynôme :

$$p(x) = x^2(x-3)$$

nous obtenons :

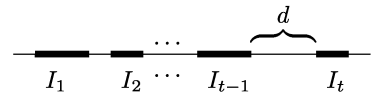
$$\mathcal{P} = [1-\sqrt{3}, 1] \cup [1+\sqrt{3}, \approx 3.2]\,.$$

Que pouvons nous dire sur $\mathcal{P}$? Puisque $p(x)$ est une fonction continue, nous savons en tout cas que $\mathcal{P}$ est la réunion d'intervalles fermés disjoints $I_1, I_2, \ldots$ et que $p(x)$ prend la valeur 2 ou -2 à chaque extrémité de l'intervalle I_j. Cela implique qu'il n'y a qu'un nombre fini d'intervalles $I_1, \ldots, I_t$ puisque $p(x)$ ne peut prendre la même valeur qu'un nombre fini de fois.

L'idée prodigieuse de Pólya a été de construire un autre polynôme $\tilde{p}(x)$ de degré n, dont le coefficient dominant est 1, tel que $\widetilde{\mathcal{P}} = \{x : |\tilde{p}(x)| \leqslant 2\}$ soit un *intervalle* de longueur au moins égale à $\ell(I_1) + \ldots + \ell(I_t)$. Le corollaire implique alors que $\ell(I_1) + \ldots + \ell(I_t) \leqslant \ell(\widetilde{\mathcal{P}}) \leqslant 4$ et l'on a terminé.

■ **Preuve du théorème 2.** Considérons $p(x) = (x-a_1)\cdots(x-a_n)$ et $\mathcal{P} = \{x \in \mathbb{R} : |p(x)| \leqslant 2\} = I_1 \cup \ldots \cup I_t$. Nous avons rangé les intervalles I_j de sorte que I_1 soit l'intervalle qui se trouve le plus à gauche et I_t celui qui se trouve le plus à droite. Nous affirmons d'abord que tout intervalle I_j contient une racine de $p(x)$. En effet, nous savons que $p(x)$ prend les valeurs 2 ou -2 aux extrémités de I_j. Si une valeur est 2 et l'autre -2, alors il existe certainement une racine dans I_j. Supposons donc que $p(x) = 2$ aux deux extrémités à la fois (le cas -2 étant analogue). Supposons que $b \in I_j$ soit un point tel que $p(x)$ atteigne son minimum dans I_j. Alors $p'(b) = 0$ et $p''(b) \geqslant 0$. Si $p''(b) = 0$, alors b est une racine multiple de $p'(x)$ donc une racine de $p(x)$ d'après le résultat 1 de l'encadré. D'autre part, si $p''(b) > 0$, alors $p(b) \leqslant 0$ d'après le résultat 2 du même encadré. Ainsi, lorsque $p(b) = 0$, nous avons notre racine et, lorsque $p(b) < 0$, nous obtenons une racine dans l'intervalle qui s'étend de b à l'une des extrémités de I_j.

Voici l'idée finale de la preuve : Soit $I_1, \ldots, I_t$ les intervalles précédents. Supposons que l'intervalle qui se trouve le plus à droite I_t contienne m racines de $p(x)$, comptées avec leur multiplicité. Si $m = n$, alors I_t est le seul intervalle (d'après ce que nous venons de montrer) et la démonstration est terminée. Supposons donc $m < n$; soit d la distance entre I_{t-1} et I_t comme l'indique la figure. Soit $b_1, \ldots, b_m$ les racines de $p(x)$ qui appartiennent à I_t et $c_1, \ldots c_{n-m}$ les racines restantes. Écrivons $p(x) = q(x)r(x)$, avec



$q(x) = (x-b_1)\cdots(x-b_m)$, $r(x) = (x-c_1)\cdots(x-c_{n-m})$, et posons $p_1(x) = q(x+d)r(x)$. Le polynôme $p_1(x)$ est à nouveau de degré n et de coefficient dominant 1. Pour $x \in I_1 \cup \ldots \cup I_{t-1}$, nous avons $|x+d-b_i| < |x-b_i|$ pour tout i donc $|q(x+d)| < |q(x)|$. Par suite :

$$|p_1(x)| \leqslant |p(x)| \leqslant 2 \qquad \text{si } x \in I_1 \cup \ldots \cup I_{t-1}$$

D'autre part, si $x \in I_t$, on voit qu'alors $|r(x-d)| \leqslant |r(x)|$ et donc :

$$|p_1(x-d)| = |q(x)||r(x-d)| \leqslant |p(x)| \leqslant 2$$

ce qui signifie que $I_t - d \subseteq \mathcal{P}_1 = \{x : |p_1(x)| \leqslant 2\}$.

En résumé, nous voyons que $\mathcal{P}_1$ contient $I_1 \cup \ldots \cup I_{t-1} \cup (I_t - d)$ et que sa longueur totale est donc supérieure ou égale à celle de $\mathcal{P}$. Remarquons maintenant qu'avec le passage de $p(x)$ à $p_1(x)$ les intervalles I_{t-1} et $I_t - d$ se confondent en un seul intervalle. Ainsi, les intervalles $J_1, \ldots, J_s$ de $p_1(x)$ formant $\mathcal{P}_1$ ont une longueur totale supérieure ou égale à $\ell(I_1) + \ldots + \ell(I_t)$ et l'intervalle le plus à droite J_s contient plus de m racines de $p_1(x)$. En répétant cette procédure au plus $t-1$ fois, nous obtenons finalement un polynôme $\tilde{p}(x)$ où $\widetilde{\mathcal{P}} = \{x : |\tilde{p}(x)| \leqslant 2\}$ est un intervalle de longueur $\ell(\widetilde{\mathcal{P}}) \geqslant \ell(I_1) + \ldots + \ell(I_t)$. La preuve est terminée. □

Deux résultats sur les polynômes à racines réelles

Soit $p(x)$ un polynôme non constant n'ayant que des racines réelles.

Proposition 1. *Si b est une racine multiple de $p'(x)$, alors b est aussi une racine de $p(x)$.*

■ **Preuve.** Soient $b_1 < \ldots < b_r$ les racines de $p(x)$ de multiplicités $s_1, \ldots, s_r$, $\sum_{j=1}^{r} s_j = n$. De $p(x) = (x-b_j)^{s_j} h(x)$ on déduit que b_j est une racine de $p'(x)$ si $s_j \geqslant 2$ et que la multiplicité de b_j dans $p'(x)$ est $s_j - 1$. En outre, il y a une racine de $p'(x)$ entre b_1 et b_2, une autre racine entre b_2 et b_3, ..., et une entre b_{r-1} et b_r. Toutes ces racines doivent être des racines *simples*, puisque $\sum_{j=1}^{r}(s_j-1)+(r-1)$ atteint déjà le degré $n-1$ de $p'(x)$. En conséquence, les racines *multiples* de $p'(x)$ ne peuvent se trouver que parmi les racines de $p(x)$. □

Proposition 2. *On a* $p'(x)^2 \geqslant p(x)p''(x)$ *pour tout* $x \in \mathbb{R}$.

■ **Preuve.** Si x est une racine a_k de p, alors le résultat est évident. Supposons que x ne soit pas une racine de p. p étant de la forme $p(x) = \prod_{k=1}^{n} (x - a_k)$, la formule de dérivation d'un produit conduit à :

$$\frac{p'(x)}{p(x)} = \sum_{k=1}^{n} \frac{1}{x - a_k}$$

et, en dérivant à nouveau, on obtient :

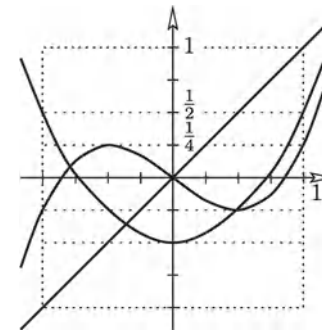
$$\frac{p''(x)p(x) - p'(x)^2}{p(x)^2} = -\sum_{k=1}^{n} \frac{1}{(x - a_k)^2} \leqslant 0 \qquad \square$$

Appendice - Le théorème de Chebyshev

Théorème. *Soit* $p(x)$ *un polynôme réel de degré* $n \geqslant 1$ *de coefficient dominant* 1. *Alors,*

$$\max_{-1 \leqslant x \leqslant 1} |p(x)| \geqslant \frac{1}{2^{n-1}}$$

Avant de commencer, examinons quelques exemples pour lesquels il y a égalité. Dans la marge sont représentés les graphes de polynômes de degrés 1, 2 et 3 pour lesquels l'égalité est réalisée. En fait, nous allons voir que pour chaque degré il existe exactement un polynôme réalisant l'égalité du théorème de Chebyshev.



Les polynômes :
$p_1(x) = x$,
$p_2(x) = x^2 - \frac{1}{2}$ et
$p_3(x) = x^3 - \frac{3}{4}x$
réalisent l'égalité dans le théorème de Chebyshev.

■ **Preuve.** Considérons un polynôme réel $p(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_0$ de coefficient dominant 1. Puisque nous nous intéressons à l'intervalle $-1 \leqslant x \leqslant 1$, posons $x = \cos\vartheta$ et notons $g(\vartheta) := p(\cos\vartheta)$ le polynôme résultant en $\cos\vartheta$,

$$g(\vartheta) = (\cos\vartheta)^n + a_{n-1}(\cos\vartheta)^{n-1} + \ldots + a_0 \qquad (1)$$

La preuve se déroule en deux étapes, chacune d'elles constituant un résultat classique et intéressant en lui-même.

(A) Nous exprimons $g(\vartheta)$ comme un *polynôme trigonométrique* en cosinus, c'est-à-dire un polynôme de la forme :

$$g(\vartheta) = b_n \cos n\vartheta + b_{n-1} \cos(n-1)\vartheta + \ldots + b_1 \cos\vartheta + b_0 \qquad (2)$$

avec $b_k \in \mathbb{R}$, et nous montrons que son coefficient dominant est $b_n = \frac{1}{2^{n-1}}$.

(B) Étant donné un polynôme trigonométrique en cosinus $h(\vartheta)$ d'ordre n (ce qui signifie que le coefficient λ_n de $\cos n\vartheta$ est non nul) :

$$h(\vartheta) \;=\; \lambda_n \cos n\vartheta + \lambda_{n-1}\cos(n-1)\vartheta + \ldots + \lambda_0 \qquad (3)$$

nous montrons que $|\lambda_n| \leqslant \max|h(\vartheta)|$, ce qui, appliqué à $g(\vartheta)$, démontrera alors le théorème.

Preuve de (A).
Pour passer de (1) à la représentation (2), nous devons exprimer toutes les puissances $(\cos\vartheta)^k$ comme des polynômes trigonométriques en cosinus. Par exemple, la formule d'addition des cosinus donne :

$$\cos 2\vartheta \;=\; \cos^2\vartheta - \sin^2\vartheta \;=\; 2\cos^2\vartheta - 1$$

et donc $\cos^2\vartheta \;=\; \frac{1}{2}\cos 2\vartheta + \frac{1}{2}$. Dans le cas d'une puissance arbitraire $(\cos\vartheta)^n$, on utilise les nombres complexes et la formule d'Euler :

$$\cos\vartheta = \frac{1}{2}(e^{i\vartheta} + e^{-i\vartheta})$$

À l'aide de la formule du binôme de Newton on trouve :

$$\begin{aligned}(\cos\vartheta)^n &= \frac{1}{2^n}(e^{i\vartheta} + e^{-i\vartheta})^n \\ &= \frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}e^{i(n-k)\vartheta}e^{-ik\vartheta}\end{aligned}$$

et en regroupant les termes deux par deux en partant de chaque extrémité de l'expression précédente et en remarquant que $\binom{n}{k} = \binom{n}{n-k}$ on est conduit à :

$$\begin{aligned}(\cos\vartheta)^n &= \frac{1}{2^n}\sum_{k=0}^{\lfloor n/2\rfloor}\binom{n}{k}(e^{ik\vartheta} + e^{-ik\vartheta}) \\ &= \frac{1}{2^{n-1}}\sum_{k=0}^{\lfloor n/2\rfloor}\binom{n}{k}\cos(k\vartheta)\end{aligned}$$

Ainsi, $(\cos\vartheta)^n$ peut s'écrire comme un polynôme trigonométrique en cosinus d'ordre n de coefficient dominant $b_n = \frac{1}{2^{n-1}}$.

Preuve de (B).
Soit $h(\vartheta)$ un polynôme trigonométrique en cosinus d'ordre n s'écrivant comme dans (3). Supposons, sans perte de généralité, que $\lambda_n > 0$. Posant $m(\vartheta) := \lambda_n \cos n\vartheta$, nous sommes conduits à :

$$m(\tfrac{k}{n}\pi) \;=\; (-1)^k\lambda_n \qquad \text{pour } k = 0, 1, \ldots, n.$$

Supposons, par l'absurde, que $\max|h(\vartheta)| < \lambda_n$. Alors :

$$m(\tfrac{k}{n}\pi) - h(\tfrac{k}{n}\pi) \;=\; (-1)^k\lambda_n - h(\tfrac{k}{n}\pi)$$

est positif pour k pair et négatif pour k impair pour $0 \leqslant k \leqslant n$. On en conclut que $m(\vartheta) - h(\vartheta)$ s'annule au moins n fois dans l'intervalle $[0, \pi]$. Cela ne peut pas se produire puisque $m(\vartheta) - h(\vartheta)$ est un polynôme trigonométrique en cosinus d'ordre $n-1$ et que le polynôme associé à cette expression trigonométrique ne peut s'annuler plus de $n-1$ fois.

La preuve de **(B)** est achevée. Nous avons donc terminé la démonstration du théorème de Chebyshev. □

Le lecteur peut facilement compléter cette analyse en montrant que $g_n(\vartheta) = \frac{1}{2^{n-1}} \cos n\vartheta$ est le *seul* polynôme trigonométrique en cosinus d'ordre n et de coefficient dominant 1 qui réalise l'égalité $\max |g(\vartheta)| = \frac{1}{2^{n-1}}$.
Les polynômes $T_n(x) = \cos n\vartheta$, $x = \cos \vartheta$ sont appelés *polynômes de Chebyshev* (de première espèce) ; ainsi, $\frac{1}{2^{n-1}} T_n(x)$ est l'unique polynôme unitaire de degré n tel que l'égalité ait lieu dans le théorème de Chebyshev.

Bibliographie

[1] P. L. CEBYCEV : *Œuvres,* Vol. I, Acad. Impériale des Sciences, St. Petersburg 1899, pp. 387-469.

[2] G. PÓLYA : *Beitrag zur Verallgemeinerung des Verzerrungssatzes auf mehrfach zusammenhängenden Gebieten,* Sitzungsber. Preuss. Akad. Wiss. Berlin (1928), 228-232 ; Collected Papers Vol. I, MIT Press 1974, 347-351.

[3] G. PÓLYA & G. SZEGŐ : *Problems and Theorems in Analysis, Vol. II,* Springer-Verlag, Berlin Heidelberg New York 1976 ; Reprint 1998.

Sur un lemme de Littlewood et Offord

Chapitre 22

Dans leur travail sur la distribution des racines des équations algébriques, Littlewood et Offord ont montré en 1943 le résultat suivant :

> *Soit $a_1, a_2, \ldots, a_n$ des nombres complexes tels que pour tout i, $|a_i| \geqslant 1$. Considérons les 2^n combinaisons linéaires $\sum_{i=1}^n \varepsilon_i a_i$ où $\varepsilon_i \in \{1, -1\}$. Alors le nombre de sommes $\sum_{i=1}^n \varepsilon_i a_i$ qui appartiennent à l'intérieur d'un disque de rayon 1 n'est pas supérieur à :*
>
> $$c \frac{2^n}{\sqrt{n}} \ln n \quad \textit{où } c \textit{ est une constante strictement positive.}$$

John E. Littlewood

Quelques années plus tard, Paul Erdős a amélioré cette borne en éliminant le terme $\ln n$. Toutefois, ce qui est plus intéressant, c'est qu'il a montré que c'est en fait une simple conséquence du théorème de Sperner (voir page 201).

Théorème de Sperner. *La taille de toute antichaîne de sous-ensembles d'un ensemble de cardinal n est au plus $\binom{n}{\lfloor n/2 \rfloor}$.*

Pour avoir une intuition de la preuve, examinons le cas où tous les a_i sont réels. Nous pouvons supposer que tous les a_i sont positifs (en changeant a_i en $-a_i$ et ε_i en $-\varepsilon_i$ chaque fois que a_i est négatif). Considérons un ensemble de combinaisons $\sum \varepsilon_i a_i$ qui appartiennent toutes à l'intérieur d'un intervalle de longueur 2. Soit $N = \{1, 2, \ldots, n\}$ l'ensemble des indices. Pour chaque $\sum \varepsilon_i a_i$, nous posons $I := \{i \in N : \varepsilon_i = 1\}$. Si $I \subsetneqq I'$ sont deux tels ensembles, nous pouvons conclure que :

$$\sum \varepsilon'_i a_i - \sum \varepsilon_i a_i \;=\; 2 \sum_{i \in I' \setminus I} a_i \;\geqslant\; 2$$

ce qui est contradictoire. Ainsi, les ensembles I forment une antichaîne et nous concluons, en utilisant le théorème de Sperner, qu'il y a au plus $\binom{n}{\lfloor n/2 \rfloor}$ combinaisons de ce type. La formule de Stirling (voir page 11) implique :

$$\binom{n}{\lfloor n/2 \rfloor} \;\leqslant\; c \frac{2^n}{\sqrt{n}} \quad \text{pour un certain } c > 0$$

Si n est pair et si tous les a_i valent 1, nous obtenons $\binom{n}{n/2}$ combinaisons $\sum_{i=1}^n \varepsilon_i a_i$ dont la somme est 0. En examinant l'intervalle $]-1, 1[$, on voit que le coefficient binomial fournit la borne *exacte*.

Dans le même article, Erdős a conjecturé que $\binom{n}{\lfloor n/2 \rfloor}$ est aussi la bonne borne pour les nombres complexes (il a seulement pu montrer la formule

avec $c\,2^n n^{-1/2}$ pour un certain c) et que la même borne est en fait valide pour des vecteurs $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ tels que $|\boldsymbol{a}_i| \geqslant 1$ dans un espace de Hilbert réel, en remplaçant le cercle de rayon 1 par la boule ouverte de rayon 1.

Erdős avait raison, mais il a fallu attendre vingt ans avant que Gyula Katona et Daniel Kleitman présentent indépendamment une preuve pour les nombres complexes (ou, ce qui est équivalent, pour le plan $\mathbb{R}^2$). Leurs preuves utilisaient explicitement le fait que le plan a pour dimension 2 : on ne voyait absolument pas comment elles pouvaient être étendues aux espaces vectoriels réels de dimension finie.

En 1970 Kleitman a montré la conjecture complète sur les espaces de Hilbert avec un argument d'une étonnante simplicité. En fait, il a même prouvé mieux. Son argument est un excellent exemple de ce qu'on peut faire lorsqu'on trouve la bonne hypothèse de récurrence.

Un mot de réconfort pour tous les lecteurs qui ne sont pas familiers avec la notion d'espace de Hilbert : nous n'avons pas véritablement besoin d'espaces de Hilbert généraux. Puisque nous travaillons uniquement avec un nombre fini de vecteurs $\boldsymbol{a}_i$, il est suffisant de considérer l'espace réel $\mathbb{R}^d$ avec le produit scalaire habituel. Voici le résultat de Kleitman.

Théorème. *Soit $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ des vecteurs de $\mathbb{R}^d$ de longueur supérieure ou égale à 1 et soit $R_1, \ldots, R_k$ k régions ouvertes de $\mathbb{R}^d$, où $|\boldsymbol{x} - \boldsymbol{y}| < 2$ pour tout $\boldsymbol{x}, \boldsymbol{y}$ appartenant à la même région R_i.*
Alors le nombre de combinaisons linéaires $\sum_{i=1}^n \varepsilon_i \boldsymbol{a}_i$, $\varepsilon_i \in \{1, -1\}$, qui peuvent appartenir à la réunion $\bigcup_i R_i$ de ces régions est au plus égal à la somme des k plus grands coefficients binomiaux $\binom{n}{j}$.
En particulier, si $k = 1$, on obtient la borne $\binom{n}{\lfloor n/2 \rfloor}$.

Avant de passer à la démonstration, il faut noter que la borne est exacte pour :

$$\boldsymbol{a}_1 = \ldots = \boldsymbol{a}_n = \boldsymbol{a} = (1, 0, \ldots, 0)$$

En effet, pour n pair, on obtient $\binom{n}{n/2}$ sommes égales à 0, $\binom{n}{n/2-1}$ sommes égales à $(-2)\boldsymbol{a}$, $\binom{n}{n/2+1}$ sommes égales à $2\boldsymbol{a}$ et ainsi de suite. En choisissant des boules de rayon 1 autour de :

$$-2\lceil \tfrac{k-1}{2} \rceil \boldsymbol{a}, \quad \ldots \quad (-2)\boldsymbol{a}, \quad \mathbf{0}, \quad 2\boldsymbol{a}, \quad \ldots \quad 2\lfloor \tfrac{k-1}{2} \rfloor \boldsymbol{a}$$

on obtient :

$$\binom{n}{\lfloor \frac{n-k+1}{2} \rfloor} + \ldots + \binom{n}{\frac{n-2}{2}} + \binom{n}{\frac{n}{2}} + \binom{n}{\frac{n+2}{2}} + \ldots + \binom{n}{\lfloor \frac{n+k-1}{2} \rfloor}$$

sommes appartenant à ces k boules ; cela constitue l'expression promise, puisque les plus grands coefficients binomiaux se répartissent vers le milieu (voir page 12). Un raisonnement similaire fonctionne aussi lorsque n est impair.

■ **Preuve.** Nous pouvons supposer, sans perte de généralité, que les régions R_i sont disjointes ; c'est ce que nous ferons désormais. La clé de la preuve est la récurrence sur les coefficients binomiaux qui nous dit comment sont reliés les plus grands coefficients binomiaux de n et $n-1$. Si l'on pose $r = \lfloor \frac{n-k+1}{2} \rfloor, s = \lfloor \frac{n+k-1}{2} \rfloor$, alors $\binom{n}{r}, \binom{n}{r+1}, \ldots, \binom{n}{s}$ sont les k plus grands coefficients binomiaux relatifs à n. La récurrence $\binom{n}{i} = \binom{n-1}{i} + \binom{n-1}{i-1}$ implique :

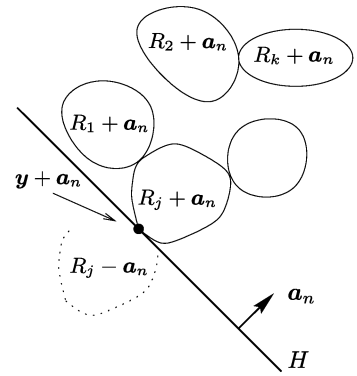
$$\begin{aligned} \sum_{i=r}^{s} \binom{n}{i} &= \sum_{i=r}^{s} \binom{n-1}{i} + \sum_{i=r}^{s} \binom{n-1}{i-1} \\ &= \sum_{i=r}^{s} \binom{n-1}{i} + \sum_{i=r-1}^{s-1} \binom{n-1}{i} \qquad (1) \\ &= \sum_{i=r-1}^{s} \binom{n-1}{i} + \sum_{i=r}^{s-1} \binom{n-1}{i} \end{aligned}$$

et un calcul facile montre que la première somme additionne les $k+1$ plus grands coefficients binomiaux $\binom{n-1}{i}$ tandis que la seconde additionne les $k-1$ plus grands.

La preuve de Kleitman consiste en une récurrence sur n, le cas $n=1$ étant trivial. À la lumière de (1) nous avons seulement besoin de montrer pour les besoins de la récurrence que les combinaisons linéaires de $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ qui appartiennent à k régions disjointes peuvent être envoyées bijectivement sur des combinaisons de $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{n-1}$ qui appartiennent à $k+1$ ou $k-1$ régions.

> ***Proposition.*** *L'une des régions translatées $R_j - \boldsymbol{a}_n$ au moins est disjointe de toutes les régions translatées $R_1 + \boldsymbol{a}_n, \ldots, R_k + \boldsymbol{a}_n$.*

Pour montrer ce résultat, considérons l'hyperplan $H = \{\boldsymbol{x} : \langle \boldsymbol{a}_n, \boldsymbol{x} \rangle = c\}$ orthogonal à $\boldsymbol{a}_n$, qui contient tous les translatés $R_i + \boldsymbol{a}_n$ du côté défini par $\langle \boldsymbol{a}_n, \boldsymbol{x} \rangle \geqslant c$, et qui touche l'adhérence d'une région, appelons la $R_j + \boldsymbol{a}_n$. Un tel hyperplan existe puisque les régions sont bornées. On a $|\boldsymbol{x} - \boldsymbol{y}| < 2$ pour tout $\boldsymbol{x} \in R_j$ et $\boldsymbol{y}$ dans l'adhérence de R_j, puisque R_j est ouvert. Nous voulons montrer que $R_j - \boldsymbol{a}_n$ se trouve de l'autre côté de H. Supposons au contraire que $\langle \boldsymbol{a}_n, \boldsymbol{x} - \boldsymbol{a}_n \rangle \geqslant c$ pour un certain $\boldsymbol{x} \in R_j$, c'est-à-dire, $\langle \boldsymbol{a}_n, \boldsymbol{x} \rangle \geqslant |\boldsymbol{a}_n|^2 + c$.



Soit $\boldsymbol{y} + \boldsymbol{a}_n$ un point où H touche $R_j + \boldsymbol{a}_n$, alors $\boldsymbol{y}$ est dans l'adhérence de R_j donc $\langle \boldsymbol{a}_n, \boldsymbol{y} + \boldsymbol{a}_n \rangle = c$, c'est-à-dire, $\langle \boldsymbol{a}_n, -\boldsymbol{y} \rangle = |\boldsymbol{a}_n|^2 - c$. Ainsi :

$$\langle \boldsymbol{a}_n, \boldsymbol{x} - \boldsymbol{y} \rangle \geqslant 2|\boldsymbol{a}_n|^2$$

et l'on déduit de l'inégalité de Cauchy-Schwarz :

$$2|\boldsymbol{a}_n|^2 \leqslant \langle \boldsymbol{a}_n, \boldsymbol{x} - \boldsymbol{y} \rangle \leqslant |\boldsymbol{a}_n||\boldsymbol{x} - \boldsymbol{y}|$$

Donc (avec $|\boldsymbol{a}_n| \geqslant 1$) on trouve $2 \leqslant 2|\boldsymbol{a}_n| \leqslant |\boldsymbol{x} - \boldsymbol{y}|$, ce qui est contradictoire.

Le reste est facile. Nous classons les combinaisons $\sum \varepsilon_i \boldsymbol{a}_i$ qui appartiennent à $R_1 \cup \ldots \cup R_k$ comme suit. Dans la classe 1 nous mettons tous les $\sum_{i=1}^{n} \varepsilon_i \boldsymbol{a}_i$ tels que $\varepsilon_n = -1$ et tous les $\sum_{i=1}^{n} \varepsilon_i \boldsymbol{a}_i$ tels que $\varepsilon_n = 1$ appartenant à R_j. Dans la classe 2 nous mettons les combinaisons restantes $\sum_{i=1}^{n} \varepsilon_i \boldsymbol{a}_i$ telles que $\varepsilon_n = 1$, qui ne sont pas dans R_j. Cela implique que les combinaisons $\sum_{i=1}^{n-1} \varepsilon_i \boldsymbol{a}_i$ correspondant à la classe 1 appartiennent à $k+1$ régions disjointes $R_1 + \boldsymbol{a}_n, \ldots, R_k + \boldsymbol{a}_n$ et $R_j - \boldsymbol{a}_n$, et que les combinaisons $\sum_{i=1}^{n-1} \varepsilon_i \boldsymbol{a}_i$ correspondant à la classe 2 appartiennent aux $k-1$ regions disjointes $R_1 - \boldsymbol{a}_n, \ldots, R_k - \boldsymbol{a}_n$ sans $R_j - \boldsymbol{a}_n$. Par récurrence, la classe 1 contient au plus $\sum_{i=r-1}^{s} \binom{n-1}{i}$ combinaisons, tandis que la classe 2 contient au plus $\sum_{i=r}^{s-1} \binom{n-1}{i}$ combinaisons ; en utilisant (1) cela termine complètement la preuve, directement issue du Grand Livre. □

Bibliographie

[1] P. ERDŐS : *On a lemma of Littlewood and Offord,* Bulletin Amer. Math. Soc. **51** (1945), 898-902.

[2] G. KATONA : *On a conjecture of Erdős and a stronger form of Sperner's theorem,* Studia Sci. Math. Hungar. **1** (1966), 59-63.

[3] D. KLEITMAN : *On a lemma of Littlewood and Offord on the distribution of certain sums,* Math. Zeitschrift **90** (1965), 251-259.

[4] D. KLEITMAN : *On a lemma of Littlewood and Offord on the distributions of linear combinaisons of vectors,* Advances Math. **5** (1970), 155-157.

[5] J. E. LITTLEWOOD & A. C. OFFORD : *On the number of real roots of a random algebraic equation III,* Mat. USSR Sb. **12** (1943), 277-285.

La fonction cotangente et l'astuce de Herglotz

Chapitre 23

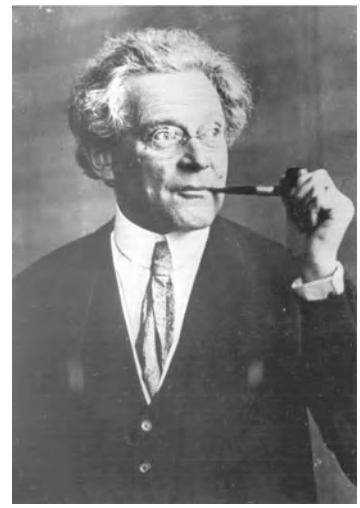
Parmi les formules qui contiennent des fonctions élémentaires, quelle est la plus intéressante ? Dans un bel article [2], dont nous suivons de près l'exposé, Jürgen Elstrodt met à la première place le développement en série de la fonction cotangente :

$$\pi \cot \pi x \;=\; \frac{1}{x} + \sum_{n=1}^{\infty} \Big(\frac{1}{x+n} + \frac{1}{x-n} \Big) \qquad (x \in \mathbb{R} \backslash \mathbb{Z})$$

Cette élégante formule a été démontrée par Euler au §178 de son *Introductio in Analysin Infinitorum*. Elle compte à coup sûr parmi ses plus beaux résultats. On peut aussi l'écrire encore plus élégamment de la manière suivante :

$$\pi \cot \pi x \;=\; \lim_{N \to \infty} \sum_{n=-N}^{N} \frac{1}{x+n} \tag{1}$$

mais il faut remarquer que l'évaluation de la somme $\sum_{n \in \mathbb{Z}} \frac{1}{x+n}$ est un peu dangereuse puisque la série n'est pas absolument convergente ; sa valeur est donc subordonnée à un choix judicieux de l'ordre de sommation.

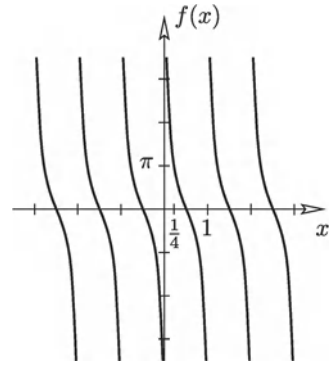


Gustav Herglotz

Nous allons déduire l'identité (1) d'un argument d'une étonnante simplicité connu sous le nom d'« astuce de Herglotz ». Pour commencer, posons :

$$f(x) \;:=\; \pi \cot \pi x \qquad g(x) \;:=\; \lim_{N \to \infty} \sum_{n=-N}^{N} \frac{1}{x+n}$$

et essayons d'établir suffisamment de propriétés communes à ces fonctions pour finalement se convaincre qu'elles doivent coïncider ...



La fonction $f(x) = \pi \cot \pi x$.

(A) Les fonctions f et g sont définies et continues en toute valeur non entière.

C'est clair pour la fonction cotangente donc pour $f(x) = \pi \cot \pi x = \pi \frac{\cos \pi x}{\sin \pi x}$ (voir figure). Pour g, nous utilisons d'abord l'identité $\frac{1}{x+n} + \frac{1}{x-n} = -\frac{2x}{n^2-x^2}$ pour écrire la formule d'Euler de la manière suivante :

$$\pi \cot \pi x \;=\; \frac{1}{x} - \sum_{n=1}^{\infty} \frac{2x}{n^2 - x^2}\,. \tag{2}$$

Ainsi, pour établir **(A)** nous devons montrer que pour chaque $x \notin \mathbb{Z}$ la série :

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

converge uniformément dans un voisinage de x.
Le terme d'indice $n = 1$ ainsi que les termes d'indice n tel que $2n - 1 \leqslant x^2$ ne posent pas problème puisqu'ils sont en nombre fini. D'autre part, si $n \geqslant 2$ et $2n - 1 > x^2$, c'est-à-dire si $n^2 - x^2 > (n-1)^2 > 0$, les termes sont majorés par :

$$0 \;<\; \frac{1}{n^2 - x^2} \;<\; \frac{1}{(n-1)^2}$$

et cette borne est non seulement valable pour la valeur x elle-même, mais aussi pour des valeurs qui se trouvent dans un voisinages de x. Enfin, le fait que $\sum \frac{1}{(n-1)^2}$ converge (vers $\frac{\pi^2}{6}$, voir page 49) assure la convergence uniforme requise.

(B) Les fonctions f et g sont toutes les deux *périodiques* de période 1, c'est-à-dire que l'on a $f(x+1) = f(x)$ et $g(x+1) = g(x)$ pour tout $x \in \mathbb{R}\backslash\mathbb{Z}$.

Comme la fonction cotangente a π pour période, f a 1 pour période (voir figure). Pour g nous raisonnons comme suit. Soit :

$$g_N(x) \;:=\; \sum_{n=-N}^{N} \frac{1}{x+n}$$

alors :

$$\begin{aligned} g_N(x+1) \;&=\; \sum_{n=-N}^{N} \frac{1}{x+1+n} \;=\; \sum_{n=-N+1}^{N+1} \frac{1}{x+n} \\ &=\; g_{N-1}(x) + \frac{1}{x+N} + \frac{1}{x+N+1} \end{aligned}$$

Ainsi, $g(x+1) = \lim_{N\to\infty} g_N(x+1) = \lim_{N\to\infty} g_{N-1}(x) = g(x)$.

(C) Les fonctions f et g sont toutes les deux *impaires*, c'est-à-dire que $f(-x) = -f(x)$ et $g(-x) = -g(x)$ pour tout $x \in \mathbb{R}\backslash\mathbb{Z}$.

La fonction f vérifie évidemment cette propriété, quant à g il suffit d'observer que : $g_N(-x) = -g_N(x)$.

Les deux derniers résultats constituent l'astuce de Herglotz : d'abord, on montre que f et g vérifient la même équation fonctionnelle, ensuite que $h := f - g$ peut être prolongée continûment à tout $\mathbb{R}$.

Formules d'addition :
$\sin(x+y) = \sin x \cos y + \cos x \sin y$
$\cos(x+y) = \cos x \cos y - \sin x \sin y$
$\Longrightarrow$ $\sin(x + \frac{\pi}{2}) = \cos x$
$\cos(x + \frac{\pi}{2}) = -\sin x$
$\sin x = 2 \sin \frac{x}{2} \cos \frac{x}{2}$
$\cos x = \cos^2 \frac{x}{2} - \sin^2 \frac{x}{2}$.

(D) Les deux fonctions f et g vérifient la même équation fonctionnelle : $f(\frac{x}{2}) + f(\frac{x+1}{2}) = 2f(x)$ et $g(\frac{x}{2}) + g(\frac{x+1}{2}) = 2g(x)$.

Pour f cela résulte des formules d'addition des fonctions sinus et cosinus :

$$\begin{aligned} f(\tfrac{x}{2}) + f(\tfrac{x+1}{2}) &= \pi\left[\frac{\cos\frac{\pi x}{2}}{\sin\frac{\pi x}{2}} - \frac{\sin\frac{\pi x}{2}}{\cos\frac{\pi x}{2}}\right] \\ &= 2\pi\frac{\cos(\frac{\pi x}{2} + \frac{\pi x}{2})}{\sin(\frac{\pi x}{2} + \frac{\pi x}{2})} = 2\,f(x) \end{aligned}$$

L'équation fonctionnelle satisfaite par g vient de l'égalité :

$$g_N(\tfrac{x}{2}) + g_N(\tfrac{x+1}{2}) = 2\,g_{2N}(x) + \frac{2}{x+2N+1}$$

qui, à son tour, résulte de l'égalité :

$$\frac{1}{\frac{x}{2}+n} + \frac{1}{\frac{x+1}{2}+n} = 2\Big(\frac{1}{x+2n} + \frac{1}{x+2n+1}\Big)$$

Examinons à présent :

$$h(x) = f(x) - g(x) = \pi\cot\pi x - \Big(\frac{1}{x} - \sum_{n=1}^{\infty}\frac{2x}{n^2-x^2}\Big) \qquad (3)$$

Nous savons que h est une fonction continue sur $\mathbb{R}\backslash\mathbb{Z}$ qui vérifie les propriétés **(B)**, **(C)**, **(D)**. Que se passe-t-il pour les valeurs entières ? En utilisant les développements en série du sinus et du cosinus ou en appliquant deux fois la règle de l'Hospital, on voit que :

$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} \pm \dots$

$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} \pm \dots$

$$\lim_{x\to 0}\Big(\cot x - \frac{1}{x}\Big) = \lim_{x\to 0}\frac{x\cos x - \sin x}{x\sin x} = 0$$

donc que :

$$\lim_{x\to 0}\Big(\pi\cot\pi x - \frac{1}{x}\Big) = 0$$

Puisque la dernière somme $\sum_{n=1}^{\infty}\frac{2x}{n^2-x^2}$ dans (3) converge vers 0 lorsque $x \longrightarrow 0$, on a en fait $\lim\limits_{x\to 0} h(x) = 0$ donc en utilisant la périodicité

$$\lim_{x\to n} h(x) = 0 \qquad \text{pour tout } n \in \mathbb{Z}$$

En résumé, on vient de montrer que :

(E) En posant $h(x) := 0$ si $x \in \mathbb{Z}$, h est une fonction continue sur $\mathbb{R}$ (tout entier) et vérifie les propriétés énoncées en **(B)**, **(C)** et **(D)**.

Nous sommes prêts pour le *coup de grâce* [1]. Comme h est une fonction périodique continue, elle admet un maximum m. Soit x_0 un point de $[0,1]$ tel que $h(x_0) = m$. On déduit de **(D)** que :

$$h(\tfrac{x_0}{2}) + h(\tfrac{x_0+1}{2}) \;=\; 2m$$

et donc que : $h(\frac{x_0}{2}) = m$. Par récurrence on obtient $h(\frac{x_0}{2^n}) = m$ pour tout n, ainsi $h(0) = m$ par continuité. Or $h(0) = 0$ donc $m = 0$, c'est-à-dire que $h(x) \leqslant 0$ pour tout $x \in \mathbb{R}$. Comme h est une fonction *impaire*, $h(x) < 0$ est impossible donc $h(x) = 0$ pour tout $x \in \mathbb{R}$. Le théorème d'Euler est démontré. □

On peut déduire de la relation (1) un grand nombre de corollaires. Le plus célèbre concerne les valeurs de la fonction Zêta de Riemann aux points entiers pairs (se reporter à l'appendice du chapitre 8) :

$$\zeta(2k) \;=\; \sum_{n=1}^{\infty} \frac{1}{n^{2k}} \qquad (k \in \mathbb{N}^*) \tag{4}$$

Pour terminer notre histoire, voyons comment Euler quelques années plus tard, en 1755, a traité la série (4). Considérons d'abord la formule (2). En multipliant (2) par x et en posant $y = \pi x$ nous trouvons, si $|y| < \pi$:

$$\begin{aligned} y\cot y &= 1 - 2\sum_{n=1}^{\infty} \frac{y^2}{\pi^2 n^2 - y^2} \\ &= 1 - 2\sum_{n=1}^{\infty} \frac{y^2}{\pi^2 n^2} \, \frac{1}{1 - \left(\frac{y}{\pi n}\right)^2} \end{aligned}$$

Le dernier facteur est la somme d'une série géométrique, donc :

$$\begin{aligned} y\cot y &= 1 - 2\sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \left(\frac{y}{\pi n}\right)^{2k} \\ &= 1 - 2\sum_{k=1}^{\infty} \Big(\frac{1}{\pi^{2k}} \sum_{n=1}^{\infty} \frac{1}{n^{2k}}\Big) y^{2k} \end{aligned}$$

Nous avons ainsi établi le résultat remarquable suivant :

Pour tout $k \in \mathbb{N}^*$, le coefficient de y^{2k} dans le développement en série entière de $y\cot y$ est égal à :

$$\big[y^{2k}\big]_{y\cot y} \;=\; -\frac{2}{\pi^{2k}} \sum_{n=1}^{\infty} \frac{1}{n^{2k}} \;=\; -\frac{2}{\pi^{2k}}\zeta(2k) \tag{5}$$

1. N.d.T. : en français dans le texte original.

Il y a un autre chemin, peut-être bien plus « canonique », pour obtenir un développement en série entière de $y \cot y$. Nous savons par l'analyse que $e^{iy} = \cos y + i \sin y$ donc que :

$$\cos y = \frac{e^{iy} + e^{-iy}}{2} \qquad \sin y = \frac{e^{iy} - e^{-iy}}{2i}$$

ce qui implique :

$$y \cot y = iy \frac{e^{iy} + e^{-iy}}{e^{iy} - e^{-iy}} = iy \frac{e^{2iy} + 1}{e^{2iy} - 1}$$

Posant $z = 2iy$, on obtient :

$$y \cot y = \frac{z}{2} \frac{e^z + 1}{e^z - 1} = \frac{z}{2} + \frac{z}{e^z - 1} \tag{6}$$

Ainsi, nous avons seulement besoin du développement en série entière de la fonction $\frac{z}{e^z - 1}$. Remarquons que cette fonction est définie et continue sur tout $\mathbb{R}$ (pour $z = 0$ on utilise le développement en série entière de la fonction exponentielle, ou encore la règle de l'Hospital qui donne la valeur 1). Nous écrivons :

$$\frac{z}{e^z - 1} =: \sum_{n \geqslant 0} B_n \frac{z^n}{n!} \tag{7}$$

Les coefficients B_n sont connus sous le nom de *nombres de Bernoulli*. Le membre gauche de (6) est une fonction *paire* (c'est-à-dire que $f(z) = f(-z)$), donc $B_n = 0$ pour $n \geqslant 3$ impair ; $B_1 = -\frac{1}{2}$ correspond donc au terme en z de (6).

À partir de :

$$\Big(\sum_{n \geqslant 0} B_n \frac{z^n}{n!}\Big)(e^z - 1) = \Big(\sum_{n \geqslant 0} B_n \frac{z^n}{n!}\Big)\Big(\sum_{n \geqslant 1} \frac{z^n}{n!}\Big) = z$$

nous obtenons, en comparant les coefficients de z^n :

$$\sum_{k=0}^{n-1} \frac{B_k}{k!(n-k)!} = \begin{cases} 1 & \text{si } n = 1 \\ 0 & \text{si } n \neq 1 \end{cases} \tag{8}$$

On peut calculer les nombres de Bernoulli par récurrence à partir de (8). La valeur $n = 1$ donne $B_0 = 1$, $n = 2$ conduit à $\frac{B_0}{2} + B_1 = 0$, c'est-à-dire $B_1 = -\frac{1}{2}$ et ainsi de suite.

n	0	1	2	3	4	5	6	7	8
B_n	1	$-\frac{1}{2}$	$\frac{1}{6}$	0	$-\frac{1}{30}$	0	$\frac{1}{42}$	0	$-\frac{1}{30}$

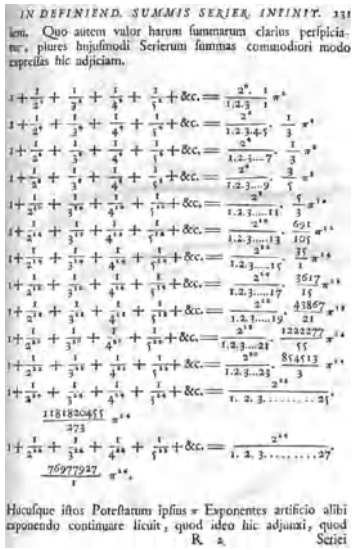
Les premiers nombres de Bernoulli.

On a presque terminé. La combinaison de (6) et (7) implique :

$$y \cot y = \sum_{k=0}^{\infty} B_{2k} \frac{(2iy)^{2k}}{(2k)!} = \sum_{k=0}^{\infty} \frac{(-1)^k 2^{2k} B_{2k}}{(2k)!} y^{2k}$$

et permet d'obtenir, avec (5), la formule d'Euler de $\zeta(2k)$:

$$\sum_{n=1}^{\infty} \frac{1}{n^{2k}} = \frac{(-1)^{k-1} 2^{2k-1} B_{2k}}{(2k)!} \pi^{2k} \qquad (k \in \mathbb{N}^*) \tag{9}$$



IN DEFINIEND. SUMMIS SERIER. INFINIT. 131

lem. Quo autem valor harum summarum clarius perſpicia-tur, plures hujuſmodi Serierum ſummas commodiori modo expreſſas hic adjiciam.

$$1+\frac{1}{2^n}+\frac{1}{3^n}+\frac{1}{4^n}+\frac{1}{5^n}+\&c.=\frac{2^0}{1.2.3}\cdot\frac{1}{1}\pi^0$$

$$1+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2}+\frac{1}{5^2}+\&c.=\frac{2^1}{1.2.3.4.5}\cdot\frac{1}{3}\pi^2$$

$$1+\frac{1}{2^4}+\frac{1}{3^4}+\frac{1}{4^4}+\frac{1}{5^4}+\&c.=\frac{2^3}{1.2.3....7}\cdot\frac{1}{3}\pi^4$$

$$1+\frac{1}{2^6}+\frac{1}{3^6}+\frac{1}{4^6}+\frac{1}{5^6}+\&c.=\frac{2^6}{1.2.3....9}\cdot\frac{3}{5}\pi^6$$

$$1+\frac{1}{2^8}+\frac{1}{3^8}+\frac{1}{4^8}+\frac{1}{5^8}+\&c.=\frac{2^7}{1.2.3....11}\cdot\frac{5}{3}\pi^8$$

$$1+\frac{1}{2^{10}}+\frac{1}{3^{10}}+\frac{1}{4^{10}}+\frac{1}{5^{10}}+\&c.=\frac{2^{10}}{1.2.3....13}\cdot\frac{691}{105}\pi^{10}$$

$$1+\frac{1}{2^{12}}+\frac{1}{3^{12}}+\frac{1}{4^{12}}+\frac{1}{5^{12}}+\&c.=\frac{2^{12}}{1.2.3....15}\cdot\frac{35}{1}\pi^{12}$$

$$1+\frac{1}{2^{14}}+\frac{1}{3^{14}}+\frac{1}{4^{14}}+\frac{1}{5^{14}}+\&c.=\frac{2^{14}}{1.2.3....17}\cdot\frac{3617}{15}\pi^{14}$$

$$1+\frac{1}{2^{16}}+\frac{1}{3^{16}}+\frac{1}{4^{16}}+\frac{1}{5^{16}}+\&c.=\frac{2^{16}}{1.2.3....19}\cdot\frac{43867}{21}\pi^{16}$$

$$1+\frac{1}{2^{18}}+\frac{1}{3^{18}}+\frac{1}{4^{18}}+\frac{1}{5^{18}}+\&c.=\frac{2^{18}}{1.2.3...21}\cdot\frac{1222277}{55}\pi^{18}$$

$$1+\frac{1}{2^{20}}+\frac{1}{3^{20}}+\frac{1}{4^{20}}+\frac{1}{5^{20}}+\&c.=\frac{2^{20}}{1.2.3...23}\cdot\frac{854513}{3}\pi^{20}$$

$$1+\frac{1}{2^{22}}+\frac{1}{3^{22}}+\frac{1}{4^{22}}+\frac{1}{5^{22}}+\&c.=\frac{2^{21}}{1.\ 2.\ 3.\ \ldots\ldots\ldots 25}\cdot\frac{1181820455}{273}\pi^{22}$$

$$1+\frac{1}{2^{24}}+\frac{1}{3^{24}}+\frac{1}{4^{24}}+\frac{1}{5^{24}}+\&c.=\frac{2^{24}}{1.\ 2.\ 3.\ \ldots\ldots\ldots 27}\cdot\frac{76977927}{1}\pi^{24}.$$

Hucuſque iſtos Poteſtatum ipſius π Exponentes artificio alibi exponendo continuare licuit, quod ideo hic adjunxi, quod

R 2 Serici

Page 131 de l'*Introductio in Analysin Infinitorum* publiée par Euler en 1748.

En examinant la table des nombres de Bernoulli, on retrouve la valeur de la somme $\sum \frac{1}{n^2} = \frac{\pi^2}{6}$ établie au chapitre 8. En outre :

$$\sum_{n=1}^{\infty}\frac{1}{n^4}=\frac{\pi^4}{90},\qquad \sum_{n=1}^{\infty}\frac{1}{n^6}=\frac{\pi^6}{945},\qquad \sum_{n=1}^{\infty}\frac{1}{n^8}=\frac{\pi^8}{9450},$$

$$\sum_{n=1}^{\infty}\frac{1}{n^{10}}=\frac{\pi^{10}}{93555},\qquad \sum_{n=1}^{\infty}\frac{1}{n^{12}}=\frac{691\,\pi^{12}}{638512875},\qquad \cdots$$

Le nombre de Bernoulli $B_{10} = \frac{5}{66}$ qui conduit à $\zeta(10)$ semble assez inoffensif mais la valeur suivante, $B_{12} = -\frac{691}{2730}$, utile pour le calcul de $\zeta(12)$, contient le grand facteur premier 691 au numérateur. Euler avait d'abord calculé quelques valeurs $\zeta(2k)$ sans avoir remarqué le lien avec les nombres de Bernoulli. Seule l'apparition de l'étrange nombre premier 691 l'a mis sur la bonne piste.

Incidemment, puisque $\zeta(2k)$ converge vers 1 lorsque $k \longrightarrow \infty$, l'équation (9) nous indique que les nombres $|B_{2k}|$ croissent très vite, ce qui n'est pas clair à partir des premières valeurs.

En revanche, on connaît très peu de choses sur les valeurs de la fonction Zêta de Riemann aux valeurs entières impaires $k \geqslant 3$ (voir page 57).

Bibliographie

[1] S. BOCHNER : *Book review of "Gesammelte Schriften" by Gustav Herglotz,* Bulletin Amer. Math. Soc. **1** (1979), 1020-1022.

[2] J. ELSTRODT : *Partialbruchzerlegung des Kotangens, Herglotz-Trick und die Weierstraßsche stetige, nirgends differenzierbare Funktion,* Math. Semesterberichte **45** (1998), 207-220.

[3] L. EULER : *Introductio in Analysin Infinitorum,* Tomus Primus, Lausanne 1748 ; Opera Omnia, Ser. 1, Vol. 8. In English : *Introduction to Analysis of the Infinite,* Book I (translated by J. D. Blanton), Springer-Verlag, New York 1988.

[4] L. EULER : *Institutiones calculi differentialis cum ejus usu in analysi finitorum ac doctrina serierum,* Petersburg 1755 ; Opera Omnia, Ser. 1, Vol. 10.

Le problème de l'aiguille de Buffon

Chapitre 24

En 1777, un noble français, Georges Louis Leclerc, Comte de Buffon, posa le problème suivant :

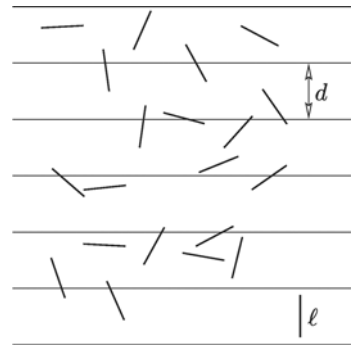
> *Supposons que l'on laisse tomber une petite aiguille sur une feuille de papier réglée. Quelle est la probabilité que l'aiguille tombe dans une position telle qu'elle traverse l'une des droites ?*

Timbre français à l'effigie du Comte de Buffon.

La probabilité dépend de la distance d entre les droites de la feuille de papier réglée et de la longueur ℓ de l'aiguille que l'on laisse tomber. Elle dépend en fait du quotient $\frac{\ell}{d}$. Nous dirons qu'une aiguille est *petite* si elle est de longueur $\ell \leqslant d$. En d'autres termes, une petite aiguille ne peut pas couper deux droites à la fois : elle parvient à toucher deux droites à la fois avec une probabilité nulle. La réponse au problème de Buffon peut paraître surprenante : elle fait intervenir le nombre π.

Théorème (« Le problème de l'aiguille de Buffon »)
Si une petite aiguille de longueur ℓ est envoyée sur du papier réglé dont les droites sont régulièrement espacées d'une distance $d \geqslant \ell$, alors la probabilité que l'aiguille se place dans une position telle qu'elle coupe l'une des droites est exactement :

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



Ce résultat signifie que l'on peut obtenir des valeurs approchées de π par l'expérience : si on laisse tomber une aiguille N fois et si on obtient une réponse positive (une intersection) dans P cas, alors $\frac{P}{N}$ doit être approximativement égal à $\frac{2}{\pi}\frac{\ell}{d}$, c'est-à-dire que l'on peut approcher π par $\frac{2\ell N}{dP}$. Le test le plus extensif (et exhaustif) a dû être fait par Lazzarini en 1901. Il prétendit même avoir construit une machine dans le but de faire tomber un bâton 3408 fois (avec $\frac{\ell}{d} = \frac{5}{6}$). Il trouva que le bâton coupa une des droites 1808 fois, ce qui conduit à l'approximation $\pi \approx 2 \cdot \frac{5}{6}\frac{3408}{1808} = 3,1415929...$, résultat correct jusqu'au sixième chiffre de π et beaucoup trop beau pour être vrai ! Les valeurs que Lazzarini a choisies conduisent directement à l'approximation bien connue $\pi \approx \frac{355}{113}$ (voir page 46). Cela explique le choix plus que suspect de 3408 et $\frac{5}{6}$, nombres tels que $\frac{5}{6}$ 3408 soit un multiple de 355 (voir [5] pour une évocation de la mystification de Lazzarini).

On peut résoudre le problème de l'aiguille en évaluant une intégrale. Nous allons le faire plus loin, et avec cette méthode nous allons aussi résoudre le

problème pour une aiguille longue. Cependant la Preuve méritant de figurer dans le Grand Livre, présentée par E. Barbier en 1860, n'a pas besoin de recourir à une intégrale. Elle consiste simplement à laisser tomber une aiguille différente...

Si on laisse tomber une aiguille *quelconque*, petite ou grande, alors le nombre attendu d'intersections est :

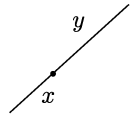
$$E \;=\; p_1 + 2p_2 + 3p_3 + \dots$$

où p_1 est la probabilité que l'aiguille se positionne avec une unique intersection, p_2 est la probabilité d'obtenir deux intersections exactement, p_3 est la probabilité de trois intersections etc. La probabilité d'obtenir au moins une intersection, ce qui est demandé dans le problème de Buffon, est donc :

$$p \;=\; p_1 + p_2 + p_3 + \dots$$

Les événements tels que l'aiguille se positionne exactement sur une droite, ou à touche par son extrémité une des droites, ont une probabilité nulle. Ils peuvent donc être ignorés dans notre propos.

D'autre part, si l'aiguille est *petite* alors la probabilité d'obtenir plus d'une intersection est nulle, $p_2 = p_3 = \dots = 0$ donc $E = p$: la probabilité que nous cherchons est simplement le nombre d'intersections attendu. Cette reformulation est extrêmement utile, parce que nous pouvons maintenant utiliser la linéarité de l'espérance (voir page 107). En effet, notons $E(\ell)$ l'espérance du nombre d'intersections obtenues en laissant tomber une aiguille droite de longueur ℓ. Si cette longueur est $\ell = x+y$ et si on considère la « partie avant » de longueur x et la « partie arrière » de longueur y de l'aiguille séparément, alors on obtient :



$$E(x+y) \;=\; E(x) + E(y)$$

puisque le nombre total d'intersections est égal au nombre de celles produites par la partie avant et par la partie arrière.

Par récurrence sur n, cette « équation fonctionnelle » implique que $E(nx) = nE(x)$ pour tout $n \in \mathbb{N}$ donc que $mE(\frac{n}{m}x) = E(m\frac{n}{m}x) = E(nx) = nE(x)$. Ainsi, nous avons $E(rx) = rE(x)$ pour tout *rationnel* $r \in \mathbb{Q}$. En outre, il est clair que $E(x)$ est monotone en $x \geqslant 0$, d'où l'on déduit que $E(x) = cx$ pour tout $x \geqslant 0$, où $c = E(1)$ est une constante.

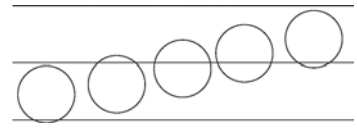
Quelle est cette constante ?

Pour répondre à cette question, nous utilisons des aiguilles de différentes formes. En effet, laissons tomber une aiguille « polygonale » de longueur ℓ, qui est constituée de morceaux de droites. Alors, le nombre d'intersections obtenu est, avec probabilité 1, la somme des nombres d'intersections produites par ces morceaux de droites. Ainsi, l'espérance du nombre d'intersections est à nouveau :

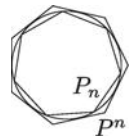
$$E \;=\; c\ell$$

par linéarité de l'espérance (il importe peu que les morceaux de droites soient reliés entre eux de manière rigide ou flexible !).

La clé de la solution de Barbier au problème de l'aiguille de Buffon est de considérer une aiguille qui soit un cercle parfait C de diamètre d, donc de longueur $x = d\pi$. Si on laisse tomber une telle aiguille sur une feuille de papier réglée, elle produit toujours exactement deux intersections !

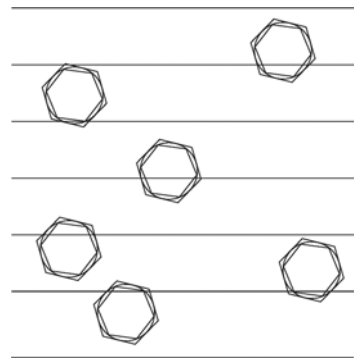


On peut approcher le cercle par des polygones. Imaginons simplement qu'avec l'aiguille circulaire C, nous laissions tomber un polygone inscrit P_n et un polygone circonscrit P^n. Toute droite qui coupe P_n coupe aussi C. Par ailleurs, si une droite coupe C, alors elle touche aussi P^n. Ainsi, l'espérance du nombre d'intersections satisfait :



$$E(P_n) \leqslant E(C) \leqslant E(P^n)$$

Toutefois, P_n et P^n sont tous les deux des polygones, donc le nombre d'intersections espéré est égal au produit de c par la longueur pour chacun d'eux, alors que ce nombre vaut 2 pour C, ainsi :



$$c\,\ell(P_n) \leqslant 2 \leqslant c\,\ell(P^n) \tag{1}$$

Comme P_n et P^n approchent tous les deux C lorsque $n \longrightarrow \infty$, on a en particulier :

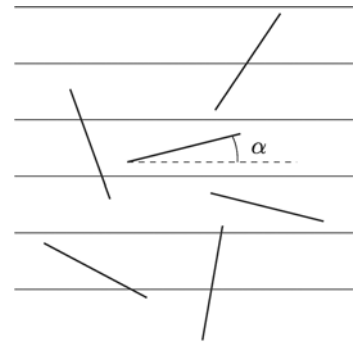
$$\lim_{n\to\infty} \ell(P_n) = d\pi = \lim_{n\to\infty} \ell(P^n)$$

et donc lorsque $n \longrightarrow \infty$ on déduit de (1) que :

$$c\,d\pi \leqslant 2 \leqslant c\,d\pi$$

ce qui conduit à $c = \frac{2}{\pi}\frac{1}{d}$. □

Il était possible d'obtenir ce résultat par le calcul ! L'astuce pour se ramener à une intégrale « facile à calculer » est de considérer d'abord la pente de l'aiguille. Disons qu'elle tombe avec un angle α par rapport à l'horizontale, où α vérifie $0 \leqslant \alpha \leqslant \frac{\pi}{2}$. On ignore le cas où l'aiguille se positionne avec une pente négative, puisque le cas est symétrique à celui de la pente positive et conduit à la même probabilité. Une aiguille qui se positionne avec un angle α a une hauteur $\ell \sin\alpha$, et la probabilité pour qu'une telle aiguille traverse une des lignes horizontales espacées d'une distance d est $\frac{\ell \sin\alpha}{d}$. Ainsi, on obtient la probabilité en calculant la moyenne sur tous les angles α possibles :



$$p = \frac{2}{\pi}\int_0^{\frac{\pi}{2}} \frac{\ell\sin\alpha}{d}d\alpha = \frac{2}{\pi}\frac{\ell}{d}\big[-\cos\alpha\big]_0^{\frac{\pi}{2}} = \frac{2}{\pi}\frac{\ell}{d}$$

Pour une aiguille longue, on obtient la même probabilité $\frac{\ell\sin\alpha}{d}$ tant que $\ell\sin\alpha \leqslant d$, c'est-à-dire dans l'intervalle $0 \leqslant \alpha \leqslant \arcsin\frac{d}{\ell}$. Cependant, pour des angles α plus grands l'aiguille *doit* traverser une ligne donc la probabilité est 1. Ainsi :

$$p = \frac{2}{\pi}\Big(\int_0^{\arcsin(\frac{d}{\ell})} \frac{\ell\sin\alpha}{d} + \int_{\arcsin(\frac{d}{\ell})}^{\frac{\pi}{2}} 1\,dx\Big)$$

$$= \frac{2}{\pi}\Big(\frac{\ell}{d}\Big[-\cos\alpha\Big]_0^{\arcsin(\frac{d}{\ell})} + \Big(\frac{\pi}{2} - \arcsin\frac{d}{\ell}\Big)\Big)$$
$$= 1 + \frac{2}{\pi}\Big(\frac{\ell}{d}\Big(1 - \sqrt{1 - \frac{d^2}{\ell^2}}\Big) - \arcsin\frac{d}{l}\Big)$$

si $\ell \geqslant d$.

La réponse n'est donc pas aussi belle pour une aiguille plus longue mais elle fournit un bel exercice : montrer (simplement par sécurité) que la formule donne $\frac{2}{\pi}$ lorsque $\ell = d$, que p est strictement croissante en fonction de ℓ et tend vers 1 lorsque $\ell \longrightarrow \infty$.

Bibliographie

[1] E. BARBIER : *Note sur le problème de l'aiguille et le jeu du joint couvert,* J. Mathématiques Pures et Appliquées (2) **5** (1860), 273-286.

[2] L. BERGGREN, J. BORWEIN & P. BORWEIN, EDS. : *Pi : A Source Book,* Springer-Verlag, New York 1997.

[3] G. L. LECLERC, COMTE DE BUFFON : *Essai d'arithmétique morale,* Appendix to "Histoire naturelle générale et particulière," Vol. 4, 1777.

[4] D. A. KLAIN & G.-C. ROTA : *Introduction to Geometric Probability,* "Lezioni Lincee", Cambridge University Press 1997.

[5] T. H. O'BEIRNE : *Puzzles and Paradoxes,* Oxford University Press, London 1965.

« Un problème ? »

Combinatoire

« Un carré latin mélancolique ».

Le principe des tiroirs et le double décompte

Chapitre 25

Certains principes mathématiques, comme ceux qui figurent dans le titre de ce chapitre, sont si simples que l'on peut penser qu'ils ne produisent que des résultats évidents. Pour se convaincre que « ce n'est pas nécessairement le cas », nous les illustrons par des exemples que Paul Erdős suggérait de présenter dans le Grand Livre. Nous retrouverons aussi certains d'entre eux dans les chapitres suivants.

Le principe des tiroirs.
Si n objets sont placés dans r boîtes, avec $r < n$, l'une des boîtes au moins contient plus d'un objet.

« Le principe des tiroirs... »

Ce résultat est évident et il n'y à rien à prouver. Dans le langage des applications, ce principe s'énonce de la manière suivante : soit N et R deux ensembles finis tels que $|N| = n > r = |R|$, et $f : N \longrightarrow R$ une application. Alors il existe $a \in R$ tel que $|f^{-1}(a)| \geqslant 2$. On peut même énoncer une inégalité plus forte ; il existe $a \in R$ tel que :

$$|f^{-1}(a)| \geqslant \left\lceil \frac{n}{r} \right\rceil. \tag{1}$$

En effet, sinon on aurait $|f^{-1}(a)| < \frac{n}{r}$ pour tout a et, par conséquent, $n = \sum_{a \in R} |f^{-1}(a)| < r\,\frac{n}{r} = n$, ce qui est impossible.

1. Nombres

Proposition. *Considérons les nombres $1, 2, 3, \ldots, 2n$ et choisissons en $n + 1$. Parmi ces $n + 1$ nombres, il en existe deux qui sont premiers entre eux.*

C'est encore un résultat évident : il doit y avoir deux nombres qui ne diffèrent que d'une unité et qui sont donc premiers entre eux.
Prenons le problème autrement.

Proposition. *Soit $A \subseteq \{1, 2, \ldots, 2n\}$ avec $|A| = n + 1$. Alors il existe deux nombres dans A tels que l'un divise l'autre.*

Ce n'est pas si clair. Comme nous l'avait confié Erdős, il avait posé cette question au jeune Lajos Pósa lors d'un dîner ; lorsque le repas fut terminé,

Ces deux résultats ne sont plus vrais si l'on remplace $n+1$ par n : il suffit de considérer respectivement les ensembles $\{2,4,6,\ldots,2n\}$ et $\{n+1,n+2,\ldots,2n\}$.

Lajos avait la réponse. C'est resté l'une des questions d'« initiation » aux mathématiques préférées d'Erdős. La solution est fournie par le principe des tiroirs. On écrit chaque nombre $a \in A$ sous la forme $a = 2^k m$, où m est un nombre impair compris entre 1 et $2n-1$. Puisqu'il y a $n+1$ nombres dans A, mais seulement n composantes impaires différentes, il doit y avoir deux nombres dans A de *même* composante impaire. Par conséquent l'un est multiple de l'autre. □

2. Suites

Voici un autre résultat parmi les préférés d'Erdős, qui figure dans un article d'Erdős et Szekeres sur les problèmes de Ramsey.

> ***Proposition.*** *Dans toute suite $a_1, a_2, \ldots, a_{mn+1}$ de $mn+1$ nombres réels distincts, il existe une sous-suite croissante :*
>
> $$a_{i_1} < a_{i_2} < \ldots < a_{i_{m+1}} \qquad (i_1 < i_2 < \ldots < i_{m+1})$$
>
> *de longueur $m+1$, ou une sous-suite décroissante :*
>
> $$a_{j_1} > a_{j_2} > \ldots > a_{j_{n+1}} \qquad (j_1 < j_2 < \ldots < j_{n+1})$$
>
> *de longueur $n+1$, ou les deux à la fois.*

Cette fois, l'application du principe des tiroirs n'est pas immédiate. Associons à chaque a_i le nombre t_i qui est la longueur de la *plus longue sous-suite croissante* qui commence par a_i. Si $t_i \geqslant m+1$ pour un certain i, nous obtenons une sous-suite croissante de longueur $m+1$. Supposons alors que $t_i \leqslant m$ pour tout i. La fonction $f : a_i \longmapsto t_i$ qui envoie $\{a_1, \ldots, a_{mn+1}\}$ dans $\{1, \ldots, m\}$ vérifie d'après (1) la propriété suivante : il existe un certain $s \in \{1, \ldots, m\}$ tel que $f(a_i) = s$ pour $\frac{mn}{m} + 1 = n+1$ nombres a_i. Soient $a_{j_1}, a_{j_2}, \ldots, a_{j_{n+1}}$ $(j_1 < \ldots < j_{n+1})$ ces nombres. Examinons deux nombres consécutifs a_{j_i} et $a_{j_{i+1}}$. Si $a_{j_i} < a_{j_{i+1}}$, on obtiendrait une sous-suite croissante de longueur s qui commence par $a_{j_{i+1}}$ et, par conséquent, une sous-suite croissante de longueur $s+1$ qui commence par a_{j_i}, ce qui est impossible puisque $f(a_{j_i}) = s$. On obtient donc une sous-suite décroissante $a_{j_1} > a_{j_2} > \ldots > a_{j_{n+1}}$ de longueur $n+1$. □

Le lecteur peut s'amuser à montrer que pour mn nombres, l'affirmation n'est plus vraie en général.

Ce résultat *a priori* simple sur les sous-suites monotones a une conséquence importante très peu évidente sur la *dimension d'un graphe*. Nous n'avons pas besoin ici de la notion de dimension pour un graphe en général, mais seulement de savoir ce qu'est la dimension d'un graphe complet K_n. Cette notion peut être formulée de la manière suivante. Soit $N = \{1, \ldots, n\}$, $n \geqslant 3$. Considérons m permutations $\pi_1, \ldots, \pi_m$ de N. On dit que les permutations π_i *représentent* K_n si pour tout triplet de nombres distincts i, j, k il existe une permutation π pour laquelle k se trouve *après* i et j. La dimension de K_n est alors le plus petit m pour lequel existe une représentation $\pi_1, \ldots, \pi_m$.

À titre d'exemple, $\dim(K_3) = 3$ puisque chacun des trois nombres doit se trouver à la fin, comme dans $\pi_1 = (1,2,3)$, $\pi_2 = (2,3,1)$, $\pi_3 = (3,1,2)$. Qu'en est-il de K_4 ? Notons d'abord que $\dim(K_n) \leqslant \dim(K_{n+1})$: il suffit de supprimer $n+1$ dans une représentation de K_{n+1}. Par suite, $\dim(K_4) \geqslant 3$; en fait, $\dim(K_4) = 3$, puisqu'il suffit de prendre :

$$\pi_1 = (1,2,3,4), \quad \pi_2 = (2,4,3,1), \quad \pi_3 = (1,4,3,2).$$

Il n'est pas si facile de prouver que $\dim(K_5) = 4$, mais ensuite, de manière surprenante, la dimension reste 4 jusqu'à $n = 12$, puis $\dim(K_{13}) = 5$. Ainsi, $\dim(K_n)$ semble être une fonction plutôt sauvage. En fait, il n'en est rien ! Lorsque n tend vers l'infini, $\dim(K_n)$ est en fait une fonction au comportement régulier. La recherche d'une borne inférieure repose sur le principe des tiroirs. On peut affirmer :

$$\dim(K_n) \geqslant \log_2\log_2 n. \tag{2}$$

Puisque, comme nous l'avons vu, $\dim(K_n)$ est une fonction monotone en n, il suffit de vérifier (2) pour $n = 2^{2^p} + 1$, c'est-à-dire de montrer que :

$$\dim(K_n) \geqslant p+1 \qquad \text{pour} \quad n = 2^{2^p} + 1.$$

Supposons, par l'absurde, que $\dim(K_n) \leqslant p$ et considérons des permutations $\pi_1, \ldots, \pi_p$ représentant $N = \{1, 2, \ldots, 2^{2^p} + 1\}$. Utilisons p fois le résultat sur les sous-suites monotones. Dans π_1 il existe une sous-suite monotone (croissante ou décroissante, cela n'a aucune importance) A_1 de longueur $2^{2^{p-1}} + 1$. Considérons cet ensemble A_1 dans π_2. En utilisant encore le résultat, on obtient une sous-suite monotone A_2 de A_1 dans π_2 de longueur $2^{2^{p-2}} + 1$; A_2 est, bien sûr, également monotone dans π_1. En continuant, on trouve finalement une sous-suite A_p de taille $2^{2^0} + 1 = 3$ monotone dans *toutes* les permutations π_i. Soit $A_p = (a,b,c)$. Alors soit $a < b < c$ soit $a > b > c$ dans *tous les* π_i. Mais c'est impossible, puisqu'il doit exister une permutation pour laquelle b arrive après a et c. □

Le comportement asymptotique exact a été mis en évidence par Joel Spencer (borne supérieure) et par Füredi, Hajnal, Rödl et Trotter (borne inférieure) :

$$\dim(K_n) \;=\; \log_2\log_2 n + (\frac{1}{2} + o(1))\log_2\log_2\log_2 n.$$

Mais l'histoire n'est pas terminée : très récemment, Morris et Hoşten ont trouvé une méthode qui, en principe, établit la valeur *exacte* de $\dim(K_n)$. En utilisant leur résultat et un ordinateur, on peut obtenir les valeurs indiquées dans la marge. Ce résultat est véritablement étonnant ! Comment peut-on décider parmi toutes les permutations d'un ensemble 1422564 éléments si l'on a besoin de 7 ou de 8 d'entre elles pour représenter $K_{1422564}$?

π_1:	1	2	3	5	6	7	8	9	10	11	12	4
π_2:	2	3	4	8	7	6	5	12	11	10	9	1
π_3:	3	4	1	11	12	9	10	6	5	8	7	2
π_4:	4	1	2	10	9	12	11	7	8	5	6	3

Ces quatre permutations représentent K_{12}.

$$\begin{aligned}
\dim(K_n) &\leqslant 4 \iff n \leqslant 12,\\
\dim(K_n) &\leqslant 5 \iff n \leqslant 81,\\
\dim(K_n) &\leqslant 6 \iff n \leqslant 2646,\\
\dim(K_n) &\leqslant 7 \iff n \leqslant 1422564.
\end{aligned}$$

3. Sommes

Paul Erdős attribuait à Andrew Vázsonyi et Marta Sved la belle application du principe des tiroirs qui suit :

> ***Proposition.*** *Soient n entiers $a_1, \ldots, a_n$, distincts ou non. Alors il existe toujours un ensemble de nombres consécutifs $a_{k+1}, a_{k+2}, \ldots, a_\ell$ dont la somme $\sum_{i=k+1}^{\ell} a_i$ est un multiple de n.*

Posons $N = \{0, a_1, a_1+a_2, \ldots, a_1+a_2+\ldots+a_n\}$ et $R = \{0, 1, \ldots, n-1\}$. Considérons l'application $f : N \to R$, où $f(m)$ est le reste de m dans la division par n. Puisque $|N| = n+1 > n = |R|$, il existe deux sommes $a_1 + \ldots + a_k$ et $a_1 + \ldots + a_\ell$ $(k < \ell)$ ayant le *même* reste, la première somme pouvant être la somme vide valant 0. Par suite :

$$\sum_{i=k+1}^{\ell} a_i \;=\; \sum_{i=1}^{\ell} a_i \;-\; \sum_{i=1}^{k} a_i$$

a pour reste 0, ce qui conclut la démonstration. □

Abordons maintenant le second principe qui consiste à compter un ensemble d'objets de deux manières différentes.

Double décompte.
Soient deux ensembles finis L et C, et soit $S \subseteq L \times C$. Chaque fois que $(p, q) \in S$, on dit que p et q sont incidents.
Si ℓ_p désigne le nombre d'éléments qui sont incidents à $p \in L$ et si c_q désigne le nombre d'éléments qui sont incidents à $q \in C$, alors :

$$\sum_{p \in L} \ell_p \;=\; |S| \;=\; \sum_{q \in C} c_q. \tag{3}$$

Une fois de plus, il n'y a rien à prouver. La première somme considère les couples dans S selon le premier indice, alors que la seconde somme considère les mêmes couples selon le deuxième indice.

On peut représenter l'ensemble S de la manière suivante. Considérons la *matrice d'incidence* $A = (a_{pq})$ de S, où les lignes et les colonnes de A sont indexées par les éléments de L et C respectivement, avec

$$a_{pq} = \begin{cases} 1 & \text{si } (p, q) \in S \\ 0 & \text{si } (p, q) \notin S \end{cases}$$

Ainsi posé, ℓ_p est la somme de la p-ième ligne de A et c_q est la somme de la q-ième colonne. Par conséquent, la première somme dans (3) additionne les éléments de A (c'est-à-dire compte les 1 dans S) en suivant les lignes alors que la seconde le fait en suivant les colonnes .

L'exemple suivant devrait éclairer cette correspondance. Considérons $L = C = \{1, 2, \ldots, 8\}$ et posons $S = \{(i, j) : i \text{ divise } j\}$. Nous obtenons la matrice qui figure dans la marge où seuls les 1 sont affichés.

L \ C	1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1	1
2		1		1		1		1
3			1			1		
4				1				1
5					1			
6						1		
7							1	
8								1

4. Encore des nombres

Examinons la matrice qui vient d'être étudiée. Le nombre de 1 dans la colonne j est précisément le nombre de diviseurs de j ; notons $t(j)$ ce nombre. On peut se demander quel est l'ordre de grandeur de ce nombre $t(j)$ *en moyenne* lorsque j varie de 1 à n. Étudions donc la quantité :

$$\bar{t}(n) = \frac{1}{n} \sum_{j=1}^{n} t(j).$$

À première vue, il semble sans espoir de trouver une estimation de $\bar{t}(n)$. Pour un nombre premier p on trouve $t(p) = 2$ alors que pour 2^k nous obtenons la (grande) valeur $t(2^k) = k + 1$. Ainsi $t(n)$ est une fonction très irrégulière et l'on peut conjecturer qu'il en est de même pour $\bar{t}(n)$. Mauvaise intuition, c'est le contraire qui est vrai ! Un double décompte fournit une réponse inespérée et simple.

n	1	2	3	4	5	6	7	8
$\bar{t}(n)$	1	$\frac{3}{2}$	$\frac{5}{3}$	2	2	$\frac{7}{3}$	$\frac{16}{7}$	$\frac{5}{2}$

Les premières valeurs de $\bar{t}(n)$.

Considérons la matrice A précédente pour les entiers varient de 1 à n. En comptant par colonnes, nous obtenons : $\sum_{j=1}^{n} t(j)$. Combien y a-t-il de 1 dans la ligne i ? C'est assez facile, les 1 correspondent aux multiples de i : $1i, 2i, \ldots$ et le dernier multiple qui ne dépasse pas n est $\lfloor \frac{n}{i} \rfloor i$. On obtient par conséquent :

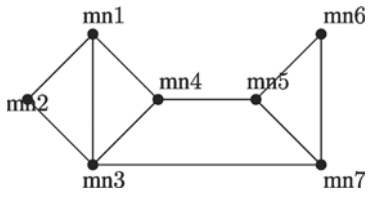
$$\bar{t}(n) \;=\; \frac{1}{n} \sum_{j=1}^{n} t(j) \;=\; \frac{1}{n} \sum_{i=1}^{n} \left\lfloor \frac{n}{i} \right\rfloor \;\leqslant\; \frac{1}{n} \sum_{i=1}^{n} \frac{n}{i} \;=\; \sum_{i=1}^{n} \frac{1}{i},$$

l'erreur dans chaque terme, en passant de $\lfloor \frac{n}{i} \rfloor$ à $\frac{n}{i}$, étant inférieure à 1. Par conséquent, l'erreur globale sur la moyenne est inférieure à 1 également. Maintenant, la dernière somme est le n-ième nombre harmonique H_n. En combinant l'encadrement $H_n - 1 < \bar{t}(n) < H_n$ avec l'estimation de H_n établie en page 11, on est conduit à :

$$\ln n - 1 \;<\; H_n - 1 - \frac{1}{n} \;<\; \bar{t}(n) \;<\; H_n \;<\; \ln n + 1.$$

Ainsi on obtient le résultat remarquable suivant : alors que le comportement de $t(n)$ est très irrégulier, la moyenne $\bar{t}(n)$ se comporte magnifiquement. $\bar{t}(n)$ vaut $\ln n$ à 1 près.

5. Graphes



Soit G un graphe fini simple dont l'ensemble de sommets est V et l'ensemble d'arêtes est E. Nous avons défini au chapitre 12 le *degré* $d(v)$ d'un sommet v comme le nombre d'arêtes qui admettent v pour extrémité. Dans l'exemple présenté sur la figure ci-contre, les sommets $1, 2, \ldots, 7$ sont respectivement de degré $3, 2, 4, 3, 3, 2, 3$.

Presque tous les livres traitant de la théorie des graphes commencent par le résultat suivant, déjà rencontré aux chapitres 12 et 18 :

$$\sum_{v \in V} d(v) \;=\; 2|E|. \tag{4}$$

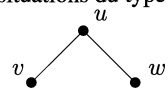
Pour démontrer ce résultat, considérons $S \subseteq V \times E$, où S est l'ensemble des paires (v, e) telles que $v \in V$ soit une extrémité de $e \in E$. En dénombrant S de deux façons différentes, on obtient d'un côté $\sum_{v \in V} d(v)$, puisque chaque sommet contribue pour $d(v)$, et de l'autre côté $2|E|$, puisque chaque arête a deux extrémités. □

Aussi simple que paraisse le résultat (4), il a plusieurs conséquences importantes. Certaines d'entre elles seront évoquées dans la suite de notre propos. Dans cette partie, nous voudrions mettre en évidence la belle application suivante à un *problème de maximum* sur les graphes. Voici le problème :

> *Soit $G = (V, E)$ un graphe à n sommets ne comportant pas de cycle de longueur 4 (noté C_4), c'est-à-dire pas de sous-graphe*  . *Quel est le nombre maximum d'arêtes que peut posséder G ?*

À titre d'exemple, le graphe à 5 sommets représenté ci-contre ne comporte pas de 4-cycle et a 6 arêtes. Le lecteur peut facilement montrer que, pour 5 sommets, le nombre maximum d'arêtes est 6, et que ce graphe est en fait le seul graphe de 5 sommets à 6 arêtes qui n'a pas de 4-cycle.

Abordons le problème général. Soit G un graphe à n sommets sans 4-cycle. Comme précédemment, on désigne par $d(u)$ le degré de u. Soit S l'ensemble des paires $(u, \{v, w\})$ telles que u soit adjacent à v et à w, avec $v \neq w$. Dénombrons S de deux façons différentes. En d'autres termes, nous dénombrons toutes les situations du type :



En sommant sur u, on trouve $|S| = \sum_{u \in V} \binom{d(u)}{2}$. D'autre part, chaque paire $\{v, w\}$ a au plus un voisin commun (à cause de la condition C_4). Par conséquent $|S| \leqslant \binom{n}{2}$, donc :

$$\sum_{u \in V} \binom{d(u)}{2} \;\leqslant\; \binom{n}{2}$$

soit encore :

$$\sum_{u \in V} d(u)^2 \;\leqslant\; n(n-1) + \sum_{u \in V} d(u) \tag{5}$$

Ensuite (c'est typique de ce genre de problème d'extremum) on applique l'inégalité de Cauchy-Schwarz aux deux vecteurs $(d(u_1), \ldots, d(u_n))$ et $(1, 1, \ldots, 1)$, obtenant ainsi :

$$\Big(\sum_{u\in V} d(u)\Big)^2 \leqslant n \sum_{u\in V} d(u)^2,$$

et par conséquent, en utilisant (5) :

$$\Big(\sum_{u\in V} d(u)\Big)^2 \leqslant n^2(n-1) + n\sum_{u\in V} d(u).$$

En invoquant (4) on trouve :

$$4\,|E|^2 \leqslant n^2(n-1) + 2n\,|E|$$

soit encore :

$$|E|^2 - \frac{n}{2}|E| - \frac{n^2(n-1)}{4} \leqslant 0$$

En résolvant l'équation du second degré correspondante, on est conduit au résultat suivant dû à Istvan Reiman.

Théorème. *Si un graphe G à n sommets ne contient pas de 4-cycle, alors*

$$|E| \leqslant \left\lfloor \frac{n}{4}\left(1 + \sqrt{4n-3}\right)\right\rfloor \tag{6}$$

Pour $n = 5$, cela donne $|E| \leqslant 6$; le graphe de l'exemple montre que l'égalité peut avoir lieu.

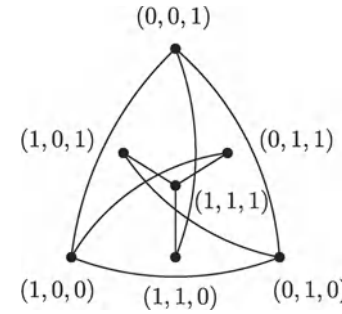
Compter de deux façons différentes a ainsi permis facilement de trouver une borne supérieure sur le nombre d'arêtes. Mais quelle est la finesse de la borne (6) en général ? Le bel exemple suivant ([2], [3], [6]) montre que c'est presque la meilleure possible. Comme souvent dans de tels problèmes, la géométrie finie montre la voie.

La présentation de cet exemple suppose que le lecteur est familier avec le corps fini $\mathbb{Z}_p$ des entiers modulo un nombre premier p (voir page 20). Considérons l'espace vectoriel $X = (\mathbb{Z}_p)^3$. On construit à partir de X le graphe G_p suivant : les sommets de G_p sont les sous-espaces de dimension 1 engendrés par les vecteurs non nuls de X, $[\boldsymbol{v}] := \mathrm{Vect}_{\mathbb{Z}_p}\{\boldsymbol{v}\}$, et deux tels sous-espaces $[\boldsymbol{v}]$ et $[\boldsymbol{w}]$ sont reliés par une arête si :

$$\langle \boldsymbol{v}, \boldsymbol{w}\rangle = v_1w_1 + v_2w_2 + v_3w_3 = 0.$$

Notons que le vecteur non nul choisi pour représenter le sous-espace n'a pas d'importance. Dans le langage de la géométrie, les sommets sont les *points* du plan projectif sur $\mathbb{Z}_p$ et $[\boldsymbol{w}]$ est adjacent à $[\boldsymbol{v}]$ si $\boldsymbol{w}$ se trouve sur la *droite polaire* de $\boldsymbol{v}$.

Par exemple, le graphe G_2 n'a pas de 4-cycle, contient 9 arêtes et atteint presque la borne 10 donnée par le théorème. Nous voulons montrer que ce résultat se produit pour tout nombre premier p.



Le graphe G_2 : ses sommets sont les sept triplets non nuls (x, y, z).

Prouvons d'abord que G_p vérifie la condition C_4. Si $[\boldsymbol{u}]$ est un voisin commun à $[\boldsymbol{v}]$ et à $[\boldsymbol{w}]$, $\boldsymbol{u}$ est solution des équations linéaires :

$$\begin{aligned} v_1x + v_2y + v_3z &= 0 \\ w_1x + w_2y + w_3z &= 0. \end{aligned}$$

Puisque $\boldsymbol{v}$ et $\boldsymbol{w}$ sont linéairement indépendants, l'espace des solutions est de dimension 1 et, par conséquent, le voisin commun $[\boldsymbol{u}]$ est unique.

Déterminons ensuite le nombre de sommets de G_p. C'est encore un double décompte qui conduit au résultat. L'espace X contient $p^3 - 1$ vecteurs non nuls. Puisque chaque sous-espace de dimension 1 contient $p - 1$ vecteurs non nuls, X a $\frac{p^3-1}{p-1} = p^2 + p + 1$ sous-espaces de dimension 1, c'est-à-dire que G_p a $n = p^2+p+1$ sommets. De même, tout sous-espace de dimension 2 contient $p^2 - 1$ vecteurs non nuls et, par conséquent, $\frac{p^2-1}{p-1} = p + 1$ sous-espaces de dimension 1.

Il reste à déterminer le nombre d'arêtes dans G_p, ou ce qui est équivalent selon (4), à déterminer les degrés. Par construction de G_p, les sommets adjacents à $[\boldsymbol{u}]$ sont les solutions de l'équation :

$$u_1x + u_2y + u_3z = 0 \tag{7}$$

L'espace des solutions de (7) est un sous-espace de dimension 2 ; il y a par conséquent $p+1$ sommets adjacents à $\boldsymbol{u}$. Cependant, il peut arriver que $[\boldsymbol{u}]$ lui-même soit une solution de (7). Dans ce cas, il y a seulement p sommets adjacents à $[\boldsymbol{u}]$.

En résumé, nous obtenons le résultat suivant : si $\boldsymbol{u}$ se trouve sur la *conique* définie par $x^2 + y^2 + z^2 = 0$, alors $d([\boldsymbol{u}]) = p$, sinon $d([\boldsymbol{u}]) = p + 1$. Il reste donc à trouver le nombre de sous-espaces de dimension 1 sur cette conique. Énonçons un résultat que nous montrerons dans un moment :

> ***Proposition.*** *L'équation $x^2+y^2+z^2 = 0$ a exactement p^2 solutions (x, y, z) et, par conséquent, (en éliminant la solution nulle) G_p a exactement $\frac{p^2-1}{p-1} = p + 1$ sommets de degré p.*

À l'aide de ce résultat, complétons notre analyse de G_p. Il y a $p+1$ sommets de degré p, donc $(p^2 + p + 1) - (p + 1) = p^2$ sommets de degré $p + 1$. En utilisant (4), on trouve :

$$\begin{aligned} |E| &= \frac{(p+1)p}{2} + \frac{p^2(p+1)}{2} = \frac{(p+1)^2p}{2} \\ &= \frac{(p+1)p}{4}\,(1 + (2p+1)) = \frac{p^2+p}{4}\,(1 + \sqrt{4p^2+4p+1}) \end{aligned}$$

En posant $n = p^2 + p + 1$, la dernière équation s'écrit :

$$|E| = \frac{n-1}{4}\,(1 + \sqrt{4n-3}),$$

ce qui est presque en accord avec (6).

Maintenant, revenons à la démonstration principale. L'argument suivant est une belle application d'algèbre linéaire utilisant des matrices symétriques et leurs valeurs propres. On retrouve la même méthode au chapitre 39, ce qui n'est pas une coïncidence : les deux démonstrations sont issues du même article d'Erdős, Rényi et Sós.

On note comme précédemment les sous-espaces de dimension 1 de X par des vecteurs $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{p^2+p+1}$, deux quelconques d'entre eux étant linéairement indépendants. De même, on peut représenter les sous-espaces de dimension 2 par le *même* ensemble de vecteurs, le sous-ensemble représenté par $\boldsymbol{u} = (u_1, u_2, u_3)$ étant l'espace de dimension 2 des solutions de l'équation $u_1 x + u_2 y + u_3 z = 0$ comme dans (7) (il ne s'agit que de l'application du principe de dualité de l'algèbre linéaire). Par conséquent, d'après (7), le sous-espace de dimension 1 représenté par $\boldsymbol{v}_i$ est contenu dans le sous-espace de dimension 2 représenté par $\boldsymbol{v}_j$, si et seulement si $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = 0$.

Considérons maintenant la matrice $A = (a_{ij})$ de dimensions $(p^2 + p + 1) \times (p^2 + p + 1)$ définie comme suit : les lignes et les colonnes de A correspondent à $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{p^2+p+1}$ (nous utilisons la même numérotation pour les lignes et les colonnes) et :

$$a_{ij} := \begin{cases} 1 & \text{si } \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = 0, \\ 0 & \text{sinon.} \end{cases}$$

A est donc une matrice symétrique réelle et $a_{ii} = 1$ si $\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle = 0$, c'est-à-dire précisément lorsque $\boldsymbol{v}_i$ se trouve sur la conique $x^2 + y^2 + z^2 = 0$. Ainsi, il ne reste plus qu'à montrer que :

$$\text{trace } A \;=\; p + 1.$$

L'algèbre linéaire nous dit que la trace est égale à la somme des valeurs propres. Et c'est ici qu'intervient l'astuce : alors que A semble compliquée, la matrice A^2 est facile à analyser. Remarquons d'abord les deux résultats suivants :

- toute ligne de A contient exactement $p+1$ fois 1. Cela implique que $p+1$ est une valeur propre de A, puisque $A\mathbf{1} = (p+1)\mathbf{1}$, où $\mathbf{1}$ est le vecteur dont toutes les composantes valent 1.
- Étant données deux lignes distinctes $\boldsymbol{v}_i$ et $\boldsymbol{v}_j$ il existe exactement une colonne ayant un 1 dans les deux lignes (la colonne correspondant à l'unique sous-espace engendré par $\boldsymbol{v}_i$ et $\boldsymbol{v}_j$).

En utilisant ces résultats, on trouve :

$$A^2 = \begin{pmatrix} p+1 & 1 & \cdots & 1 \\ 1 & p+1 & & \vdots \\ \vdots & & \ddots & \\ 1 & \cdots & & p+1 \end{pmatrix} \;=\; pI \;+\; J$$

où I est la matrice identité et J la matrice dont tous les coefficients sont des 1. D'autre part, J a pour valeurs propres $p^2 + p + 1$ (de multiplicité

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

La matrice de G_2 .

1) et 0 (de multiplicité p^2+p). Par conséquent, A^2 a pour valeurs propres $p^2+2p+1=(p+1)^2$ de multiplicité 1 et p de multiplicité p^2+p. Puisque A est réelle et symétrique, donc diagonalisable, A admet la valeur propre $p+1$ ou $-(p+1)$ et p^2+p valeurs propres égales à $\pm\sqrt{p}$. D'après la première remarque faite précédemment, la première valeur propre doit être $p+1$. Supposons que $\sqrt{p}$ ait une multiplicité r et $-\sqrt{p}$ une multiplicité s, alors :

$$\text{trace } A \;=\; (p+1)+r\sqrt{p}-s\sqrt{p}.$$

Nous sommes arrivés au bout de nos peines : comme la trace est entière, on doit avoir $r=s$, donc trace $A=p+1$. □

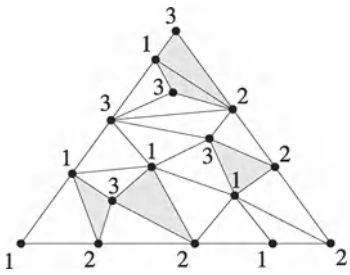
6. Le lemme de Sperner

En 1912, Luitzen Brouwer a publié son célèbre théorème du point fixe :

> *Toute fonction continue $f : B^n \longrightarrow B^n$ d'une boule de dimension n dans elle même admet un point fixe (c'est-à-dire un point $\boldsymbol{x} \in B^n$ tel que $f(\boldsymbol{x})=\boldsymbol{x}$).*

Pour la dimension 1, c'est-à-dire pour un intervalle, ce résultat se déduit facilement du théorème des valeurs intermédiaires, mais pour des dimensions plus grandes, la preuve de Brouwer repose sur un raisonnement sophistiqué. Par conséquent, il fut très surprenant de voir qu'en 1928 le jeune Emanuel Sperner (il avait 23 ans à l'époque) montre un résultat combinatoire simple à partir duquel on peut établir le théorème du point fixe de Brouwer ainsi que l'invariance de la dimension par bijection continue. Et, que demander de plus, le lemme ingénieux de Sperner est assorti d'une preuve également magnifique : un double décompte.

Nous allons étudier le lemme de Sperner et le théorème de Brouwer qui s'en déduit dans le premier cas intéressant, celui de la dimension $n=2$. Le lecteur ne devrait pas avoir de difficultés à étendre les démonstrations à des dimensions plus grandes en raisonnant par récurrence sur la dimension.

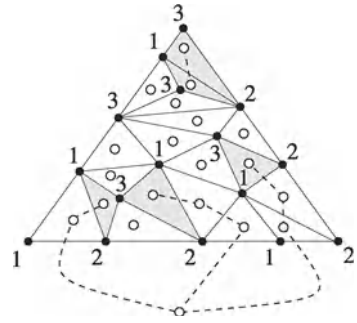


Les triangles de trois couleurs différentes sont ombrés.

Lemme de Sperner.
Supposons qu'un « grand » triangle de sommets V_1, V_2, V_3 soit triangulé (c'est-à-dire décomposé en un nombre fini de « petits » triangles qui s'ajustent ensemble arête par arête).
Supposons que les sommets de la triangulation aient des « couleurs » choisies dans l'ensemble $\{1,2,3\}$ telles que V_i soit de la couleur i (pour tout i) et supposons que seules les couleurs i et j soient utilisées pour les sommets qui se trouvent le long de l'arête allant de V_i à V_j (pour $i \neq j$), tandis que les sommets intérieurs peuvent être coloriés arbitrairement avec 1, 2 ou 3. Alors il doit y avoir dans la triangulation un petit triangle « tricolore », c'est-à-dire dont les trois sommets sont de couleurs différentes.

■ **Preuve.** Nous allons montrer un résultat plus fort : non seulement le nombre de triangles tricolores est non nul, mais il est toujours *impair*.

Considérons le graphe dual de la triangulation, mais sans retenir toutes ses arêtes : on retient seulement celles qui coupent une arête du graphe initial ayant ses extrémités de couleurs (différentes) 1 et 2. On obtient ainsi un « graphe dual partiel » dont chaque sommet a pour degré 1 s'il correspond à un triangle tricolore, un degré 2 s'il correspond à un triangle dans lesquels seules les deux couleurs 1 et 2 apparaissent, et un degré 0 s'il correspond à un triangle qui ne présente pas les couleurs 1 et 2. Ainsi, seuls les triangles tricolores correspondent aux sommets de degré impair (de degré 1).



Le sommet du graphe dual qui correspond à l'extérieur de la triangulation a un degré impair : en effet, le long de la grande arête de V_1 à V_2, il y a un nombre impair de changements entre 1 et 2. Un nombre impair d'arêtes du graphe dual partiel coupent donc cette grande arête, alors que les autres grandes arêtes ne peuvent pas avoir les deux couleurs 1 et 2.

Maintenant, comme le nombre de sommets impairs d'un graphe fini est pair (d'après l'équation (4)), le nombre de petits triangles ayant trois couleurs différentes (qui correspondent aux sommets impairs intérieurs du graphe dual) est impair. □

On déduit facilement le théorème de Brouwer de ce lemme.

■ **Preuve du théorème du point fixe de Brouwer (pour $n = 2$).** Soit Δ le triangle de $\mathbb{R}^3$ de sommets $\boldsymbol{e}_1 = (1,0,0)$, $\boldsymbol{e}_2 = (0,1,0)$ et $\boldsymbol{e}_3 = (0,0,1)$. Il suffit de montrer que toute application continue $f : \Delta \longrightarrow \Delta$ a un point fixe, puisque Δ est homéomorphe à la boule B_2 de dimension 2.

On note $\delta(\mathcal{T})$ la longueur maximale d'une arête d'une triangulation $\mathcal{T}$. On peut facilement construire une suite infinie de triangulations $\mathcal{T}_1, \mathcal{T}_2, \ldots$ de Δ telle que la suite des diamètres maximaux $\delta(\mathcal{T}_k)$ converge vers 0. On peut obtenir une telle suite par une construction explicite ou par récurrence, en prenant par exemple pour $\mathcal{T}_{k+1}$ la subdivision barycentrique de $\mathcal{T}_k$.

Pour chaque triangulation, on définit une 3-coloration de ses sommets $\boldsymbol{v}$ en posant $\lambda(\boldsymbol{v}) := \min\{i : f(\boldsymbol{v})_i < v_i\}$, c'est-à-dire que $\lambda(\boldsymbol{v})$ est le plus petit indice i tel que la i-ième coordonnée de $f(\boldsymbol{v}) - \boldsymbol{v}$ soit négative. En supposant que f n'a pas de point fixe, une telle définition est valide. Pour s'en convaincre, il suffit de remarquer que chaque $\boldsymbol{v} \in \Delta$ est dans le plan $x_1 + x_2 + x_3 = 1$ et, par conséquent, $\sum_i v_i = 1$. Donc si $f(\boldsymbol{v}) \neq \boldsymbol{v}$, l'une au moins des coordonnées de $f(\boldsymbol{v}) - \boldsymbol{v}$ doit être négative (et l'une au moins doit être positive).

Vérifions que cette coloration satisfait les hypothèses du lemme de Sperner. D'abord, le sommet $\boldsymbol{e}_i$ doit recevoir la couleur i, puisque la seule composante négative possible de $f(\boldsymbol{e}_i) - \boldsymbol{e}_i$ est la i-ième composante. De plus, si $\boldsymbol{v}$ est sur l'arête opposée de $\boldsymbol{e}_i$, alors $v_i = 0$, la i-ième composante de $f(\boldsymbol{v}) - \boldsymbol{v}$ n'est pas négative et ainsi $\boldsymbol{v}$ n'est pas de couleur i.

Le lemme de Sperner permet maintenant d'affirmer qu'il y a dans chaque triangulation $\mathcal{T}_k$ un triangle tricolore $\{\boldsymbol{v}^{k:1}, \boldsymbol{v}^{k:2}, \boldsymbol{v}^{k:3}\}$ avec $\lambda(\boldsymbol{v}^{k:i}) = i$. La suite de points $(\boldsymbol{v}^{k:1})_{k \geqslant 1}$ ne converge pas nécessairement, mais puisque le simplexe Δ est compact, il existe une sous-suite de cette suite qui converge. Après avoir remplacé la suite de triangulations $\mathcal{T}_k$ par la sous-suite cor-

respondante (encore désignée par $\mathcal{T}_k$ pour simplifier) nous pouvons supposer que $(\boldsymbol{v}^{k:1})_k$ converge vers un point $\boldsymbol{v} \in \Delta$. Mais la distance de $\boldsymbol{v}^{k:2}$ et $\boldsymbol{v}^{k:3}$ à $\boldsymbol{v}^{k:1}$ est au plus égale à la longueur de la maille $\delta(\mathcal{T}_k)$, qui converge vers 0. Ainsi, les suites $(\boldsymbol{v}^{k:2})$ et $(\boldsymbol{v}^{k:3})$ convergent vers le *même* point $\boldsymbol{v}$.

Où se trouve $f(\boldsymbol{v})$? Nous savons que la première coordonnée $f(\boldsymbol{v}^{k:1})$ est inférieure à celle de $\boldsymbol{v}^{k:1}$ pour tout k. Puisque f est continue, la première coordonnée de $f(\boldsymbol{v})$ est inférieure ou égale à celle de $\boldsymbol{v}$. Le même raisonnement s'applique pour les deuxième et troisième coordonnées. Par conséquent, aucune des coordonnées de $f(\boldsymbol{v}) - \boldsymbol{v}$ n'est positive, ce qui contredit l'hypothèse $f(\boldsymbol{v}) \neq \boldsymbol{v}$. □

Bibliographie

[1] L. E. J. BROUWER : *Über Abbildungen von Mannigfaltigkeiten,* Math. Annalen **71** (1912), 97-115.

[2] W. C. BROWN : *On graphs that do not contain a Thomsen graph,* Canadian Math. Bull. **9** (1966), 281-285.

[3] P. ERDŐS, A. RÉNYI & V. SÓS : *On a problem of graph theory,* Studia Sci. Math. Hungar. **1** (1966), 215-235.

[4] P. ERDŐS & G. SZEKERES : *A combinatorial problem in geometry,* Compositio Math. (1935), 463-470.

[5] S. HOŞTEN & W. D. MORRIS : *The order dimension of the complete graph,* Discrete Math. **201** (1999), 133-139.

[6] I. REIMAN : *Über ein Problem von K. Zarankiewicz,* Acta Math. Acad. Sci. Hungar. **9** (1958), 269-273.

[7] J. SPENCER : *Minimal scrambling sets of simple orders,* Acta Math. Acad. Sci. Hungar. **22** (1971), 349-353.

[8] E. SPERNER : *Neuer Beweis für die Invarianz der Dimensionszahl und des Gebietes,* Abh. Math. Sem. Hamburg **6** (1928), 265-272.

[9] W. T. TROTTER : *Combinatorics and Partially Ordered Sets : Dimension Theory,* John Hopkins University Press, Baltimore and London 1992.

Pavages de rectangles

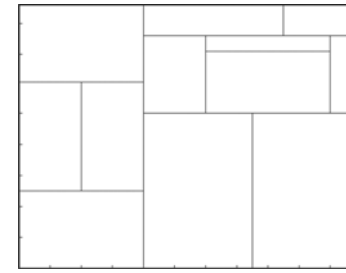
Chapitre 26

Certains théorèmes présentent une caractéristique particulière : l'énoncé est élémentaire mais la démonstration paraît très difficile jusqu'à ce que l'on ouvre une porte magique et que tout devienne simple et limpide.

Le résultat suivant, dû à Nicolaas de Bruijn, en est un exemple typique.

> **Théorème.** *Si un rectangle est pavé par des rectangles qui présentent tous au moins un côté de longueur entière, alors ce rectangle présente lui aussi au moins un côté de longueur entière.*

Par pavage d'un rectangle R, on entend un recouvrement de R à l'aide de rectangles $T_1, \ldots, T_m$ d'intérieurs disjoints comme on le voit sur la figure de droite.



Le grand rectangle a des côtés de longueurs respectives 11 et 8.5.

En fait, de Bruijn a démontré le résultat suivant à propos de la meilleure façon de ranger des copies d'un rectangle de dimensions $a \times b$ pour paver un rectangle de dimensions $c \times d$: si a,b,c et d sont des entiers, alors a et b doivent chacun diviser l'un des deux nombres c et d. Cela résulte de l'application répétée du théorème plus général énoncé ci-dessus au petit rectangle réduit par homothétie d'un facteur $\frac{1}{a}$ d'une part, puis d'un facteur $\frac{1}{b}$ d'autre part. Le petit rectangle se retrouve dans chaque cas avec un côté de longueur 1 et donc $\frac{c}{a}$ ou $\frac{d}{a}$ doit être un entier.

La première idée qui vient à l'esprit pour démontrer le théorème consiste à procéder par récurrence sur le nombre de petits rectangles. On peut effectivement y parvenir mais il faut faire preuve de beaucoup de soin et la démonstration n'est pas un modèle d'élégance. Dans un article très plaisant, Stan Wagon passe en revue pas moins de quatorze démonstrations différentes. Nous en avons retenu trois ; aucune d'entre elles ne fait appel au principe de récurrence. La première démonstration, due à de Bruijn, utilise un calcul d'analyse particulièrement astucieux. La deuxième, proposée par Richard Rochberg et Sherman Stein, est une version discrétisée de la première qui devient ainsi encore plus simple. Cependant, la démonstration la plus remarquable est probablement la troisième, suggérée par Mike Paterson. Elle consiste à effectuer un dénombrement de deux manières différentes et elle tient presqu'en une ligne.

Dans la suite, nous supposons que les côtés du grand rectangle R sont respectivement parallèles aux axes des abscisses et des ordonnées et que son sommet inférieur gauche est placé sur l'origine $(0, 0)$. Les petits rectangles T_i ont alors, eux aussi, leurs côtés qui sont parallèles aux axes.

■ **Première démonstration.** Soit T un rectangle quelconque du plan, T s'étendant entre a et b sur l'axe des abscisses et entre c et d sur l'axe des ordonnées. Voici l'astuce de de Bruijn. Considérons l'intégrale double sur T suivante :

$$I := \int_c^d \int_a^b e^{2\pi i(x+y)} dx\, dy. \tag{1}$$

Comme :

$$\int_c^d \int_a^b e^{2\pi i(x+y)} dx\, dy \;=\; \int_a^b e^{2\pi ix} dx \cdot \int_c^d e^{2\pi iy} dy,$$

il s'ensuit que l'intégrale I est nulle si et seulement si au moins l'un des deux facteurs $\int_a^b e^{2\pi ix} dx$ ou $\int_c^d e^{2\pi iy} dy$ est égal à 0.
Nous allons montrer que :

$$\int_a^b e^{2\pi ix} dx = 0 \quad \Longleftrightarrow \quad b-a \text{ est un entier.} \tag{2}$$

$\iint_R f(x,y) = \sum_i \iint_{T_i} f(x,y)$
Additivité de l'intégrale.

Nous aurons alors terminé la démonstration. En effet, d'après la nature du pavage (les petits rectangles ont au moins un côté de longueur entière), chaque $\iint_{T_i}$ est égal à 0. Donc, tenant compte de l'additivité de l'intégrale, $\iint_R$ est nul aussi et donc R présente au moins un côté de longueur entière.
Il reste à vérifier (2). De :

$$\begin{aligned} \int_a^b e^{2\pi ix} dx &= \frac{1}{2\pi i} e^{2\pi ix}\Big|_a^b = \frac{1}{2\pi i}(e^{2\pi ib - 2\pi ia}) \\ &= \frac{e^{2\pi ia}}{2\pi i}(e^{2\pi i(b-a)} - 1)\,, \end{aligned}$$

on déduit que :

$$\int_a^b e^{2\pi ix} dx = 0 \quad \Longleftrightarrow \quad e^{2\pi i(b-a)} = 1.$$

Comme $e^{2\pi ix} = \cos 2\pi x + i \sin 2\pi x$ la précédente équation est équivalente à :

$$\cos 2\pi(b-a) = 1 \;\text{ et }\; \sin 2\pi(b-a) = 0.$$

Or $\cos x = 1$ si et seulement si x est un multiple entier de 2π, il faut et il suffit donc que $b-a \in \mathbb{Z}$ (ce qui assure aussi que $\sin 2\pi(b-a) = 0$). □

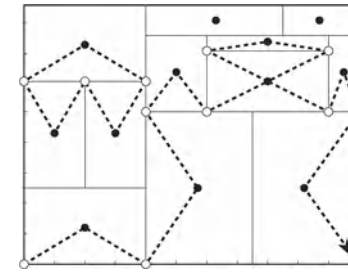
■ **Deuxième démonstration.** On colorie le plan à la manière d'un échiquier avec des carrés noirs ou blancs de taille $\frac{1}{2} \times \frac{1}{2}$, en commençant avec un carré noir en $(0, 0)$.

L'hypothèse sur le pavage (chaque petit rectangle a au moins un côté de longueur entière) implique que chaque rectangle T_i doit contenir autant de blanc que de noir. Par conséquent, le grand rectangle R doit, lui aussi, contenir autant de blanc que de noir.

Ceci implique que R a un côté de longueur entière car sinon on pourrait le découper en quatre parties, trois d'entre elles ayant un nombre égal de blanc et de noir, alors que la quatrième, située dans le coin supérieur droit, ne respecte pas cet équilibre. En effet, si $x = a - \lfloor a \rfloor$, $y = b - \lfloor b \rfloor$, avec $0 < x, y < 1$, la quantité de noir est supérieure à la quantité de blanc. Cette situation est illustrée par la figure ci-contre. □

La quantité de noir dans le rectangle du coin est $\min(x, \frac{1}{2}) \cdot \min(y, \frac{1}{2}) + \max(x - \frac{1}{2}, 0) \cdot \max(y - \frac{1}{2}, 0)$, ce qui est toujours supérieur à $\frac{1}{2}xy$.

■ **Troisième démonstration.** Soit C l'ensemble des sommets du pavage (les coins) dont les deux coordonnées sont entières (ainsi, par exemple, $(0, 0) \in C$) et soit T l'ensemble des pavés (les petits rectangles) du pavage. On forme un graphe biparti G, d'ensemble de sommets $C \cup T$, en joignant par une arête chaque coin $c \in C$ à tous les pavés dont il est un coin. L'hypothèse implique que chaque pavé est relié à 0, 2 ou 4 coins dans C car si un coin de pavé est dans C, il en va nécessairement de même pour le coin qui est son homologue à l'autre bout du côté entier originaire de ce sommet. G est donc constitué d'un nombre *pair* d'arêtes. Considérons maintenant C. Chaque $c \in C$ qui n'est pas un coin de R est relié à un nombre *pair* de pavés. Toutefois, le coin $(0, 0)$ n'est relié qu'à un seul pavé. Donc il doit exister un autre $c \in C$ de degré impair, et cet élément c ne peut être qu'un des autres coins de R. □

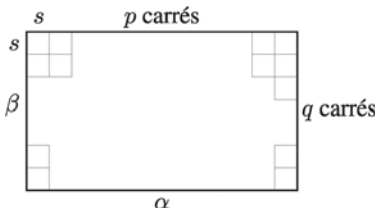


Ici le graphe biparti G est représenté avec des sommets blancs quand ils sont dans C, noirs quand ils sont dans T et des arêtes en pointillés.

Ces trois preuves peuvent être très facilement adaptées pour obtenir une version n-dimensionnelle du résultat de de Bruijn : si un parallélépipède n-dimensionnel R est pavé par des parallélépipèdes ayant chacun au moins un côté de longueur entière, alors R présente au moins un côté de longueur entière.

Nous allons rester en dimension 2 pour la discussion qui suit et considérer un « résultat compagnon » de celui de de Bruijn. Il a été proposé par Max Dehn bien des années plus tôt ; il semble d'allure similaire mais fait appel à des idées très différentes.

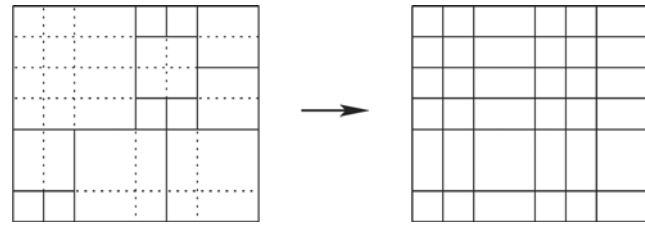
> **Théorème.** *Un rectangle peut être pavé par des carrés si et seulement si les longueurs de ses côtés sont commensurables (c'est-à-dire si le rapport des longueurs de ses côtés est un nombre rationnel).*



Désignons respectivement par α et β les longueurs des côtés du rectangle R où $\frac{\alpha}{\beta} \in \mathbb{Q}$, c'est-à-dire $\frac{\alpha}{\beta} = \frac{p}{q}$ avec $p, q \in \mathbb{N}$. Posons $s := \frac{\alpha}{p} = \frac{\beta}{q}$, on peut alors paver R avec des copies du carré $s \times s$ comme le montre la figure ci-contre.

Pour démontrer l'implication réciproque, Max Dehn a eu recours a un argument élégant qu'il avait déjà utilisé dans sa solution au troisième problème de Hilbert (voir chapitre 9). Les articles qu'il rédigea à ce propos parurent à un an d'intervalle dans les *Mathematische Annalen* [*Annales mathématiques*].

■ **Preuve.** Supposons que R est pavé par des carrés qui peuvent être de différentes tailles. Par homothétie, on peut supposer que R est un rectangle de dimensions $a \times 1$. Supposons que $a \notin \mathbb{Q}$ et montrons que l'on obtient une contradiction. La première étape consiste à étendre les côtés des carrés à toute la longueur et à toute la largeur de R, comme indiqué dans la figure suivante.



Le quadrillage ainsi obtenu décompose R en un certain nombre de petits rectangles qui constituent un pavage étendu de R. Notons $a_1, a_2, \ldots, a_M$ les longueurs de leurs côtés respectifs (peu importe l'ordre d'énumération) et considérons l'ensemble :

$$A := \{1, a, a_1, \ldots, a_M\} \subset \mathbb{R}\,.$$

Maintenant arrive un peu d'algèbre linéaire. Appelons $V(A)$ le $\mathbb{Q}$-espace vectoriel des combinaisons linéaires à coefficients rationnels des éléments de A. Remarquons que $V(A)$ contient toutes les longueurs des côtés du pavage original puisque chacune de ces longueurs s'écrit comme somme de certains a_i. Comme le nombre a n'est pas rationnel, la famille $\{1, a\}$ est libre et on peut la compléter en une base B de $V(A)$. Écrivons :

$$B = \{b_1 = 1, b_2 = a, b_3, \ldots, b_m\}\,.$$

f est une application $\mathbb{Q}$-linéaire si et seulement si :
$f(q_1 b_1 + \cdots + q_m b_m) =$
$q_1 f(b_1) + \cdots + q_m f(b_m)$
pour tous $q_1, \ldots, q_m \in \mathbb{Q}$.

Soit $f : V(A) \to \mathbb{R}$ l'application $\mathbb{Q}$-linéaire définie par l'image qu'elle donne de la base B :

$$f(1) := 1, \quad f(a) := -1, \quad \text{et} \quad f(b_i) := 0 \ \text{ pour } \ i \geqslant 3\,.$$

On adopte alors la définition — non standard ! — suivante pour déterminer « l'aire » d'un rectangle. Pour $c, d \in V(A)$ « l'aire » du rectangle $c \times d$ est définie par :

$$\text{aire}(\underset{c}{\square} d) = f(c) f(d).$$

On termine alors la preuve en trois étapes rapides :

(1) aire $\left(\underset{c_1}{\square}\underset{c_2}{\square} d\right) = \text{aire}\left(\underset{c_1}{\square} d\right) + \text{aire}\left(\underset{c_2}{\square} d\right)$.
Cela résulte immédiatement de la linéarité de f. Un résultat analogue s'obtient pour les bandes verticales.

(2) $\text{aire}(R) = \sum_{\text{carrés}} \text{aire}\left(\square\right)$, où la somme s'étend sur tous les carrés du pavage.
Remarquons simplement que, d'après (1), $\text{aire}(R)$ est égal à la somme des aires de tous les petits rectangles du pavage étendu. Comme chacun de ces rectangles est situé dans un carré du pavage initial et un seul, on voit, toujours par (1), que cette somme est aussi égale au membre de droite de (2).

(3) Par ailleurs, par application de la définition de « l'aire », on obtient :

$$\text{aire}(R) = f(a)f(1) = -1,$$

alors que pour un carré de côté t, $\text{aire}\left(\underset{t}{\square}\right) = f(t)^2 \geqslant 0$ et donc

$$\text{aire}(R) = \sum_{\text{carrés}} \text{aire}\left(\square\right) \geqslant 0,$$

ce qui conduit à la contradiction recherchée. □

Nous recommandons vivement le très bel article [1] rédigé par Federico Ardila et Richard Stanley à ceux qui voudraient faire quelques excursions supplémentaires dans le monde des pavages.

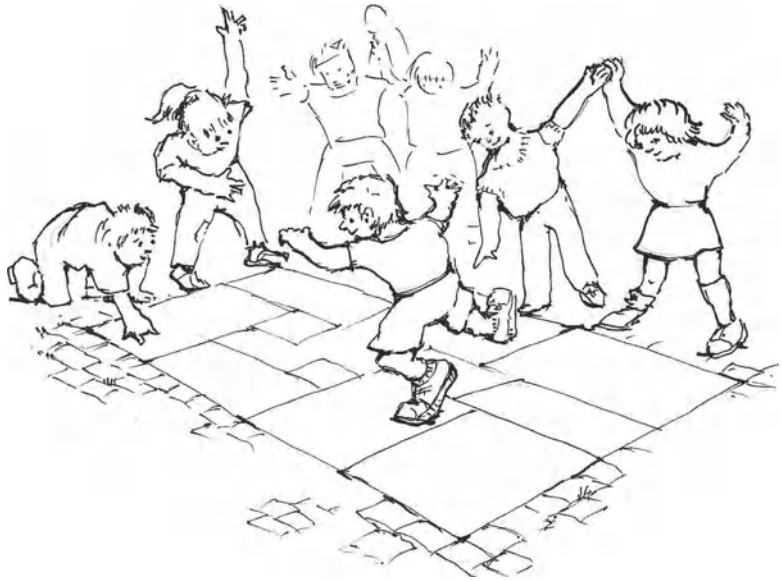
Bibliographie

[1] F. ARDILA & R. P. STANLEY : *Tilings,* Math. Intelligencer (4) **32** (2010), 32-43.

[2] N. G. DE BRUIJN : *Filling boxes with bricks,* Amer. Math. Monthly **76** (1969), 37-40.

[3] M. DEHN : *Über die Zerlegung von Rechtecken in Rechtecke,* Mathematische Annalen **57** (1903), 314-332.

[4] S. WAGON : *Fourteen proofs of a result about tiling a rectangle,* Amer. Math. Monthly **94** (1987), 601-617.



« La nouvelle marelle : évitez les entiers ! »

Trois théorèmes célèbres sur les ensembles finis

Chapitre 27

Ce chapitre est consacré à un thème fondamental de combinatoire : les propriétés et les cardinaux de familles particulières $\mathcal{F}$ de sous-ensembles d'un ensemble fini $N = \{1, 2, \ldots, n\}$. Nous commençons par deux résultats classiques dans ce domaine : les théorèmes de Sperner et d'Erdős-Ko-Rado. Ces deux résultats ont en commun d'avoir été démontrés plusieurs fois et d'avoir ouvert un nouveau domaine en théorie combinatoire des ensembles. Dans les deux cas, un raisonnement par récurrence semble naturel, mais les arguments évoqués sont très différents et vraiment inspirés.

Emanuel Sperner

En 1928 Emanuel Sperner s'est posé et a résolu la question suivante : supposons que l'on se donne l'ensemble $N = \{1, 2, \ldots, n\}$. Appelons *antichaîne* une famille $\mathcal{F}$ de sous-ensembles de N telle qu'aucun ensemble de $\mathcal{F}$ ne contienne un autre ensemble de la famille $\mathcal{F}$. Quelle est la taille maximale d'une antichaîne ? Il est clair que la famille $\mathcal{F}_k$ de tous les k-ensembles vérifie la propriété d'antichaîne avec $|\mathcal{F}_k| = \binom{n}{k}$. En cherchant le maximum des coefficients binomiaux (voir page 12) on voit qu'il y a une antichaîne de taille $\binom{n}{\lfloor n/2 \rfloor} = \max_k \binom{n}{k}$. Le théorème de Sperner affirme qu'il n'y en n'a pas de plus grande.

Théorème 1. *La taille maximale d'une antichaîne d'un n-ensemble est :* $\binom{n}{\lfloor n/2 \rfloor}$.

■ **Preuve.** Parmi toutes les preuves, celle qui suit, due à Lubell, est probablement la plus courte et la plus élégante. Soit $\mathcal{F}$ une antichaîne arbitraire. Nous devons montrer que $|\mathcal{F}| \leqslant \binom{n}{\lfloor n/2 \rfloor}$. La clé de la preuve consiste à considérer des *chaînes* de sous-ensembles $\varnothing = C_0 \subset C_1 \subset C_2 \subset \ldots \subset C_n = N$, où $|C_i| = i$ pour $i = 0, \ldots, n$. Combien y a-t-il de chaînes ? Il est clair que l'on obtient une chaîne en ajoutant un par un les éléments de N. Il y a donc exactement autant de chaînes qu'il y a de permutations de N, à savoir $n!$. Ensuite, étant donné un ensemble $A \in \mathcal{F}$ déterminons combien de ces chaînes contiennent A. C'est encore facile. Pour passer de $\varnothing$ à A, on doit ajouter les éléments de A un par un, et, pour passer de A à N, on doit ajouter les éléments restants. Ainsi, si A contient k éléments, en considérant tous ces couples de chaînes reliées entre elles, on voit qu'il y a précisément $k!(n-k)!$ chaînes de ce type. Notons qu'aucune chaîne ne peut traverser deux ensembles différents A et B de $\mathcal{F}$, puisque $\mathcal{F}$ est une antichaîne.

Pour terminer la preuve, notons m_k le nombre de k-ensembles de $\mathcal{F}$. On a $|\mathcal{F}| = \sum_{k=0}^{n} m_k$. On déduit de notre discussion que le nombre de chaînes

qui traversent un élément de $\mathcal{F}$ est :

$$\sum_{k=0}^{n} m_k \, k! \, (n-k)!$$

et que cette expression ne peut dépasser le nombre $n!$ de *toutes* les chaînes. Par conséquent :

$$\sum_{k=0}^{n} m_k \frac{k!(n-k)!}{n!} \leqslant 1 \qquad \text{ou} \qquad \sum_{k=0}^{n} \frac{m_k}{\binom{n}{k}} \leqslant 1.$$

En remplaçant les dénominateurs par le plus grand coefficient binomial, on obtient :

$$\frac{1}{\binom{n}{\lfloor n/2 \rfloor}} \sum_{k=0}^{n} m_k \leqslant 1 \qquad \text{c'est à dire} \qquad |\mathcal{F}| = \sum_{k=0}^{n} m_k \leqslant \binom{n}{\lfloor n/2 \rfloor}$$

et la démonstration est terminée. □

On vérifie que la famille de tous les $\frac{n}{2}$-ensembles pour n pair, les deux familles de tous les $\frac{n-1}{2}$-ensembles et de tous les $\frac{n+1}{2}$-ensembles pour n impair, sont les *seules* antichaînes de taille maximale !

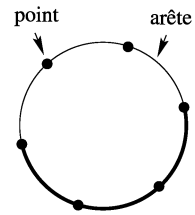
Le deuxième résultat présenté ici est de nature totalement différente. Considérons encore l'ensemble $N = \{1, \ldots, n\}$. On dit qu'une famille $\mathcal{F}$ de sous-ensembles est une *famille intersectante* si deux ensembles quelconques de $\mathcal{F}$ ont au moins un élément commun. Il est presque immédiat que la taille de la plus grande famille intersectante est 2^{n-1}. Si $A \in \mathcal{F}$, son complémentaire $A^c = N \backslash A$ a une intersection vide avec A et ne peut donc pas être dans $\mathcal{F}$. Nous concluons qu'une famille intersectante contient au plus la moitié des 2^n sous-ensembles, c'est à dire $|\mathcal{F}| \leqslant 2^{n-1}$. D'autre part, en considérant la famille de tous les ensembles qui contiennent un élément fixé, par exemple la famille $\mathcal{F}_1$ de tous les ensembles qui contiennent 1, on voit que $|\mathcal{F}_1| = 2^{n-1}$; le problème est réglé.

Posons maintenant la question suivante : quelle taille peut avoir une famille intersectante $\mathcal{F}$ si tous les ensembles de $\mathcal{F}$ ont la même taille k ? Nous appellerons de telles familles des *k-familles intersectantes*. Pour éviter les banalités, nous supposons $n \geqslant 2k$ car sinon deux k-ensembles quelconques ont nécessairement une intersection non vide et il n'y a rien à prouver. En reprenant l'idée précédente, nous sommes certains d'obtenir une telle famille $\mathcal{F}_1$ en considérant tous les k-ensembles qui contiennent un élément fixé, par exemple 1. Il est donc clair que nous obtenons tous les ensembles de $\mathcal{F}_1$ en ajoutant à 1 tous les $(k-1)$-sous-ensembles de $\{2, 3, \ldots, n\}$. Par conséquent $|\mathcal{F}_1| = \binom{n-1}{k-1}$. Est-il possible de faire mieux ? Le théorème d'Erdős-Ko-Rado affirme que non.

Théorème 2. *La taille maximale d'une k-famille intersectante dans un n-ensemble est $\binom{n-1}{k-1}$ lorsque $n \geqslant 2k$.*

Paul Erdős, Chao Ko et Richard Rado ont trouvé ce résultat en 1938, mais il ne fut publié que 23 ans plus tard. Depuis, on en a donné une multitude de variantes et de preuves, mais l'argument suivant que l'on doit à Gyula Katona est particulièrement élégant.

■ **Preuve.** La clé de la preuve est le simple lemme suivant qui, à première vue, semble totalement sans rapport avec le problème. Considérons un cercle C divisé par n points en n arêtes. Nous appelons *arc* de longueur k la réunion de $k+1$ points consécutifs et des k arêtes situées entre eux.



Un cercle C avec $n = 6$. Les arêtes en gras dessinent un arc de longueur 3.

Lemme. *Soit $n \geqslant 2k$ et soient t arcs distincts $A_1, \ldots, A_t$ de longueur k, tels que deux arcs quelconques aient une arête commune. Alors $t \leqslant k$.*

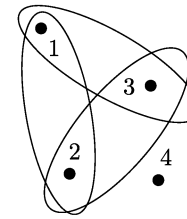
Remarquons d'abord que tout point de C est l'extrémité d'un arc au plus. En effet, si A_i et A_j avaient une extrémité commune v, ils devraient partir dans des directions différentes (puisqu'ils sont distincts). Cependant, ils ne peuvent alors pas avoir une arête en commun puisque $n \geqslant 2k$. Fixons A_1. Comme tout A_i $(i \geqslant 2)$ a une arête en commun avec A_1, l'une des extrémités de A_i est un point intérieur à A_1. Puisque ces extrémités doivent être distinctes, comme nous venons de le voir, et, puisque A_1 contient $k-1$ points intérieurs, il y a au plus $k-1$ autres arcs, donc au plus k arcs en tout. □

Poursuivons maintenant la preuve du théorème d'Erdős-Ko-Rado. Soit $\mathcal{F}$ une famille k-intersectante. Considérons un cercle C à n points et n arêtes comme précédemment, considérons une permutation circulaire quelconque $\pi = (a_1, a_2, \ldots, a_n)$ et écrivons les nombres a_i à coté des arêtes de C dans le sens des aiguilles d'une montre. Comptons les ensembles $A \in \mathcal{F}$ qui apparaissent comme k nombres *consécutifs* sur C. Comme $\mathcal{F}$ est une famille intersectante, le lemme implique que l'on obtient au plus k ensembles de ce type. Puisque cela se produit pour toute permutation circulaire et puisqu'il y a $(n-1)!$ permutations circulaires, on obtient par ce procédé au plus :

$$k(n-1)!$$

ensembles de $\mathcal{F}$ qui apparaissent comme des éléments consécutifs de permutations circulaires. Combien de fois compte-t-on un ensemble fixé $A \in \mathcal{F}$? C'est assez facile : A apparaît dans π si les k éléments de A apparaissent consécutivement dans un certain ordre. Par conséquent, il y a $k!$ possibilités d'écrire A et $(n-k)!$ façons d'ordonner les éléments restants. Un ensemble fixé A apparaît donc dans $k!(n-k)!$ permutations circulaires exactement ; par conséquent :

$$|\mathcal{F}| \leqslant \frac{k(n-1)!}{k!(n-k)!} = \frac{(n-1)!}{(k-1)!(n-1-(k-1))!} = \binom{n-1}{k-1} \quad \square$$



Une famille intersectante pour $n = 4$, $k = 2$.

On peut encore se demander si les familles qui contiennent un élément fixé sont les seules k-familles intersectantes maximales. Ce n'est certainement pas vrai pour $n = 2k$. Par exemple, pour $n = 4$ et $k = 2$ la famille $\{1,2\}, \{1,3\}, \{2,3\}$ a aussi pour taille $\binom{3}{1} = 3$. Plus généralement, pour $n = 2k$ les k-familles intersectantes les plus grandes sont de taille $\frac{1}{2}\binom{n}{k} = \binom{n-1}{k-1}$ et sont composées de la manière suivante : on considère tous les couples composés d'une k-partie A et de son complémentaire $N \backslash A$ et l'on choisit l'une des deux k-parties pour appartenir à la famille. Toutefois, pour

$n > 2k$ les familles qui contiennent un élément fixé sont en fait les seules. Le lecteur est invité à essayer de le démontrer lui-même.

Venons-en enfin au troisième résultat qui est probablement le théorème fondamental le plus important de la théorie des ensembles finis, le « théorème du mariage » que Philip Hall a démontré en 1935. Il ouvre la porte à ce qui est appelé aujourd'hui la théorie du couplage et à une grande variété d'applications. Nous traiterons certaines d'entre elles tout au long de cette partie.

Considérons un ensemble fini X et une collection $A_1, \ldots, A_n$ de sous-ensembles de X (non nécessairement distincts). Nous dirons qu'une suite $x_1, \ldots, x_n$ est un *système de représentants distincts* (SRD en abrégé) de $\{A_1, \ldots, A_n\}$ si les x_i sont des éléments distincts de X et si $x_i \in A_i$ pour tout i. Bien sûr un tel système peut très bien ne pas exister. C'est le cas, par exemple, lorsqu'un des ensembles A_i est vide. Le théorème de Hall fournit une condition pour qu'un SRD existe.

« Un mariage collectif ».

Avant d'énoncer le résultat, voici l'interprétation humaine qui lui a valu le nom folklorique de *théorème du mariage* : considérons un ensemble de filles $\{1, \ldots, n\}$ et un ensemble de garçons X. Chaque fois que $x \in A_i$, alors une fille i et un garçon x sont enclins à se marier ; A_i est simplement l'ensemble des alliances possibles de la fille i. Un SRD représente ainsi un mariage collectif où chaque fille épouse un garçon qu'elle aime.

Revenons aux ensembles, voici le théorème.

Théorème 3. *Soit $A_1, \ldots, A_n$ une collection de sous-ensembles d'un ensemble fini X. Alors il existe un système de représentants distincts si et seulement si la réunion de m sous-ensembles quelconques A_i contient au moins m éléments, pour $1 \leqslant m \leqslant n$.*

Il est clair que la condition est nécessaire : si m ensembles A_i contiennent à eux tous moins de m éléments, alors ces m ensembles ne peuvent certainement pas être représentés par des éléments distincts. Le fait surprenant (qui confère au théorème son universalité) est que cette condition évidente est aussi suffisante. La preuve originale de Hall était plutôt compliquée ; plusieurs nouvelles preuves différentes furent proposées depuis. Celle qui semble la plus naturelle est due à Easterfield ; elle a été redécouverte par Halmos et Vaughan.

■ **Preuve.** On procède par récurrence sur n. Pour $n = 1$, il n'y a rien à prouver. Soit $n > 1$; supposons que $\{A_1, \ldots, A_n\}$ vérifie la condition du théorème (que l'on appelera (H) dans la suite). On dit qu'une collection de ℓ ensembles A_i, $1 \leqslant \ell < n$, est une *famille critique* si sa réunion a un cardinal égal à ℓ. Deux cas se présentent.

Cas 1: Il n'y a pas de famille critique.

Choisissons un élément $x \in A_n$. Supprimons x de X et considérons la collection $A'_1, \ldots, A'_{n-1}$ avec $A'_i = A_i \backslash \{x\}$. Puisqu'il n'y a pas de famille

critique, la réunion de m ensembles quelconques A'_i contient au moins m éléments. Un simple raisonnement par récurrence sur n montre qu'il existe un SRD $x_1, \ldots, x_{n-1}$ de $\{A'_1, \ldots, A'_{n-1}\}$. En y adjoignant $x_n = x$, cela fournit un SRD pour la collection originale.

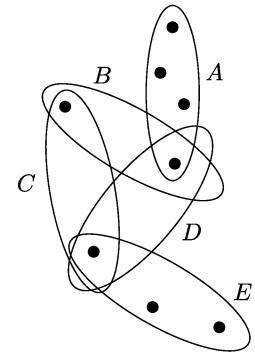
Cas 2: Il existe au moins une famille critique.

Après éventuelle réindexation, on peut supposer que $\{A_1, \ldots, A_\ell\}$ est une famille critique. Alors $\bigcup_{i=1}^{\ell} A_i = \widetilde{X}$ avec $|\widetilde{X}| = \ell$. Puisque $\ell < n$, on déduit par récurrence l'existence d'un SRD pour $A_1, \ldots, A_\ell$, c'est-à-dire qu'il existe une façon d'indexer $x_1, \ldots, x_\ell$ de $\widetilde{X}$ telle que $x_i \in A_i$ pour tout $i \leqslant \ell$.

Considérons maintenant la collection restante $A_{\ell+1}, \ldots, A_n$ et prenons m de ces ensembles. Puisque la réunion de $A_1, \ldots, A_\ell$ et de ces m ensembles contient au plus $\ell + m$ éléments par la condition (H), les m ensembles contiennent au plus m éléments hors de $\widetilde{X}$. En d'autres termes, la condition (H) est vérifiée pour la famille :

$$A_{\ell+1}\backslash\widetilde{X},\ \ldots\ , A_n\backslash\widetilde{X}$$

La récurrence fournit à présent un SRD pour $A_{\ell+1}, \ldots, A_n$ qui évite $\widetilde{X}$. En la combinantavec $x_1, \ldots, x_\ell$ on obtient un SRD pour tous les ensembles A_i.



$\{B, C, D\}$ est une famille critique.

Comme nous l'avons indiqué, le théorème de Hall a été à l'origine du vaste domaine de la théorie du couplage [6]. Parmi les nombreuses variantes et ramifications, énonçons un résultat particulièrement attractif que le lecteur est invité à démontrer lui-même :

> *Supposons que les ensembles $A_1, \ldots, A_n$ aient tous un cardinal $k \geqslant 1$ et supposons qu'aucun élément ne soit contenu dans plus de k ensembles. Alors il existe k SRD tels que pour tout i, les k représentants de A_i soient distincts et forment ainsi l'ensemble A_i.*

Un joli résultat qui devrait ouvrir de nouveaux horizons sur les possibilités de mariage.

Bibliographie

[1] T. E. EASTERFIELD : *A combinatorial algorithm,* J. London Math. Soc. **21** (1946), 219-226.

[2] P. ERDŐS, C. KO & R. RADO : *Intersection theorems for systems of finite sets,* Quart. J. Math. (Oxford), Ser. (2) **12** (1961), 313-320.

[3] P. HALL : *On representatives of subsets,* J. London Math. Soc. **10** (1935), 26-30.

[4] P. R. HALMOS & H. E. VAUGHAN : *The marriage problem,* Amer. J. Math. **72** (1950), 214-215.

[5] G. KATONA : *A simple proof of the Erdős-Ko-Rado theorem,* J. Combinatorial Theory, Ser. B **13** (1972), 183-184.

[6] L. LOVÁSZ & M. D. PLUMMER : *Matching Theory,* Akadémiai Kiadó, Budapest 1986.

[7] D. LUBELL : *A short proof of Sperner's theorem,* J. Combinatorial Theory **1** (1966), 299.

[8] E. SPERNER : *Ein Satz über Untermengen einer endlichen Menge,* Math. Zeitschrift **27** (1928), 544-548.

Mélanger un jeu de cartes

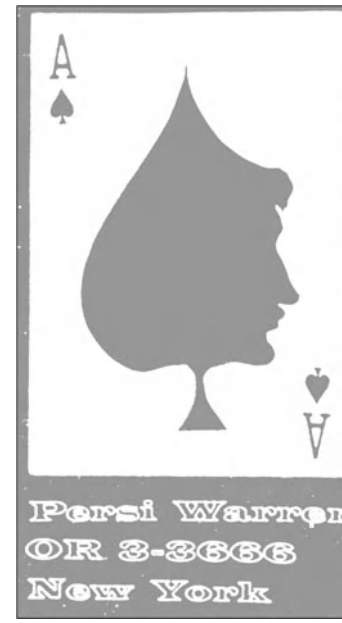
Chapitre 28

Combien de fois faut-il mélanger un jeu de cartes pour que l'on puisse considérer que les cartes sont dans un ordre aléatoire ?

L'étude de processus aléatoires est une chose courante dans la vie quotidienne (« Combien de temps faut-il pour se rendre à l'aéroport aux heures de pointe ? ») comme en mathématiques. Bien évidemment, si l'on pose les questions de manière pertinente on favorise grandement l'obtention de réponses pertinentes à ce type de problèmes. Dans l'exemple que nous nous proposons d'étudier, en l'occurrence comment bien mélanger un jeu de cartes, cela consiste :

- à spécifier le nombre de cartes composant le jeu (disons $n = 52$ cartes) ;
- à préciser le procédé mis en œuvre pour mélanger (nous analyserons une méthode qui consiste à placer au hasard dans le jeu la carte supérieure du tas avant de considérer une méthode plus réaliste et plus efficace consistant à mélanger le jeu à l'américaine) ;
- à dire précisément ce que l'on entend par « aléatoire » ou « presque aléatoire »

Ainsi, notre but dans ce chapitre est-il d'analyser le processus consistant à mélanger un jeu de cartes en se fondant sur une méthode due à Edgar N. Gilbert et Claude Shannon (1955, inédit), reprise par Jim Reeds (1981, inédit) et qui est présentée dans une étude du statisticien David Aldous et du magicien devenu mathématicien Persi Diaconis [1]. Nous n'allons pas démontrer le résultat le plus fin selon lequel il suffit de mélanger les cartes 7 fois à l'américaine pour obtenir un jeu bien mélangé (c'est-à-dire dans lequel l'ordre des cartes est presqu'aléatoire) mais nous allons démontrer que 12 est un majorant. Nous allons, chemin faisant, découvrir quelques idées extrêmement brillantes : le concept de *critère d'arrêt*, celui de *temps d'arrêt*, un lemme qui affirme que le temps d'arrêt majore la mesure de variation totale, le lemme d'inversion de Reed ; ceci nous amènera à interpréter le fait de mélanger des cartes comme un « tri inversé ». La question se réduira finalement à deux problèmes combinatoires classiques : le problème de la collection de vignettes et le paradoxe des anniversaires. Commençons donc par étudier ces deux problèmes.



La carte de visite de Persi Diaconis alors qu'il exerçait la profession de magicien. Répondant à une interview, il avait dit : « Si vous dites que vous êtes professeur à Stanford, les gens vous considèrent avec respect. Si vous dites que vous préparez des tours de magie, ils refusent de vous présenter à leur fille ».

Le paradoxe des anniversaires

On considère une assemblée de n personnes prises au hasard — par exemple les auditeurs d'un cours ou d'un séminaire. Quelle est la probabilité pour que les dates anniversaires de toutes ces personnes soient distinctes ? Avec les hypothèses simplificatrices habituelles (365 jours dans l'année, pas d'effet saisonnier, pas de jumeaux dans l'assemblée) la probabilité est :

$$p(n) \;=\; \prod_{i=1}^{n-1} \Big(1 - \frac{i}{365}\Big),$$

valeur qui est inférieure à $\frac{1}{2}$ dès que n est supérieur ou égal à 23 (c'est ce que l'on a coutume d'appeler « le paradoxe des anniversaires »). Elle est inférieur à $0,09$ pour $n = 42$ et vaut exactement 0 pour $n > 365$ (conformément au « principe des tiroirs », voir chapitre 25). La formule précédente est facile à établir : si les i premières personnes considérées ont des dates anniversaires distinctes, alors la probabilité pour que la $(i+1)$-ième personne ait encore une date anniversaire distincte des dates du groupe des i premières personnes est $1 - \frac{i}{365}$ puisqu'il reste $365 - i$ dates libres.

De manière analogue, si n boules sont placées indépendamment et aléatoirement dans K boîtes, alors la probabilité qu'aucune boîte ne contienne plus d'une boule est :

$$p(n, K) \;=\; \prod_{i=1}^{n-1} \Big(1 - \frac{i}{K}\Big).$$

La collection de vignettes

Les enfants achètent des vignettes représentant des chanteurs célèbres (ou des joueurs de football) pour compléter des albums. Les vignettes sont vendues dans des sachets opaques de sorte qu'on ne peut pas savoir au moment de l'achat quelles vignettes on va obtenir (on suppose ici pour simplifier qu'un sachet contient une vignette). Si la collection comporte n vignettes différentes, combien faut-il acheter de vignettes pour espérer avoir complété la collection ?

De manière analogue, si l'on tire aléatoirement des boules dans une urne contenant n boules indiscernables, et que l'on replace la boule dans l'urne après le tirage (tirage avec remise) et que l'on mélange le contenu de l'urne, combien de tirages faut-il effectuer en moyenne pour être certain d'avoir extrait de l'urne chacune des boules au moins une fois ?

Si l'on a déjà réussi à extraire k boules différentes de l'urne, alors la probabilité pour qu'apparaisse au tirage suivant une boule déjà présente dans la liste des k premières est $\frac{k}{n}$. Ainsi, la probabilité de devoir recourir à exactement s tirages avant d'obtenir une boule distincte des k déjà rencontrées est

$(\frac{k}{n})^{s-1}(1-\frac{k}{n})$. Le nombre moyen de tirages que l'on doit effectuer avant d'espérer obtenir une nouvelle boule est donc :

$$\sum_{s\geqslant 1}\Big(\frac{k}{n}\Big)^{s-1}\Big(1-\frac{k}{n}\Big)s \;=\; \frac{1}{1-\frac{k}{n}},$$

le calcul étant effectué en s'aidant du résultat évoqué ci-contre dans la marge. Ainsi, le nombre moyen de tirages nécessaires pour avoir rencontré *chacune* des boules au moins une fois est :

$$\sum_{k=0}^{n-1}\frac{1}{1-\frac{k}{n}} \;=\; \frac{n}{n}+\frac{n}{n-1}+\cdots+\frac{n}{2}+\frac{n}{1} \;=\; nH_n \;\approx\; n\ln(n),$$

en s'appuyant sur l'estimation du n-ième nombre harmonique H_n établie en page 11. La réponse au problème de la collection de vignettes est donc qu'il faut s'attendre à devoir faire à peu près $n\ln(n)$ achats de vignettes.

La majoration que nous allons maintenant établir concerne les cas où nous aurons besoin de recourir à un nombre de tirages nettement plus élevé que $n\ln(n)$ pour parvenir à nos fins. Si l'on désigne par V_n le nombre aléatoire de tirages effectués jusqu'à obtenir, pour la première fois, au moins une fois chaque boule (on vient de voir que l'espérance de V_n vérifie $E[V_n]\approx n\ln(n)$), alors pour $n\geqslant 1$ et $c\geqslant 0$, la probabilité pour que l'on ait besoin de plus de $m:=\lceil n\ln(n)+cn\rceil$ tirages vérifie :

$$\mathrm{Prob}\big[V_n>m\big] \;\leqslant\; e^{-c}.$$

En effet, si A_i désigne l'événement « la boule i n'est pas sortie lors des m premiers tirages », alors :

$$\begin{aligned}\mathrm{Prob}\big[V_n>m\big] \;&=\; \mathrm{Prob}\Big[\bigcup_i A_i\Big] \;\leqslant\; \sum_i \mathrm{Prob}\big[A_i\big]\\ &=\; n\Big(1-\frac{1}{n}\Big)^m \;<\; ne^{-m/n} \;\leqslant\; e^{-c}.\end{aligned}$$

Considérons à présent un jeu de n cartes. On peut indexer les cartes de 1 à n en fonction de leur ordre d'apparition, de sorte que la carte portant le numéro 1 est celle qui se trouve en haut de la pile alors que celle portant le numéro n se trouve au bas de la pile. Nous notons désormais $\mathfrak{S}_n$ le groupe symétrique d'ordre n, c'est-à-dire l'ensemble des permutations de $\{1,\ldots,n\}$. *Mélanger* le jeu de cartes revient à appliquer *aléatoirement* un certain nombre de *permutations* à l'ordre des cartes. De manière idéale, cela sous-entend que chacune des permutations $\pi\in\mathfrak{S}_n$ appliquée à la séquence initiale $(1,2,\ldots,n)$ à la même probabilité $\frac{1}{n!}$ d'être appliquée. Ainsi, l'application d'une permutation transforme l'ordre des cartes en une nouvelle séquence $\pi=(\pi(1),\pi(2),\ldots,\pi(n))$ parfaitement aléatoire. Toutefois, ce n'est pas ce qui se produit dans la pratique. En effet, lorsque l'on mélange un jeu de cartes, ce sont seulement « certaines » permutations qui se produisent, et pas toujours avec la même probabilité. On mélange donc plusieurs fois le jeu et on espère qu'à l'issue de ces opérations le jeu est au moins rangé dans un ordre que l'on peut considérer comme « proche » d'un ordre aléatoire.

$$\begin{aligned}&\sum_{s\geqslant 1}x^{s-1}(1-x)s \;=\\ &\quad=\; \sum_{s\geqslant 1}x^{s-1}s \;-\; \sum_{s\geqslant 1}x^s s\\ &\quad=\; \sum_{s\geqslant 0}x^s(s+1) \;-\; \sum_{s\geqslant 0}x^s s\\ &\quad=\; \sum_{s\geqslant 0}x^s \;=\; \frac{1}{1-x},\end{aligned}$$

où la dernière égalité s'obtient par sommation d'une série géométrique (voir page 42).

Un petit calcul montre que la fonction $x\mapsto\big(1-\frac{1}{x}\big)^x$ est croissante sur $]1,+\infty[$ et qu'elle tend vers $1/e$ en $+\infty$. Il en résulte que $\big(1-\frac{1}{n}\big)^n<\frac{1}{e}$ pour tout $n\geq 1$.

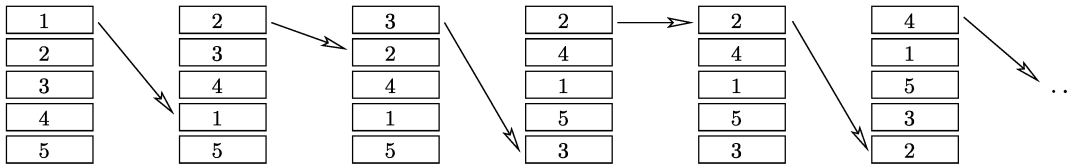
« La carte supérieure est insérée au hasard dans le jeu ».

Mélanges à partir de la carte du haut de la pile

On peut mélanger un jeu de cartes très simplement en partant de la carte du haut de la pile : le jeu étant rangé en tas, on prend la carte supérieure du tas et on l'insère au hasard dans le reste du jeu à l'une des n places possibles, ces places étant équiprobables de probabilité $\frac{1}{n}$. Ainsi, on applique à l'ordre initial des cartes une permutation de la forme :

$$\tau_i \;=\; \big(2, 3, \ldots, \underset{\displaystyle\downarrow}{i}, 1, i{+}1, \ldots, n\big)$$

où $1 \leqslant i \leqslant n$. Après une telle manipulation, l'ordre des cartes ne peut être considéré comme aléatoire et l'on s'attend à ce qu'il soit nécessaire de recourir à un grand nombre de manipulations de ce type pour parvenir à un ordre aléatoire des cartes. Une succession typique de mélanges effectués à partir de l'insertion aléatoire de la carte supérieure peut ressembler à ce qui suit (pour $n = 5$) :



La question est à présent de savoir comment on mesure le fait d'avoir obtenu un ordre « presqu'aléatoire » ? Les probabilistes ont mis au point la « mesure de variation totale » comme une mesure impitoyable du caractère aléatoire. On considère la distribution de probabilité sur les $n!$ ordres possibles du jeu, ou, ce qui revient au même, sur les $n!$ permutations $\sigma \in \mathfrak{S}_n$ conduisant à ces ordres.

Par exemple, la distribution de départ, I, est définie par :

$$\begin{aligned} \mathsf{I}(\mathrm{id}) &= 1, \\ \mathsf{I}(\pi) &= 0, \quad \text{si } \pi \neq \mathrm{id}, \end{aligned}$$

et la distribution uniforme U est définie par :

$$\mathsf{U}(\pi) \;=\; \tfrac{1}{n!}, \quad \text{pour tout } \pi \in \mathfrak{S}_n.$$

La *mesure de variation totale* entre deux distributions de probabilité Q_1 et Q_2 est définie comme :

$$\|\mathsf{Q}_1 - \mathsf{Q}_2\| \;:=\; \tfrac{1}{2} \sum_{\pi \in \mathfrak{S}_n} |\mathsf{Q}_1(\pi) - \mathsf{Q}_2(\pi)|.$$

En posant $\mathsf{Q}_i(S) := \sum_{\pi \in S} \mathsf{Q}_i(\pi)$, on peut montrer que :

$$\|\mathsf{Q}_1 - \mathsf{Q}_2\| \;=\; \max_{S \subseteq \mathfrak{S}_n} |\mathsf{Q}_1(S) - \mathsf{Q}_2(S)|,$$

en tenant compte de $\sum_\pi \mathsf{Q}_1(\pi) = \sum_\pi \mathsf{Q}_2(\pi) = 1$. En effet, la majoration

$$\max_{S \subseteq \mathfrak{S}_n} |\mathsf{Q}_1(S) - \mathsf{Q}_2(S)| \leqslant \|\mathsf{Q}_1 - \mathsf{Q}_2\|$$

est facile à établir et l'égalité s'obtient en considérant le cas particulier $S := \{\pi \in \mathfrak{S}_n : \mathsf{Q}_1(\pi) > \mathsf{Q}_2(\pi)\}$. Il est manifeste que $0 \leqslant \|\mathsf{Q}_1 - \mathsf{Q}_2\| \leqslant 1$. Dans la suite, « être presqu'aléatoire » sera interprété comme « être à petite distance (à petite mesure de variation totale) de la distribution uniforme ». Ici, la distance séparant la distribution initiale I et la distribution uniforme U est très proche de 1 :

$$\|\mathsf{I} - \mathsf{U}\| \;=\; 1 - \tfrac{1}{n!}.$$

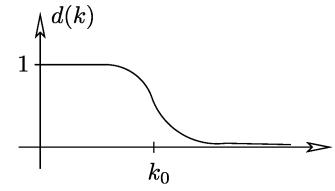
Après insertion aléatoire de la carte supérieure dans le tas, les choses ne s'améliorent pas beaucoup :

$$\|\mathsf{Top} - \mathsf{U}\| \;=\; 1 - \tfrac{1}{(n-1)!}.$$

La distribution de probabilité sur $\mathfrak{S}_n$ que l'on obtient après avoir répété k fois une insertion aléatoire de la carte supérieure sera notée Top^{*k}. La question est de savoir comment se comporte $\|\mathsf{Top}^{*k} - \mathsf{U}\|$ lorsque k augmente, c'est-à-dire au fur et à mesure que l'on mélange le jeu. Que se passe-t-il pour d'autres façons de mélanger ? La théorie générale (en particulier, les chaînes de Markov sur les groupes finis ; voir, par exemple, Behrends [3]) établit que pour des grandes valeurs de k la distance (de variation totale) $d(k) := \|\mathsf{Top}^{*k} - \mathsf{U}\|$ tend vers 0 exponentiellement. Cependant, cette théorie ne rend pas compte du phénomène de « coupure » que l'on observe en pratique : après un certain nombre k_0 d'opérations, $d(k)$ décroît rapidement vers 0. La représentation graphique présentée dans la marge illustre ce phénomène.

Pour les joueurs de cartes, la question n'est pas « à combien précisément se trouve-t-on de la distribution uniforme après avoir mélangé un million de fois ? » mais « est-ce que sept mélanges suffisent ? ».

(Aldous & Diaconis [1])



Critère d'arrêt uniforme fort

L'étonnant critère d'arrêt « uniforme fort » tel qu'il a été introduit par Aldous et Diaconis permet de transcrire les notions essentielles. Imaginons qu'un directeur de casino observe attentivement l'opération consistant à mélanger les cartes et qu'il note les permutations appliquées au jeu à chaque étape de sorte qu'il puisse dire « ON ARRÊTE ! » après un certain nombre d'étapes dépendant des permutations qu'il a vues. Il disposerait ainsi d'un « critère d'arrêt » permettant d'interrompre le mélange des cartes. Un tel critère ne doit dépendre que des opérations effectuées (aléatoirement) sur le jeu. Le critère d'arrêt est dit *uniforme fort* si la condition suivante est satisfaite pour tout $k \geqslant 0$:

> ***Si** le processus est interrompu après exactement k étapes, **alors** les permutations associées à l'ordre potentiel des cartes (résultat des k étapes) ont une distribution (exactement) uniforme.*

Soit T le nombre d'étapes exécutées jusqu'à ce que le critère d'arrêt se déclenche et soit X_k la permutation de $\mathfrak{S}_n$ associée à l'ordre dans lequel

Probabilités conditionnelles

La *probabilité conditionnelle*

$$\mathrm{Prob}[A \mid B]$$

désigne la probabilité de l'événement A sous la condition que l'événement B se réalise. C'est aussi la probabilité que les deux événements se réalisent divisée par la probabilité que B se réalise :

$$\mathrm{Prob}[A \mid B] \;=\; \frac{\mathrm{Prob}[A \wedge B]}{\mathrm{Prob}[B]}.$$

se trouve le jeu de cartes après la k-ième opération de mélange. T et X_k sont des variables aléatoires et le critère d'arrêt est uniforme fort si pour toutes les valeurs possibles de k :

$$\text{Prob}\big[X_k = \pi \mid T = k\big] \;=\; \frac{1}{n!} \quad \text{pour tout } \pi \in \mathfrak{S}_n.$$

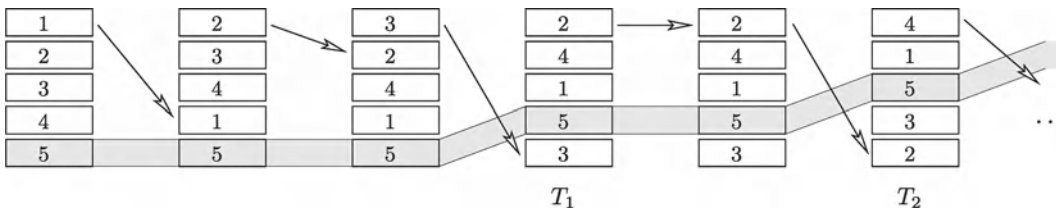
Cette expression est remarquable et utile à trois titres.

1. Elle prouve que des critères d'arrêt uniforme forts existent ; ils peuvent même être très simples dans certains cas.
2. Les cas en question peuvent être analysés : Prob$[T > k]$ conduit souvent à des problèmes combinatoires simples.
3. Le résultat précédent conduit à des bornes supérieures sur des distances du type $d(k) = \|\mathsf{Top}^{*k} - \mathsf{U}\|$.

Par exemple, le procédé consistant à insérer la carte supérieure dans le jeu admet comme critère d'arrêt uniforme fort :

> « ARRÊTER dès que la carte du bas de la pile (celle numérotée n) est à son tour insérée dans le tas pour la première fois »

En effet, si l'on suit la position de la n-ième carte au fur et à mesure des opérations :



on constate que tout au long du processus, l'ordre des cartes situées au-dessous de cette carte obéit à une loi uniforme. Ainsi, lorsque la n-ième carte a atteint le sommet de la pile et qu'elle va être insérée au hasard dans le jeu, tout l'ordre de tout le tas obéit à une loi uniforme. On ne connaît pas précisément le moment où cet événement se produit (seul le directeur du casino le sait).

Désignons par T_i la variable aléatoire qui décompte le nombre d'opérations de mélange effectuées avant que pour la première fois i cartes reposent sous la carte numéro n (celle qui était initialement au bas de la pile). Il nous faut calculer la distribution de :

$$T \;=\; T_n \;=\; T_1 + (T_2 - T_1) + \ldots + (T_{n-1} - T_{n-2}) + (T_n - T_{n-1}).$$

Chaque terme de la somme correspond à un problème de collection de vignettes : $T_i - T_{i-1}$ indique le nombre d'opérations qui ont été effectuées avant que la carte du haut de la pile n'atteigne l'une des i places possibles se trouvant en-dessous de la carte numéro n. C'est donc aussi le nombre d'achats effectués pour que la collection de vignettes passe de la $(n-i)$-ième image à la $(n-i+1)$-ième. Soit V_i le nombres de vignettes achetées jusqu'à obtention de i vignettes distinctes. Alors :

$$V_n \;=\; V_1 + (V_2 - V_1) + \ldots + (V_{n-1} - V_{n-2}) + (V_n - V_{n-1}),$$

et nous avons vu que $\text{Prob}[T_i - T_{i-1} = j] = \text{Prob}[V_{n-i+1} - V_{n-i} = j]$ pour tout i et tout j. Ainsi, le problème de la collection de vignettes et l'opération consistant à mélanger un jeu de cartes (en plaçant la carte supérieure au hasard dans le jeu) conduisent aux mêmes séquences de processus aléatoires indépendants (ils diffèrent en ce qu'ils se produisent en ordre inverse l'un de l'autre). Le déclenchement du critère d'arrêt uniforme fort pour mélanger les cartes ne nécessitera donc plus de $k = \lceil n\ln(n) + cn \rceil$ étapes qu'avec la probabilité :

$$\text{Prob}\big[T > k\big] \;\leqslant\; e^{-c}.$$

Cela signifie qu'après $k = \lceil n\ln(n) + cn \rceil$ étapes, le jeu de cartes est « bien mélangé » puisqu'il vérifie :

$$d(k) \;=\; \|\mathsf{Top}^{*k} - \mathsf{U}\| \;\leqslant\; e^{-c},$$

d'après le lemme, simple mais essentiel, qui suit.

Lemme. *Soit* $\mathsf{Q} : \mathfrak{S}_n \longrightarrow \mathbb{R}$ *une distribution de probabilité quelconque définissant un procédé* Q^{*k} *pour mélanger les cartes avec un critère d'arrêt uniforme fort de temps d'arrêt* T*. Alors, pour tout* $k \geqslant 0$*,*

$$\|\mathsf{Q}^{*k} - \mathsf{U}\| \;\leqslant\; \textit{Prob}\big[T > k\big].$$

■ **Preuve.** Si X est une variable aléatoire à valeurs dans $\mathfrak{S}_n$, dont la distribution de probabilité est Q, on écrit $\mathsf{Q}(S)$ pour désigner la probabilité que X prenne une valeur dans $S \subseteq \mathfrak{S}_n$. En d'autres termes, $\mathsf{Q}(S) = \text{Prob}[X \in S]$, et dans le cas de la distribution uniforme $\mathsf{Q} = \mathsf{U}$ on trouve :

$$\mathsf{U}(S) \;=\; \text{Prob}\big[X \in S\big] \;=\; \frac{|S|}{n!}.$$

Pour chaque sous-ensemble $S \subseteq \mathfrak{S}_n$, on obtient la probabilité que le jeu de cartes soit ordonné selon une certaine permutation S de la manière suivante :

$$\begin{aligned}
\mathsf{Q}^{*k}(S) \;&=\; \text{Prob}[X_k \in S] \\
&=\; \sum_{j\leqslant k} \text{Prob}[X_k \in S \;\wedge\; T = j] \;+\; \text{Prob}[X_k \in S \;\wedge\; T > k] \\
&=\; \sum_{j\leqslant k} \mathsf{U}(S)\,\text{Prob}[T = j] \;+\; \text{Prob}[X_k \in S \,|\, T > k]\cdot\text{Prob}[T > k] \\
&=\; \mathsf{U}(S)\,(1 - \text{Prob}[T > k]) \;+\; \text{Prob}[X_k \in S \,|\, T > k]\cdot\text{Prob}[T > k] \\
&=\; \mathsf{U}(S) \;+\; \big(\text{Prob}[X_k \in S \,|\, T > k] - \mathsf{U}(S)\big)\cdot\text{Prob}[T > k].
\end{aligned}$$

On en tire :

$$|\mathsf{Q}^{*k}(S) - \mathsf{U}(S)| \;\leqslant\; \text{Prob}[T > k]$$

puisque la quantité :

$$\text{Prob}[X_k \in S \,|\, T > k] \;-\; \mathsf{U}(S)$$

est obtenue comme différence de deux probabilités si bien qu'elle est de valeur absolue inférieure à 1. □

À ce stade nous avons terminé l'analyse de la procédure consistant à mélanger un jeu de cartes en insérant la carte supérieure au hasard dans le jeu : nous avons établi une majoration du nombre d'itérations du procédé nécessaires pour que l'ordre du jeu puisse être considéré comme aléatoire.

Théorème 1. *Soit $c \geqslant 0$ et $k := \lceil n\ln(n)+cn \rceil$. Alors après k itérations du procédé de mélange à partir de la carte du haut, la distance (de variation totale) entre le résultat et la distribution uniforme vérifie :*

$$d(k) \;:=\; \|\mathsf{Top}^{*k} - \mathsf{U}\| \;\leqslant\; e^{-c}.$$

On peut aussi montrer que la distance $d(k)$ reste grande si l'on un nombre d'itérations k nettement inférieur à $n\ln(n)$ parce qu'un nombre inférieur de mélanges ne permet pas de défaire l'ordre initial sur les cartes appartenant à la partie inférieure du tas.

Bien sûr, mélanger un jeu de cartes selon le procédé que nous avons étudié dans cette partie, c'est-à-dire en plaçant la carte supérieure au hasard dans le tas, est un procédé extrêmement inefficace. Il faudrait, en effet, répéter le procédé $n\ln(n) \approx 205$ fois pour qu'un jeu de $n = 52$ cartes puisse être considéré comme convenablement mélangé. Nous nous intéressons donc désormais à un procédé pour mélanger les cartes qui soit plus efficace et plus réaliste.

Mélange à l'américaine

« Mélange à l'américaine ».

Mélanger un jeu à l'américaine [1] consiste à séparer le jeu en deux paquets et à entrelacer les cartes des deux paquets selon un motif irrégulier. À cet effet, chaque paquet est tenu par les extrémités à l'aide des pouces à l'intérieur et des majeurs à l'extérieur. Les index sont sur le dessus des paquets et plaquent les paquets sur la table de jeu. Les pouces soulèvent et effeuillent simultanément les deux paquets. Les deux paquets s'entrecroisent. Pour terminer le mélange, les paquets sont poussés l'un vers l'autre pour ne plus former qu'un seul tas.

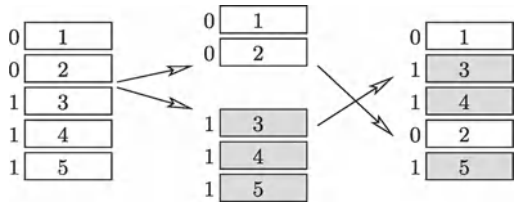
Un mélange à l'américaine effectue un certain nombre de permutations sur les cartes du paquet. On considère que les cartes du paquet sont numérotées de 1 à n, 1 étant le numéro de la carte du haut du paquet. Les mélanges à l'américaine correspondent exactement aux permutations $\pi \in \mathfrak{S}_n$ telles que la séquence :

$$(\pi(1), \pi(2), \ldots, \pi(n))$$

puisse se décomposer en deux sous-séquences croissantes entrelacées (sauf dans le cas de la permutation identité où l'on n'a qu'une séquence). Il y a

1. N.d.T. : le « mélange à l'américaine », appelé *riffle shuffle* en anglais (littéralement « mélange mitraillette »), probablement à cause du bruit que font les cartes pendant l'opération, est aussi connu en français sous le nom de « mélange à la queue d'aronde ».

exactement $2^n - n$ façons de procéder à un mélange à l'américaine pour un jeu de n cartes.



En fait, si le paquet est coupé de sorte que les t cartes de dessus sont tenues dans la main droite ($0 \leqslant t \leqslant n$) et les $n - t$ cartes de l'autre paquet sont tenues dans la main gauche, alors il y a $\binom{n}{t}$ façons d'entrelacer les deux paquets, chacune d'elle engendrant une permutation distincte — à ceci près que pour chaque valeur de t il est possible que l'on obtienne la permutation identité.

À ce stade, il n'est pas évident de savoir quelle distribution de probabilité il convient de plaquer sur le mélange à l'américaine. Il n'y a pas de réponse unique dans la mesure où, par exemple, les joueurs amateurs ne vont pas procéder de la même façon que des croupiers professionnels. Cependant, le modèle suivant, proposé en premier par Edgar N. Gilbert and Claude Shannon en 1955 (à l'époque membre du légendaire département de « mathématiques des communications » des Bell Labs) présente plusieurs avantages :

- il est élégant, simple et paraît naturel ;
- il modélise très bien la manière dont un amateur va procéder à un mélange à l'américaine ;
- et nous avons une chance de pouvoir l'analyser.

Voici trois descriptions qui caractérisent la même distribution de probabilité Rif sur $\mathfrak{S}_n$:

1. $\mathsf{Rif} : \mathfrak{S}_n \longrightarrow \mathbb{R}$ est définie par :

$$\mathsf{Rif}(\pi) \;:=\; \begin{cases} \frac{n+1}{2^n} & \text{si } \pi = \mathrm{id}, \\ \frac{1}{2^n} & \text{si } \pi \text{ consiste en deux séquences croissantes}, \\ 0 & \text{sinon.} \end{cases}$$

2. Couper le jeu de sorte que le premier paquet contienne t cartes avec une probabilité $\frac{1}{2^n}\binom{n}{t}$, prendre ce paquet dans la main droite et prendre l'autre paquet dans la main gauche. À présent, considérant que l'on a r cartes dans la main droite et ℓ cartes dans la main gauche, faire tomber la carte du bas du tas tenu en main droite avec une probabilité $\frac{r}{r+\ell}$ et la carte du bas du tas tenu en main gauche avec une probabilité $\frac{\ell}{r+\ell}$. Recommencer.
3. Un *mélange inverse* consiste à prendre un sous-ensemble des cartes du jeu, à les extraire du jeu et à les poser au-dessus du paquet restant, tout en conservant l'ordre relatif de chaque paquet. Un tel réarrangement du jeu est caractérisé par le sous-ensemble choisi, on considère que tous les sous-ensembles sont équiprobables.

Les mélanges à l'américaine inverses correspondent aux permutations telles que la séquence $\pi = (\pi(1), \ldots, \pi(n))$ soit croissante à l'exception d'une « décroissance » (seule la permutation identité ne présente pas de décroissance).

De manière équivalente, on peut affecter de manière aléatoire et indépendante, une étiquette 0 ou 1 à chaque carte avec la probabilité $\frac{1}{2}$ et choisir de déplacer les cartes auxquelles on a affecté un 0 vers le haut du paquet.

On vérifie assez facilement que ces trois façons de procéder conduisent aux mêmes distributions de probabilité. Pour (1) $\Longleftrightarrow$ (3), il suffit d'observer que l'on obtient la permutation identité chaque fois que les cartes affectées d'un 0 sont au-dessus des cartes qui sont affectées d'un 1.

On vient de définir le modèle, il faut maintenant l'analyser. Combien de fois faut-il répéter le mélange à l'américaine pour obtenir un résultat presqu'aléatoire ? Nous n'allons pas établir le résultat optimal mais allons tout de même trouver une réponse assez précise en combinant trois arguments :

(1) on va analyser les mélanges inverses plutôt que les mélanges à l'américaine directs ;
(2) on va établir un critère d'arrêt uniforme fort pour les mélanges inverses ;
(3) on montre que la clé de l'analyse des mélanges inverses est fournie par le paradoxe des anniversaires.

Théorème 2. *Après avoir exécuté k mélanges à l'américaine sur un jeu de n cartes, la distance de variation totale vérifie :*

$$\|\mathsf{Rif}^{*k} - \mathsf{U}\| \leqslant 1 - \prod_{i=1}^{n-1} \left(1 - \frac{i}{2^k}\right).$$

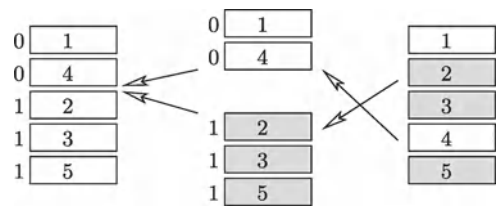
■ **Preuve.**
(1) Il est possible d'analyser les mélanges inverses et on peut essayer de voir à quelle vitesse ils font passer de la distribution initiale à une distribution proche de la distribution uniforme. Les mélanges à l'américaine inverses correspondent à la distribution de probabilité définie par $\overline{\mathsf{Rif}}(\pi) := \mathsf{Rif}(\pi^{-1})$.

Toute permutation admet une unique inverse et $\mathsf{U}(\pi) = \mathsf{U}(\pi^{-1})$, ce qui permet d'écrire :

$$\|\mathsf{Rif}^{*k} - \mathsf{U}\| = \|\overline{\mathsf{Rif}}^{*k} - \mathsf{U}\|.$$

C'est ce qu'on appelle le lemme d'inversion de Reed.

(2) À l'occasion d'un mélange inverse, chaque carte se voit affectée d'un 0 ou d'un 1 :

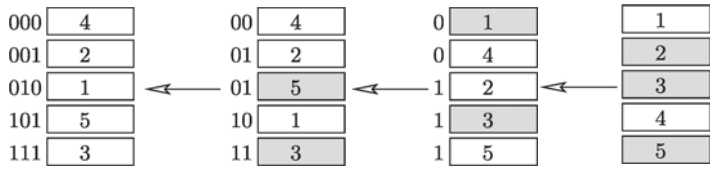


Si l'on peut se souvenir des valeurs affectées à chaque carte — disons, en les écrivant sur la carte —, alors, après k opérations de mélange inverse, chaque carte s'est vu affecter une chaîne de k chiffres (0 ou 1). Nous choisissons alors le critère d'arrêt suivant :

« ARRÊTER dès que toutes les cartes ont des chaînes distinctes. »

Lorsque cela se produit, les cartes du jeu sont *triées* en fonction des nombres binaires $b_k b_{k-1} \dots b_2 b_1$, où b_i est le bit que la carte a collecté au moment du i-ième mélange inverse. Comme ces bits sont parfaitement aléatoires et indépendants, le critère d'arrêt est uniforme fort.

Dans l'exemple suivant, pour $n = 5$ cartes, on a besoin de recourir à $T = 3$ mélanges inverses avant de s'arrêter :



(3) Le temps T requis par le critère d'arrêt est distribué selon le paradoxe des anniversaires pour $K = 2^k$. On place deux cartes dans une même boîte si elles ont la même étiquette $b_k b_{k-1} \dots b_2 b_1 \in \{0,1\}^k$. Il y a donc $K = 2^k$ boîtes et la probabilité pour qu'une boîte reçoive plus d'une carte est :

$$\text{Prob}[T > k] \;=\; 1 - \prod_{i=1}^{n-1} \left(1 - \frac{i}{2^k}\right),$$

et, comme nous l'avons vu, cette quantité majore la mesure de variation totale $\|\mathsf{Rif}^{*k} - \mathsf{U}\| = \|\,\overline{\mathsf{Rif}}^{*k} - \mathsf{U}\|$. □

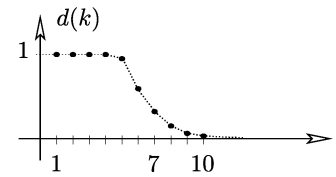
Finalement, combien de fois faut-il mélanger le jeu ? Pour de grandes valeurs de n, on aura besoin de mélanger plus de $k = 2\log_2(n)$ fois. En effet, en posant $k := 2\log_2(cn)$ où $c \geqslant 1$, on trouve (après quelques calculs d'analyse élémentaire) que $P[T > k] \approx 1 - e^{-\frac{1}{2c^2}} \approx \frac{1}{2c^2}$.

Plus précisément, pour $n = 52$ cartes, la majorante du théorème 2 vérifie $d(10) \leqslant 0.73$, $d(12) \leqslant 0.28$, $d(14) \leqslant 0.08$, si bien que $k = 12$ conduit à un processus « suffisamment aléatoire ». Cependant, on ne procède jamais à 12 opérations de mélange « dans la pratique » ! En fait, il n'est pas nécessaire d'en faire autant, comme le montre une analyse plus précise (dont les résultats sont présentés dans la marge). L'analyse des mélanges à l'américaine fait encore aujourd'hui l'objet de débats, notamment pour savoir ce qui constitue une bonne mesure de ce qui peut être considéré comme « suffisamment aléatoire ». Diaconis [4] est une bonne introduction aux récents développements concernant cette question.

Après trois mélanges à l'américaine, un jeu de cartes trié peut paraître bien mélangé mais il ne l'est pas du tout. Martin Gardner [5] (chapitre 7) décrit un certain nombre de tours de cartes étonnants qui exploitent l'ordre caché du jeu de cartes.

k	$d(k)$
1	1.000
2	1.000
3	1.000
4	1.000
5	0.952
6	0.614
7	0.334
8	0.167
9	0.085
10	0.043

La mesure de variation totale après k mélanges à l'américaine, d'après [2].



Bibliographie

[1] D. ALDOUS & P. DIACONIS : *Shuffling cards and stopping times,* Amer. Math. Monthly **93** (1986), 333-348.

[2] D. BAYER & P. DIACONIS : *Trailing the dovetail shuffle to its lair,* Annals Applied Probability **2** (1992), 294-313.

[3] E. BEHRENDS : *Introduction to Markov Chains,* Vieweg, Braunschweig/ Wiesbaden 2000.

[4] P. DIACONIS : *Mathematical developments from the analysis of riffle shuffling,* in : "Groups, Combinatorics and Geometry. Durham 2001" (A. A. Ivanov, M. W. Liebeck and J. Saxl, eds.), World Scientific, Singapore 2003, pp. 73-97.

[5] M. GARDNER : *Mathematical Magic Show,* Knopf, New York/Allen & Unwin, London 1977.

[6] E. N. GILBERT : *Theory of Shuffling,* Technical Memorandum, Bell Laboratories, Murray Hill NJ, 1955.

« Suffisamment aléatoire ? »

Chemins dans les treillis et déterminants

Chapitre 29

L'essence même des mathématiques consiste à démontrer des théorèmes et c'est bien ce que font les mathématiciens : ils démontrent des théorèmes. Toutefois, à dire vrai, ce qu'ils aimeraient vraiment faire une fois dans leur vie, c'est démontrer un *lemme*, comme le lemme de Fatou en analyse, le lemme de Gauss en théorie des nombres, ou le lemme de Burnside-Frobenius en combinatoire.

Mais qu'est-ce qui fait d'une assertion mathématique un lemme célèbre ? D'abord, elle doit être applicable à une grande variété de situations ainsi qu'à des problèmes apparemment sans rapport. Ensuite, le résultat doit être, une fois qu'on l'a compris, tout à fait évident, la réaction du lecteur étant de l'ordre de la jalousie : pourquoi n'ai-je pas remarqué cela avant ? Enfin, sur un plan esthétique, le lemme, et particulièrement sa démonstration, doit être beau !

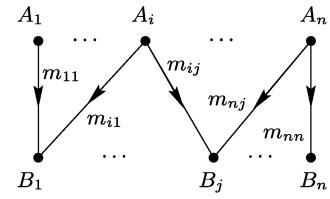
Nous examinons dans ce chapitre une de ces merveilles du raisonnement mathématique, le lemme d'Ira Gessel et Gérard Viennot, devenu classique en énumération combinatoire depuis sa parution en 1985. Une version similaire avait été obtenue auparavant par Bernt Lindström.

Le point de départ se trouve dans la représentation habituelle du déterminant d'une matrice en termes de permutations. Soit $M = (m_{ij})$ une matrice réelle $n \times n$. Alors :

$$\det M \;=\; \sum_{\sigma} \varepsilon(\sigma)\, m_{1\sigma(1)}\, m_{2\sigma(2)} \cdots m_{n\sigma(n)} \qquad (1)$$

où σ parcourt toutes les permutations de l'ensemble $\{1, 2, \ldots, n\}$ et où la signature $\varepsilon(\,\sigma)$ de σ est 1 ou -1 selon que σ est un produit pair ou impair de transpositions .

Passons maintenant aux graphes, plus précisément aux graphes *bipartis orientés et pondérés*. Représentons les lignes de M par les sommets A_1, …, A_n et les colonnes par les sommets B_1, …, B_n. Étant donnée une paire (i, j), on trace un arc de A_i vers B_j et en lui attribuant le poids m_{ij}, comme indiqué sur la figure.



Voici l'interprétation de la formule (1) en termes de graphes :

- Le membre gauche est le déterminant de la *matrice-chemin* M, dont le coefficient (i, j) est le *poids* de *l'unique* chemin orienté de A_i vers B_j.
- Le membre droit est la somme pondérée (signée) sur tous les *systèmes de chemins à sommets disjoints* de $\mathcal{A} = \{A_1, \ldots, A_n\}$ à $\mathcal{B} = \{B_1, \ldots, B_n\}$. Un tel système $\mathcal{P}_\sigma$ est déterminé par les chemins :

$$A_1 \to B_{\sigma(1)},\;\; \ldots \;\;, A_n \to B_{\sigma(n)},$$

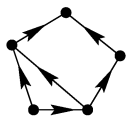
et le *poids* du système de chemins $\mathcal{P}_\sigma$ est le produit des poids de chaque chemin :

$$w(\mathcal{P}_\sigma) \;=\; w(A_1 \to B_{\sigma(1)}) \;\cdots\; w(A_n \to B_{\sigma(n)}).$$

Avec cette interprétation, (1) se lit :

$$\det M \;=\; \sum_\sigma \varepsilon(\sigma)\, w(\mathcal{P}_\sigma).$$

Quel est le résultat de Gessel et Viennot ? C'est la généralisation naturelle de (1) d'un graphe biparti à un graphe quelconque. C'est précisément cette caractéristique qui ouvre un champ d'application aussi large à ce lemme. En outre, sa preuve est prodigieusement simple et élégante.



Un graphe orienté acyclique.

Réunissons d'abord les notions nécessaires. On se donne un graphe fini orienté acyclique $G = (V, E)$, le terme *acyclique* signifiant qu'il n'y a pas de cycles orientés dans G. En particulier, il n'y a qu'un nombre fini de chemins orientés entre deux sommets A et B ; on compte tous les chemins triviaux $A \to A$ de longueur 0. Chaque arête e porte un poids $w(e)$. Si P est un chemin orienté de A vers B, brièvement noté $P : A \to B$, on définit le *poids* de P par :

$$w(P) \;:=\; \prod_{e \in P} w(e),$$

qui vérifie par définition $w(P) = 1$ si P est un chemin de longueur 0.

Soient maintenant $\mathcal{A} = \{A_1, \ldots, A_n\}$ et $\mathcal{B} = \{B_1, \ldots, B_n\}$ deux ensembles de n sommets, $\mathcal{A}$ et $\mathcal{B}$ n'étant pas nécessairement disjoints. Associons à $\mathcal{A}$ et $\mathcal{B}$ la *matrice chemin* $M = (m_{ij})$ définie par :

$$m_{ij} \;:=\; \sum_{P : A_i \to B_j} w(P).$$

Un *système de chemins* $\mathcal{P}$ *de* $\mathcal{A}$ *vers* $\mathcal{B}$ est la donnée d'une permutation σ et de n chemins $P_i : A_i \to B_{\sigma(i)}$, où $i = 1, \ldots, n$. On notera $\varepsilon(\mathcal{P}) = \varepsilon(\sigma)$. Le *poids* de $\mathcal{P}$ est le produit des poids des chemins :

$$w(\mathcal{P}) \;=\; \prod_{i=1}^{n} w(P_i), \tag{2}$$

c'est-à-dire le produit des poids de toutes les arêtes du système de chemins.

Enfin on dit que le système de chemins $\mathcal{P} = (P_1, \ldots, P_n)$ est à *sommets disjoints* si deux chemins de $\mathcal{P}$ n'ont aucun sommet commun.

Lemme. *Soit $G = (V, E)$ un graphe acyclique fini orienté et pondéré, $\mathcal{A} = \{A_1, \ldots, A_n\}$ et $\mathcal{B} = \{B_1, \ldots, B_n\}$ deux ensembles de n sommets, et M la matrice des chemins de $\mathcal{A}$ vers $\mathcal{B}$. Alors :*

$$\det M \;=\; \sum_{\substack{\mathcal{P},\ \text{système de chemins} \\ \text{à sommets disjoints}}} \varepsilon(\mathcal{P})\, w(\mathcal{P}). \tag{3}$$

■ **Preuve.** Le terme $\varepsilon(\sigma)\, m_{1\sigma(1)} \cdots m_{n\sigma(n)}$ dans $\det(M)$ peut s'écrire :

$$\varepsilon(\sigma) \Big(\sum_{P_1 : A_1 \to B_{\sigma(1)}} w(P_1) \Big) \cdots \Big(\sum_{P_n : A_n \to B_{\sigma(n)}} w(P_n) \Big)$$

En sommant sur σ on déduit immédiatement de (2) que :

$$\det M \;=\; \sum_{\mathcal{P}} \varepsilon(\mathcal{P})\, w(\mathcal{P})$$

où $\mathcal{P}$ parcourt *tous* les systèmes de chemins de $\mathcal{A}$ vers $\mathcal{B}$ (à sommets disjoints ou non). Par conséquent, pour arriver à (3), nous devons simplement montrer que :

$$\sum_{\mathcal{P} \in N} \varepsilon(\mathcal{P})\, w(\mathcal{P}) \;=\; 0\,, \tag{4}$$

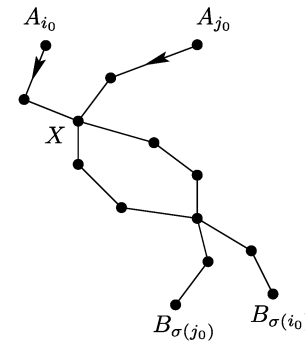
où N est l'ensemble de tous les systèmes de chemins qui ne sont *pas* à sommets disjoints. L'argument est d'une rare beauté : on exhibe une involution $\pi : N \to N$ (sans point fixe) telle que pour $\mathcal{P}$ et $\pi\mathcal{P}$

$$w(\pi\mathcal{P}) \;=\; w(\mathcal{P}) \qquad \text{et} \qquad \varepsilon(\pi\mathcal{P}) \;=\; -\varepsilon(\mathcal{P})$$

Il est clair que ce résultat implique (4) et donc la formule (3) du lemme.

L'involution π est définie très naturellement. Soit $\mathcal{P} \in N$ avec $P_i : A_i \to B_{\sigma(i)}$. Par définition, certaines paires de chemins ont un sommet commun.

- Soit i_0 l'indice minimum tel que P_{i_0} partage des sommets avec un autre chemin.
- Soit X le premier sommet commun de ce type sur le chemin P_{i_0}.
- Soit j_0 l'indice minimum ($j_0 > i_0$) tel que P_{j_0} ait le sommet X en commun avec P_{i_0}.

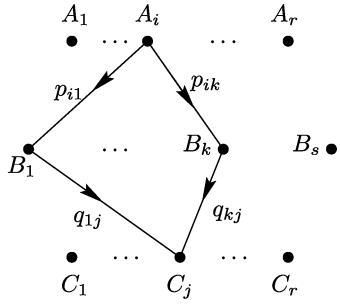


Construisons maintenant le nouveau système $\pi\mathcal{P} = (P'_1, \ldots, P'_n)$ de la manière suivante.

- On pose $P'_k = P_k$ pour tous $k \neq i_0, j_0$.
- Le nouveau chemin P'_{i_0} va de A_{i_0} à X le long de P_{i_0} ; il continue ensuite vers $B_{\sigma(j_0)}$ le long de P_{j_0}. De même, P'_{j_0} va de A_{j_0} à X le long de P_{j_0} et continue vers $B_{\sigma(i_0)}$ le long de P_{i_0}.

Il est clair que $\pi(\pi\mathcal{P}) = \mathcal{P}$, puisque l'indice i_0, le sommet X et l'indice j_0 sont les mêmes qu'auparavant. En d'autres termes, en appliquant π deux fois, on revient aux anciens chemins P_i. Ensuite, puisque $\pi\mathcal{P}$ et $\mathcal{P}$ utilisent exactement les mêmes arêtes, on a $w(\pi\mathcal{P}) = w(\mathcal{P})$. Finalement, puisque la nouvelle permutation σ' est obtenue en multipliant σ par la transposition (i_0, j_0), on a $\varepsilon(\pi\mathcal{P}) = -\varepsilon(\mathcal{P})$. □

Le lemme de Gessel-Viennot permet de retrouver toutes les propriétés fondamentales des déterminants, simplement en étudiant les graphes appropriés. Considérons un exemple particulièrement frappant, la formule de Binet-Cauchy, qui donne une généralisation très utile de la règle du produit pour les déterminants.



Théorème. *Si P est une matrice $(r \times s)$ et Q une matrice $(s \times r)$, $r \leqslant s$, alors :*

$$\det(PQ) \;=\; \sum_{\mathcal{Z}} (\det P_{\mathcal{Z}})(\det Q_{\mathcal{Z}}),$$

où $P_{\mathcal{Z}}$ est la sous-matrice $(r \times r)$ de P dont les colonnes sont définies par $\mathcal{Z}$ et $Q_{\mathcal{Z}}$, la sous-matrice $(r \times r)$ de Q constituée des lignes correspondantes $\mathcal{Z}$.

■ **Preuve.** Comme précédemment, faisons correspondre à P le graphe biparti sur $\mathcal{A}$ et $\mathcal{B}$, et, de même, faisons correspondre à Q le graphe biparti sur $\mathcal{B}$ et $\mathcal{C}$. Considérons maintenant le graphe concaténé représenté sur la figure de gauche ; observons que le coefficient m_{ij} de la matrice des chemins M de $\mathcal{A}$ vers $\mathcal{C}$ est exactement $m_{ij} = \sum_k p_{ik} q_{kj}$. On a donc $M = PQ$.
Puisque les systèmes de chemins à sommets disjoints de $\mathcal{A}$ vers $\mathcal{C}$ dans le graphe concaténé correspondent aux paires de systèmes de $\mathcal{A}$ vers $\mathcal{Z}$ et de $\mathcal{Z}$ vers $\mathcal{C}$, le résultat se déduit immédiatement du lemme, en remarquant que $\varepsilon(\sigma\tau) = \varepsilon(\sigma).\varepsilon(\tau)$. □

Le lemme de Gessel-Viennot est aussi la source d'un grand nombre de résultats qui relient des déterminants à des propriétés énumératives. La recette est toujours la même : on interprète la matrice M comme une matrice de chemins et on essaie de calculer le membre droit de (3). En guise d'illustration, nous allons considérer le problème initialement étudié par Gessel et Viennot, qui les a conduits à leur lemme :

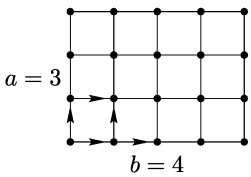
> *Soient $a_1 < a_2 < \ldots < a_n$ et $b_1 < b_2 < \ldots < b_n$ deux ensembles d'entiers naturels. On veut calculer le déterminant de la matrice $M = (m_{ij})$, où m_{ij} est le coefficient binomial $\binom{a_i}{b_j}$.*

En d'autres termes, Gessel et Viennot étudiaient les déterminants de matrices carrées arbitraires extraites du triangle de Pascal. Par exemple :

$$\det \begin{pmatrix} \binom{3}{1} & \binom{3}{3} & \binom{3}{4} \\ \binom{4}{1} & \binom{4}{3} & \binom{4}{4} \\ \binom{6}{1} & \binom{6}{3} & \binom{6}{4} \end{pmatrix} \;=\; \det \begin{pmatrix} 3 & 1 & 0 \\ 4 & 4 & 1 \\ 6 & 20 & 15 \end{pmatrix}$$

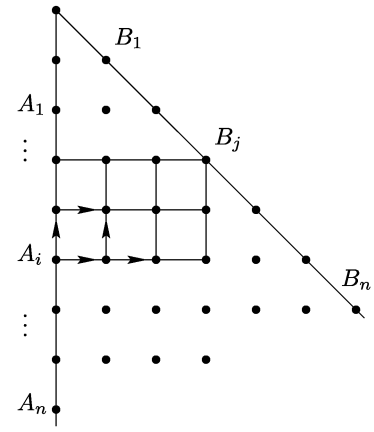
est défini par les coefficients du triangle de Pascal écrits en gras dans la marge.

Avant de donner la solution du problème, rappelons un résultat bien connu qui relie les coefficients binomiaux aux chemins dans les treillis. Considérons un treillis $a \times b$ comme celui représenté ci-contre. Alors, si les seuls pas autorisés pour les chemins sont ceux qui vont vers le haut (nord) et vers la droite (est), le nombre de chemins allant du coin inférieur gauche au coin supérieur droit est $\binom{a+b}{a}$.



La démonstration de ce résultat est facile : chaque chemin est composé d'une suite arbitraire de b pas vers l'est et a pas vers le nord ; il peut donc être codé par une suite de la forme NENEEEN, composée de $a+b$ lettres, a N et b E. Le nombre de telles chaînes est le nombre de façons de choisir a positions de lettres N sur un total de $a+b$ positions, c'est donc $\binom{a+b}{a} = \binom{a+b}{b}$.

Examinons maintenant la figure de représentée ci-contre,où A_i est le point $(0, -a_i)$ et B_j le point $(b_j, -b_j)$.



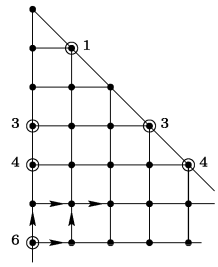
Le nombre de chemins qui vont de A_i à B_j sur cette grille (et qui n'utilisent que des pas vers le nord et vers l'est) est, d'après ce que nous venons de voir $\binom{b_j+(a_i-b_j)}{b_j} = \binom{a_i}{b_j}$. En d'autres termes, la matrice de coefficients binomiaux M est exactement la matrice des chemins de $\mathcal{A}$ à $\mathcal{B}$ dans le graphe de treillis orienté dont toutes les arêtes ont un poids égal à 1, et dont toutes les arêtes sont orientées vers le nord ou vers l'est. Par conséquent, pour calculer $\det M$ nous pouvons appliquer le lemme de Gessel-Viennot. Un peu d'attention montre que chaque système de chemins à sommets disjoints $\mathcal{P}$ de $\mathcal{A}$ à $\mathcal{B}$ doit être composé de chemins $P_i : A_i \to B_i$ pour tout i. Ainsi, la seule permutation possible est l'identité, dont le signe est 1, et nous obtenons le beau résultat suivant :

$$\det\left(\binom{a_i}{b_j}\right) \;=\; \#\text{ systèmes de chemins à sommets disjoints de } \mathcal{A} \text{ vers } \mathcal{B}$$

En particulier, cela implique le fait (loin d'être évident) que $\det M$ est toujours positif, puisque le membre droit de l'égalité *dénombre quelque chose*. En fait, on déduit du lemme de Gessel-Viennot que $\det M = 0$ si et seulement si $a_i < b_i$ pour un certain i.

Par exemple,

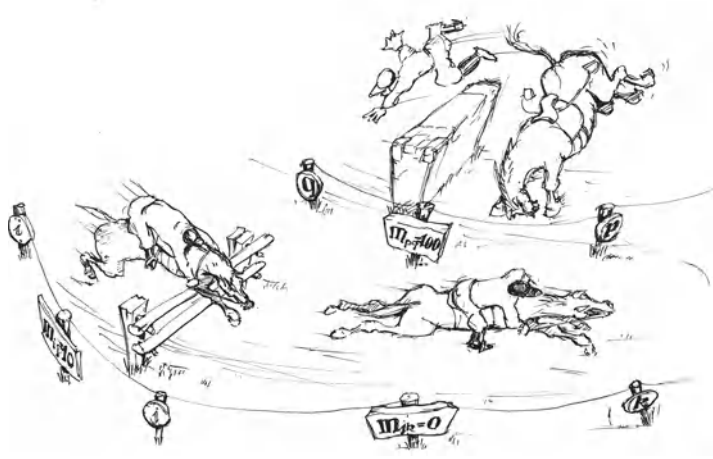
$$\det\begin{pmatrix} \binom{3}{1} & \binom{3}{3} & \binom{3}{4} \\ \binom{4}{1} & \binom{4}{3} & \binom{4}{4} \\ \binom{6}{1} & \binom{6}{3} & \binom{6}{4} \end{pmatrix} = \#\text{ systèmes de chemins à sommets disjoints dans}$$



Bibliographie

[1] I. M. GESSEL & G. VIENNOT : *Binomial determinants, paths, and hook length formulae,* Advances in Math. **58** (1985), 300-321.

[2] B. LINDSTRÖM : *On the vector representation of induced matroids,* Bulletin London Math. Soc. **5** (1973), 85-90.

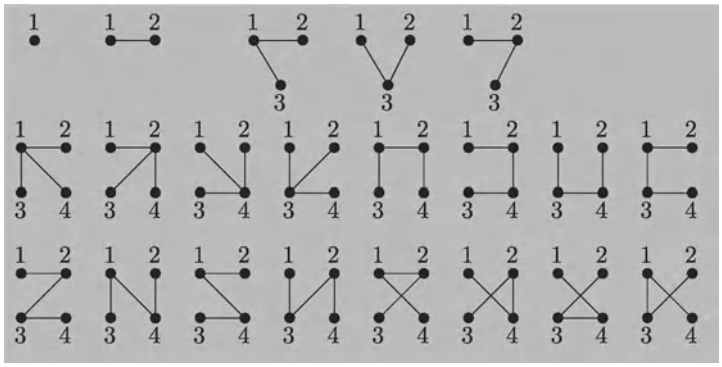


« Chemins dans des treillis ».

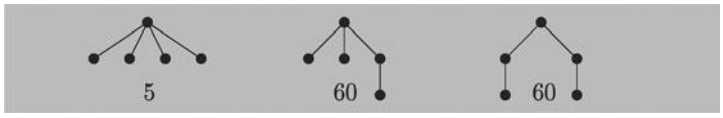
La formule de Cayley pour le nombre d'arbres

Chapitre 30

Une des plus belles formules en combinatoire énumérative est relative au nombre d'arbres étiquetés. Considérons l'ensemble $N = \{1, 2, \ldots, n\}$. Combien d'arbres différents peut-on former sur cet ensemble de sommets ? Soit T_n ce nombre. Le décompte « à la main » donne $T_1 = 1$, $T_2 = 1$, $T_3 = 3$, $T_4 = 16$, avec les arbres représentés dans le tableau suivant :



Il faut remarquer que l'on considère ici des arbres *étiquetés*. Ainsi, par exemple, bien qu'il y ait un seul arbre d'ordre 3 à isomorphisme de graphe près, il y a 3 arbres étiquetés différents (on commence par choisir le sommet intérieur). Pour $n = 5$, il y a trois arbres non isomorphes :



Il est clair qu'il y a 5 étiquetages différents pour le premier arbre et qu'il y a $\frac{5!}{2} = 60$ étiquetages pour le second et le troisième. Nous obtenons donc $T_5 = 125$. Cela suffit pour conjecturer que $T_n = n^{n-2}$; c'est précisément le résultat établi par Cayley.

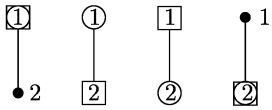
Théorème. *Il y a n^{n-2} arbres étiquetés différents sur n sommets.*

Cette belle formule a donné lieu à des démonstrations aussi belles les unes que les autres, qui reposent sur des techniques diverses allant de la combinatoires à l'algèbre. Nous esquisserons trois d'entre elles avant de présenter celle qui est à ce jour la plus belle de toutes.

Arthur Cayley

■ **Première preuve (Bijection).** La méthode classique, qui est aussi la plus directe, consiste à trouver une bijection de l'ensemble de tous les arbres à n sommets sur un autre ensemble de cardinal connu égal à n^{n-2}. Naturellement, on pense immédiatement à l'ensemble de tous les $(n-2)$-uplets $(a_1, \ldots, a_{n-2})$ tels que $1 \leqslant a_i \leqslant n$. On veut donc coder de manière unique chaque arbre T par un $(n-2)$-uplet $(a_1 \ldots, a_{n-2})$. Ce code, trouvé par Prüfer, figure dans la plupart des livres de théorie des graphes.

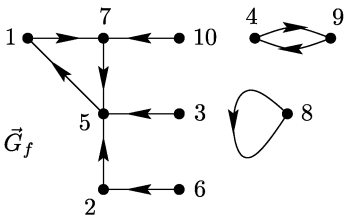
Nous présentons ici une autre preuve due à Joyal, qui construit une bijection, moins connue mais d'une élégance et d'une simplicité similaires. À cet effet, nous ne considérons pas seulement des arbres t sur $N = \{1, \ldots, n\}$ mais des arbres auxquels on adjoint deux sommets distincts, l'*extrémité de gauche* ◯ et l'*extrémité de droite* □, qui coïncident éventuellement. Soit $\mathcal{T}_n = \{(t; ◯, □)\}$ ce nouvel ensemble ; il est alors clair que $|\mathcal{T}_n| = n^2 T_n$.



Les quatre arbres de $\mathcal{T}_2$.

Notre but est donc de montrer que $|\mathcal{T}_n| = n^n$. Or n^n est le cardinal de l'ensemble N^N de toutes les applications de N dans N. La formule est donc établie si on exhibe une bijection de N^N sur $\mathcal{T}_n$.

Soit $f : N \longrightarrow N$ une application quelconque. Représentons f par un graphe orienté $\vec{G}_f$ en plaçant des arcs de i vers $f(i)$.



Par exemple, l'application :

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 7 & 5 & 5 & 9 & 1 & 2 & 5 & 8 & 4 & 7 \end{pmatrix}$$

est représentée par le graphe orienté figurant dans la marge.

Examinons une composante de $\vec{G}_f$. Puisqu'il y a exactement une arête qui part de chaque sommet, cette composante contient autant de sommets que d'arêtes donc exactement un cycle orienté. Soit $M \subseteq N$ la réunion des ensembles de sommets de ces cycles. Un moment de réflexion permet de voir que M est l'*unique* sous-ensemble maximal de N tel que la restriction de f à M soit une bijection sur M. Écrivons :

$$f|_M = \begin{pmatrix} a & b & \ldots & z \\ f(a) & f(b) & \ldots & f(z) \end{pmatrix}$$

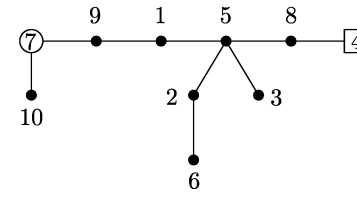
où les nombres $a, b, \ldots, z$ de la première ligne apparaissent dans un ordre naturel. Cela fournit un ordre $f(a), f(b), \ldots, f(z)$ de M déduit de la seconde ligne. $f(a)$ est l'extrémité gauche et $f(z)$ est l'extrémité droite.

On construit à présent l'arbre t correspondant à l'application f de la manière suivante : on place $f(a), \ldots, f(z)$ dans cet ordre comme un *chemin* de $f(a)$ à $f(z)$ et on complète avec les sommets manquants en les plaçant comme dans $\vec{G}_f$ (sans tenir compte de l'orientation c'est-à-dire en supprimant les flèches).

Avec l'exemple précédent, on obtient : $M = \{1, 4, 5, 7, 8, 9\}$,

$$f|_M = \begin{pmatrix} 1 & 4 & 5 & 7 & 8 & 9 \\ 7 & 9 & 1 & 5 & 8 & 4 \end{pmatrix}$$

et donc l'arbre t représenté ci-contre.



On voit immédiatement comment inverser cette correspondance : étant donné un arbre t, on regarde l'unique chemin P de l'extrémité gauche à l'extrémité droite. Cela fournit l'ensemble M et l'application $f|_M$. Les correspondances restantes $i \to f(i)$ sont alors réalisées par les uniques chemins de i à P. □

■ **Deuxième preuve (Algèbre Linéaire).** On peut considérer T_n comme le nombre d'arbres recouvrants du graphe complet K_n. Examinons maintenant un graphe simple connexe arbitraire G sur $V = \{1, 2, \ldots, n\}$; notons $t(G)$ le nombre d'arbres recouvrants minimaux ; alors $T_n = t(K_n)$. Le résultat évoqué qui suit est le *théorème de l'arbre-matrice* de Kirchhoff (voir [1]). Considérons la matrice d'incidence $B = (b_{ie})$ de G, dont les lignes sont indexées par V, les colonnes par E, avec $b_{ie} = 1$ ou 0 suivant que $i \in e$ ou $i \notin e$. Remarquons que $|E| \geqslant n-1$ puisque G est connexe. Remplaçons dans chaque colonne un des deux 1 par -1 de manière arbitraire (ce qui équivaut à orienter G) et appelons C la matrice ainsi obtenue. $M = CC^T$ est alors une matrice $(n \times n)$ symétrique avec les degrés $d_1, \ldots, d_n$ sur la diagonale principale.

Proposition. *On a $t(G) = \det M_{ii}$ pour tout $i = 1, \ldots, n$, où M_{ii} se déduit de M en supprimant la i-ième ligne et la i-ième colonne.*

■ **Preuve.** La clé de la preuve est le théorème de Binet-Cauchy démontré au chapitre précédent : si P est une matrice $(r \times s)$ et Q une matrice $(s \times r)$, $r \leqslant s$, alors $\det(PQ)$ est égal à la somme des produits des déterminants des sous-matrices $(r \times r)$ correspondantes, où « correspondantes » signifie que l'on prend les mêmes indices pour les r colonnes de P et les r lignes de Q.

Pour M_{ii}, cela signifie que :

$$\det M_{ii} \;=\; \sum\nolimits_N \det N \cdot \det N^T \;=\; \sum\nolimits_N (\det N)^2$$

où N parcourt toutes les $(n-1) \times (n-1)$ sous-matrices de $C \backslash \{\text{ligne } i\}$. Les $n-1$ colonnes de N correspondent à un sous-graphe de G ayant $n-1$ arêtes sur n sommets. Il reste à montrer que :

$$\det N = \begin{cases} \pm 1 & \text{si ces arêtes engendrent un arbre} \\ 0 & \text{sinon} \end{cases}$$

Supposons que les $n-1$ arêtes n'engendrent pas d'arbre. Il existe alors une composante qui ne contient pas i. Puisque les lignes correspondantes de cette composante ont une somme nulle, elles sont linéairement dépendantes donc $\det N = 0$.

Supposons maintenant que les colonnes de N engendrent un arbre. Il existe alors un sommet $j_1 \neq i$ de degré 1 ; soit e_1 l'arête incidente. En supprimant j_1, e_1 nous obtenons un arbre avec $n-2$ arêtes. Il y a encore un sommet $j_2 \neq i$ de degré 1 avec une arête incidente e_2. On continue de cette façon

« Une méthode non standard pour compter les arbres : placer un chat dans chaque arbre, promener son chien et compter le nombre de fois où il aboie. »

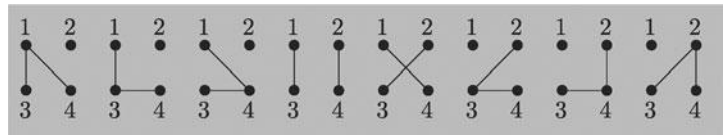
jusqu'à ce que l'on ait déterminé $j_1, j_2, \ldots, j_{n-1}$ et $e_1, e_2, \ldots, e_{n-1}$ tels que $j_i \in e_i$. Permutons maintenant les lignes et les colonnes pour amener j_k à la k-ième ligne et e_k à la k-ième colonne. Puisque, par construction, $j_k \notin e_\ell$ si $k < \ell$, nous voyons que la nouvelle matrice N' est triangulaire inférieure avec tous les éléments de la diagonale principale égaux à ± 1. Ainsi, $\det N = \pm \det N' = \pm 1$, ce qui est le résultat désiré.

Dans le cas particulier $G = K_n$ il est clair que l'on obtient :

$$M_{ii} = \begin{pmatrix} n-1 & -1 & \ldots & -1 \\ -1 & n-1 & & -1 \\ \vdots & & \ddots & \vdots \\ -1 & -1 & \ldots & n-1 \end{pmatrix}$$

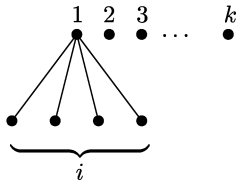
et l'on calcule facilement $\det M_{ii} = n^{n-2}$. □

■ **Troisième preuve (Récurrence).** Une autre méthode classique en combinatoire énumérative consiste à établir une récurrence. On doit essentiellement l'idée suivante à Riordan et Rényi. Pour trouver la récurrence appropriée, on considère un problème plus général qui apparaît déjà dans l'article de Cayley. Soit A un k-ensemble arbitraire de sommets. On désigne par $T_{n,k}$ le nombre de forêts étiquetées sur $\{1, \ldots, n\}$ constituées de k arbres où les sommets de A apparaissent dans des arbres différents. Il est clair que seul le cardinal k de l'ensemble A, et non la nature de ses éléments, joue un rôle. Remarquons que $T_{n,1} = T_n$.



Par exemple, $T_{4,2} = 8$ pour $A = \{1, 2\}$.

Considérons une telle forêt F avec $A = \{1, 2, \ldots, k\}$ et supposons que 1 soit adjacent à i sommets, comme indiqué dans la figure représentée ci-contre. En supprimant 1, ces i sommets en association avec $2, \ldots, k$ engendrent $T_{n-1,k-1+i}$ forêts. Comme on peut choisir arbitrairement les sommets i parmi les $n - k$ sommets différents de $1, \ldots k$, on en déduit, pour $n \geqslant k \geqslant 1$, que :



$$T_{n,k} = \sum_{i=0}^{n-k} \binom{n-k}{i} T_{n-1,k-1+i} \tag{1}$$

où on a posé $T_{0,0} = 1$, $T_{n,0} = 0$ pour $n > 0$. Il est nécessaire de choisir $T_{0,0} = 1$ pour que $T_{n,n} = 1$.

Proposition.

$$T_{n,k} = k\, n^{n-k-1} \tag{2}$$

et donc en particulier, :

$$T_{n,1} = T_n = n^{n-2}.$$

■ **Preuve.** En utilisant la relation (1) et en procédant par récurrence, on trouve :

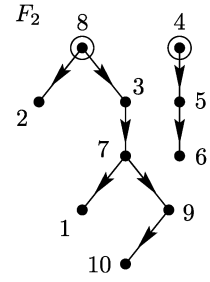
$$\begin{aligned}
T_{n,k} &= \sum_{i=0}^{n-k} \binom{n-k}{i} (k-1+i)(n-1)^{n-1-k-i} \qquad (i \to n-k-i) \\
&= \sum_{i=0}^{n-k} \binom{n-k}{i} (n-1-i)(n-1)^{i-1} \\
&= \sum_{i=0}^{n-k} \binom{n-k}{i} (n-1)^{i} - \sum_{i=1}^{n-k} \binom{n-k}{i} i(n-1)^{i-1} \\
&= n^{n-k} - (n-k) \sum_{i=1}^{n-k} \binom{n-1-k}{i-1} (n-1)^{i-1} \\
&= n^{n-k} - (n-k) \sum_{i=0}^{n-1-k} \binom{n-1-k}{i} (n-1)^{i} \\
&= n^{n-k} - (n-k) n^{n-1-k} \;=\; k\, n^{n-1-k}
\end{aligned}$$

□

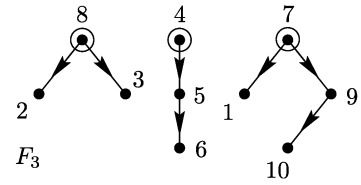
■ **Quatrième preuve (Double décompte).** La merveilleuse idée suivante, due à Jim Pitman, donne la formule de Cayley ainsi que sa généralisation (2) sans récurrence ni bijection : elle fait seulement appel à un double décompte astucieux.

Une *forêt enracinée* sur $\{1, \ldots, n\}$ est une forêt et un choix de racine dans chaque arbre qui la compose. Soit $\mathcal{F}_{n,k}$ l'ensemble de toutes les forêts enracinées constituées de k arbres enracinés. En particulier, $\mathcal{F}_{n,1}$ est l'ensemble de tous les arbres enracinés.

Remarquons que $|\mathcal{F}_{n,1}| = nT_n$, puisqu'il y a dans chaque arbre n choix de racine. Nous considérons maintenant $F_{n,k} \in \mathcal{F}_{n,k}$ comme un graphe *orienté* où toutes les arêtes sont orientées depuis les racines. On dit qu'une forêt F *contient* une autre forêt F' si F contient F' en tant que graphe orienté. Il est clair que si F contient proprement F', alors F a moins de composantes que F'. La figure montre deux telles forêts, les racines se trouvant en haut.



Voici l'idée cruciale. On dit qu'une suite $F_1, \ldots, F_k$ de forêts est une *suite raffinante* si $F_i \in \mathcal{F}_{n,i}$ et F_i contient F_{i+1}, pour tout i.



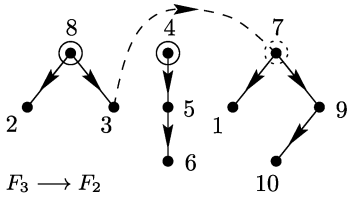
Soit maintenant F_k une forêt fixée dans $\mathcal{F}_{n,k}$. On note :

- $N(F_k)$ le nombre d'arbres enracinés qui contiennent F_k et
- $N^*(F_k)$ le nombre de suites raffinantes qui se terminent dans F_k.

Dénombrons $N^*(F_k)$ de deux façons différentes, d'abord en commençant par un arbre et ensuite en commençant par F_k. Supposons que $F_1 \in \mathcal{F}_{n,1}$ contienne F_k. Puisque l'on peut supprimer les $k-1$ arêtes de $F_1 \backslash F_k$ dans un ordre quelconque pour obtenir une suite raffinante de F_1 à F_k, on a

$$N^*(F_k) \;=\; N(F_k)\,(k-1)! \tag{3}$$

Commençons maintenant de l'autre côté. Pour produire un F_{k-1} à partir de F_k, il faut ajouter une arête orientée, allant d'un sommet quelconque a



vers l'une quelconque des $k-1$ racines des arbres qui ne contiennent pas a (voir figure pour le passage de F_3 à F_2 : on ajoute l'arête 3●→●7). Il y a donc $n(k-1)$ choix possibles. De la même manière, pour F_{k-1} nous pouvons produire une arête orientée qui va d'un sommet quelconque b vers l'une quelconque des $k-2$ racines des arbres qui ne contiennent pas b. Cela laisse $n(k-2)$ choix possibles. En continuant de cette manière, on obtient :

$$N^*(F_k) \;=\; n^{k-1}(k-1)!, \tag{4}$$

et l'on trouve, à l'aide de la relation (3), la relation simple et inattendue :

$$N(F_k) = n^{k-1} \qquad \text{pour } \textit{tout } F_k \in \mathcal{F}_{n,k}.$$

Pour $k = n$, F_n se compose seulement de n sommets isolés. Par conséquent, $N(F_n)$ dénombre *tous* les arbres enracinés. Finalement $|\mathcal{F}_{n,1}| = n^{n-1}$; on obtient donc la formule de Cayley. □

En fait, cette démonstration apporte bien plus. La formule (4) donne pour $k = n$:

$$\#\{\text{suites raffinantes } (F_1, F_2, \ldots, F_n)\} \;=\; n^{n-1}(n-1)! \tag{5}$$

Pour $F_k \in \mathcal{F}_{n,k}$, on note $N^{**}(F_k)$ le nombre de ces suites raffinantes $F_1, \ldots, F_n$ dont le k-ième terme est F_k. Il est clair que c'est $N^*(F_k)$ fois le nombre de façons de choisir $(F_{k+1}, \ldots, F_n)$. Cependant, ce dernier nombre est $(n-k)!$ puisque l'on peut supprimer les $n-k$ arêtes de F_k de n'importe quelle façon, et donc :

$$N^{**}(F_k) \;=\; N^*(F_k)(n-k)! \;=\; n^{k-1}(k-1)!(n-k)! \tag{6}$$

Puisque ce nombre ne dépend pas du choix de F_k, la division de (5) par (6) donne le nombre de forêts enracinées à k arbres :

$$|\mathcal{F}_{n,k}| \;=\; \frac{n^{n-1}(n-1)!}{n^{k-1}(k-1)!(n-k)!} \;=\; \binom{n}{k} k\, n^{n-1-k}$$

Il y a $\binom{n}{k}$ façons de choisir k racines, on a montré une nouvelle fois la formule $T_{n,k} = kn^{n-k-1}$ sans recourir à une récurrence.

Terminons par une note historique. L'article de Cayley de 1889 a été précédé par les travaux de Carl W. Borchardt (1860), ce que reconnut Cayley lui-même. Un résultat équivalent apparut même plus tôt dans un article de James J. Sylvester (1857), voir [2, chapitre 3]. La nouveauté apportée par l'article de Cayley réside dans le recours à la théorie des graphes et depuis le théorème est associé à son nom.

Bibliographie

[1] M. AIGNER : *Combinatorial Theory,* Springer-Verlag, Berlin Heidelberg New York 1979 ; Reprint 1997.

[2] N. L. BIGGS, E. K. LLOYD & R. J. WILSON : *Graph Theory 1736-1936,* Clarendon Press, Oxford 1976.

[3] A. CAYLEY : *A theorem on trees,* Quart. J. Pure Appl. Math. **23** (1889), 376-378 ; Collected Mathematical Papers Vol. 13, Cambridge University Press 1897, 26-28.

[4] A. JOYAL : *Une théorie combinatoire des séries formelles,* Advances in Math. **42** (1981), 1-82.

[5] J. PITMAN : *Coalescent random forests,* J. Combinatorial Theory, Ser. A **85** (1999), 165-193.

[6] H. PRÜFER : *Neuer Beweis eines Satzes über Permutationen,* Archiv der Math. u. Physik (3) **27** (1918), 142-144.

[7] A. RÉNYI : *Some remarks on the theory of trees.* MTA Mat. Kut. Inst. Kozl. (Publ. math. Inst. Hungar. Acad. Sci.) **4** (1959), 73-85 ; Selected Papers Vol. 2, Akadémiai Kiadó, Budapest 1976, 363-374.

[8] J. RIORDAN : *Forests of labeled trees,* J. Combinatorial Theory **5** (1968), 90-103.

Identités et bijections

Chapitre 31

On considère le produit infini $(1+x)(1+x^2)(1+x^3)(1+x^4)\cdots$ que l'on développe selon la méthode habituelle sous la forme d'une série $\sum_{n\geqslant 0} a_n x^n$ en regroupant les termes correspondant à la même puissance de x. On trouve alors pour les premiers termes :

$$\prod_{k\geqslant 1}(1+x^k) = 1 + x + x^2 + 2x^3 + 2x^4 + 3x^5 + 4x^6 + 5x^7 + \dots \quad (1)$$

On constate, par exemple, que $a_6 = 4$ et $a_7 = 5$, et l'on peut penser (à juste titre) que a_n tend vers l'infini avec n.

Le produit $(1-x)(1-x^2)(1-x^3)(1-x^4)\cdots$, qui s'écrit pourtant aussi simplement que le précédent, conduit à un phénomène plus subtil. Si l'on développe, on trouve :

$$\prod_{k\geqslant 1}(1-x^k) \;=\; 1 - x - x^2 + x^5 + x^7 - x^{12} - x^{15} + x^{22} + x^{26} - \dots \quad (2)$$

Il semble que tous les coefficients soient égaux à 1, -1 ou 0. Cette assertion est-elle vraie et, le cas échéant, selon quel motif se répartissent les valeurs en question ?

L'étude de la convergence de séries et de produits infinis joue un rôle central en analyse depuis les débuts de la discipline ; des contributions sur ce thème ont été faites par les plus grands noms, de Leonhard Euler à Srinivasa Ramanujan.

Cependant, en écrivant des identités comme (1) et (2), on ne se pose pas la question de la convergence, on manipule simplement les coefficients. On dit que l'on travaille avec des séries ou des produits « formels ». Dans ce contexte, nous allons montrer comment des arguments combinatoires permettent de prouver des identités qui, de prime abord, semblent difficiles à établir.

La notion fondamentale qui apparaît ici est celle de *partition d'un entier naturel*. On appelle partition de l'entier n toute suite $\lambda_1, \lambda_2, \dots, \lambda_t$ décroissante d'entiers telle que :

$$\lambda : \;\; n \;=\; \lambda_1 + \lambda_2 + \dots + \lambda_t \quad \text{avec} \quad \lambda_1 \geqslant \lambda_2 \geqslant \dots \geqslant \lambda_t \geqslant 1$$

t est la longueur de la partition. Dans la suite, nous appellerons partition de n toute décomposition de n en somme d'entiers naturels non nuls et laisserons au lecteur le soin d'ordonner les termes de cette somme pour en faire une suite conforme à la définition. Nous noterons $P(n)$ l'ensemble de

$5 = 5$
$5 = 4 + 1$
$5 = 3 + 2$
$5 = 3 + 1 + 1$
$5 = 2 + 2 + 1$
$5 = 2 + 1 + 1 + 1$
$5 = 1 + 1 + 1 + 1 + 1$

Les sept partitions de 5.

toutes les partitions de n et $p(n) := |P(n)|$ le nombre de ces partitions, avec la convention $p(0) = 1$. Que viennent faire les partitions dans notre problème ? Considérons le produit infini de séries suivant :

$$(1+x+x^2+x^3+\dots)(1+x^2+x^4+x^6+\dots)(1+x^3+x^6+x^9+\dots)\cdots \quad (3)$$

dont le k-ième facteur est $(1+x^k+x^{2k}+x^{3k}+\dots)$. Quel est le coefficient de x^n lorsqu'on développe ce produit en une série $\sum_{n\geqslant 0} a_n x^n$? Un peu de réflexion permet de se convaincre que ce coefficient n'est autre que le nombre de façons différentes d'écrire n comme une somme :

$$\begin{aligned} n &= n_1\cdot 1 + n_2\cdot 2 + n_3\cdot 3 + \dots \\ &= \underbrace{1+\dots+1}_{n_1} + \underbrace{2+\dots+2}_{n_2} + \underbrace{3+\dots+3}_{n_3} + \dots \end{aligned}$$

Le coefficient recherché est donc le nombre $p(n)$ de partitions de n. Comme la somme de la série géométrique $1 + x^k + x^{2k} + \dots$ est égale à $\frac{1}{1-x^k}$, on vient de prouver une première identité :

$$\prod_{k\geqslant 1} \frac{1}{1-x^k} = \sum_{n\geqslant 0} p(n)\, x^n . \quad (4)$$

Mieux encore, on constate que le facteur $\frac{1}{1-x^k}$ est celui qui correspond à la contribution de k dans une partition de n. Ainsi, si l'on exclut $\frac{1}{1-x^k}$ du produit du premier membre de l'égalité (4), alors k n'apparaît pas dans les partitions du second membre. C'est ainsi que l'on obtient immédiatement :

$$\prod_{i\geqslant 1} \frac{1}{1-x^{2i-1}} = \sum_{n\geqslant 0} p_o(n)\, x^n, \quad (5)$$

où $p_o(n)$ est le nombre de partitions de n dont les décompositions ne font intervenir que des nombres *impairs*. Une identité analogue peut être établie pour les décompositions ne faisant intervenir que des entiers *pairs*.

À ce stade de l'exposé, il est facile de prévoir ce que sera le n-ième coefficient du produit infini $\prod_{k\geqslant 1}(1+x^k)$. Cette fois nous ne considérons que les partitions dans lesquelles l'un quelconque des termes k apparaît au plus une fois. Autrement dit, le produit infini de (1) est développé en :

$$\prod_{k\geqslant 1}(1+x^k) = \sum_{n\geqslant 0} p_d(n)\, x^n , \quad (6)$$

où $p_d(n)$ désigne le nombre de partitions de n en termes tous *distincts*.

C'est maintenant que la méthode des séries formelles révèle toute sa puissance. Puisque $1-x^2 = (1-x)(1+x)$, nous pouvons écrire :

$$\prod_{k\geqslant 1}(1+x^k) = \prod_{k\geqslant 1}\frac{1-x^{2k}}{1-x^k} = \prod_{k\geqslant 1}\frac{1}{1-x^{2k-1}}$$

$6 = 5+1$
$6 = 3+3$
$6 = 3+1+1+1$
$6 = 1+1+1+1+1+1$

Partitions de 6 ne faisant intervenir que des entiers impairs : $p_o(6) = 4$.

$7 = 7$
$7 = 5+1+1$
$7 = 3+3+1$
$7 = 3+1+1+1+1$
$7 = 1+1+1+1+1+1+1$

$7 = 7$
$7 = 6+1$
$7 = 5+2$
$7 = 4+3$
$7 = 4+2+1$

Les partitions de 7 en entiers impairs (en haut) et en entiers distincts (en bas) : $p_o(7) = p_d(7) = 5$.

car tous les facteurs $1-x^{2i}$ s'éliminent. Les produits infinis introduits en (5) et en (6) sont donc identiques et il en va de même pour les séries auxquels ils sont associés, si bien que l'on est conduit au joli résultat suivant :

$$p_o(n) \;=\; p_d(n) \qquad \text{pour tout } n \geqslant 0. \tag{7}$$

Une égalité aussi étonnante requiert qu'on l'établisse directement par le biais d'une bijection — du moins c'est le point de vue de tout combinatoriste qui se respecte.

Problème. *Soit $P_o(n)$ (respectivement $P_d(n)$) l'ensemble des partitions de n en entiers impairs (respectivement en entiers distincts). Trouver une bijection de $P_o(n)$ sur $P_d(n)$.*

Plusieurs bijections répondant à la question sont connues. Celle que nous présentons dans la suite, due à J. W. L. Glaisher (1907), est peut-être la plus pure. Soit λ une partition de n en entiers impairs. En regroupant les termes on peut écrire :

$$\begin{aligned} n &= \underbrace{\lambda_1 + \ldots + \lambda_1}_{n_1} + \underbrace{\lambda_2 + \ldots + \lambda_2}_{n_2} + \ldots + \underbrace{\lambda_t + \ldots + \lambda_t}_{n_t} \\ &= n_1 \cdot \lambda_1 + n_2 \cdot \lambda_2 + \ldots + n_t \cdot \lambda_t. \end{aligned}$$

On écrit à présent $n_1 = 2^{m_1} + 2^{m_2} + \ldots + 2^{m_r}$ selon sa décomposition en base 2 et l'on fait de même pour tous les n_i. On trouve alors une nouvelle partition de n :

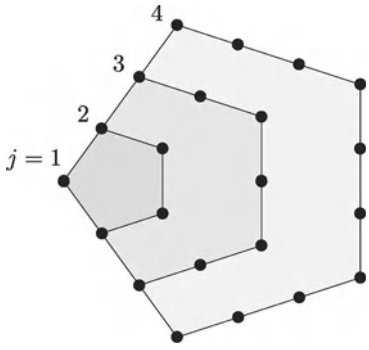
$$\lambda' : \;\; n = 2^{m_1}\lambda_1 + 2^{m_2}\lambda_1 + \ldots + 2^{m_r}\lambda_1 + 2^{k_1}\lambda_2 + \ldots$$

Il faut vérifier que λ' appartient à $P_d(n)$ et que $\phi : \lambda \longmapsto \lambda'$ est bien une bijection. Ces deux points sont faciles à établir. Si $2^a\lambda_i = 2^b\lambda_j$, alors $2^a = 2^b$ car λ_i et λ_j sont tous les deux impairs. Ainsi, $\lambda_i = \lambda_j$. Il en résulte que λ' appartient à $P_d(n)$. Par ailleurs, si l'on peut écrire $n = \mu_1 + \mu_2 + \ldots + \mu_s$ où tous les termes μ_i sont distincts, on peut proposer un schéma pour trouver l'antécédent de cette décomposition par ϕ en regroupant les termes μ_i présentant la même puissance de 2 et en en déduisant la décomposition en somme d'entiers impairs, comme cela est montré dans l'exemple développé ci-contre dans la marge.

La manipulation de produits formels a donc conduit à l'égalité $p_o(n) = p_d(n)$ pour les partitions d'entiers ; nous avons alors donné une autre démonstration de cette égalité en exhibant une bijection. Nous allons procéder en sens inverse : nous allons établir une égalité concernant les partitions à l'aide d'une bijection et en déduire une identité. Notre objectif est cette fois de trouver la forme du développement obtenu dans l'égalité (2).

Par exemple :
$\lambda : 25 = 5{+}5{+}5{+}3{+}3{+}1{+}1{+}1{+}1$
est envoyé par ϕ sur :
$\lambda' : 25 = (2{+}1)5 + (2)3 + (4)1$
$\quad = 10 + 5 + 6 + 4$
$\quad = 10 + 6 + 5 + 4\,.$

On écrit :
$\lambda' : 30 = 12 + 6 + 5 + 4 + 3$
sous la forme :
$30 = 4(3{+}1) + 2(3) + 1(5{+}3)$
$\quad = (1)5 + (4{+}2{+}1)3 + (4)1$
et l'on obtient pour $\phi^{-1}(\lambda')$ la partition :
$\lambda : 30 = 5+3+3+3+3+3+3+3+1+1+1+1$
en entiers impairs.



Nombres pentagonaux.

Considérons :

$$1 - x - x^2 + x^5 + x^7 - x^{12} - x^{15} + x^{22} + x^{26} - x^{35} - x^{40} + \dots$$

À l'exception du terme constant, les exposants semblent s'apparier par signes et si l'on prend le degré du premier terme de la paire à chaque fois, on trouve la suite :

$$1 \quad 5 \quad 12 \quad 22 \quad 35 \quad 51 \quad 70 \quad \dots$$

bien connue d'Euler. Il s'agit de la suite des *nombres pentagonaux* $f(j)$ dont le nom est suggéré par la figure présentée ci-contre dans la marge. On montre facilement que $f(j) = \frac{3j^2-j}{2}$ et que $\bar{f}(j) = \frac{3j^2+j}{2}$ pour le second élément de chaque paire. En d'autres termes, nous conjecturons, comme le fit autrefois Euler, que l'on peut établir l'égalité suivante :

Théorème.

$$\prod_{k\geqslant 1}(1-x^k) \;=\; 1+\sum_{j\geqslant 1}(-1)^j\Big(x^{\frac{3j^2-j}{2}}+x^{\frac{3j^2+j}{2}}\Big). \qquad (8)$$

Euler a prouvé ce remarquable résultat en faisant des calculs sur des séries formelles mais nous allons en donner une preuve digne de passer à la postérité en exhibant une bijection. Nous commençons par remarquer que, selon (4), le produit $\prod_{k\geqslant 1}(1-x^k)$ est exactement l'inverse de la série $\sum_{n\geqslant 0} p(n)x^n$ associée aux partitions. Ainsi, si l'on pose $\prod_{k\geqslant 1}(1-x^k) =: \sum_{n\geqslant 0} c(n)x^n$, on trouve :

$$\Big(\sum_{n\geqslant 0} c(n)x^n\Big)\cdot\Big(\sum_{n\geqslant 0} p(n)x^n\Big) \;=\; 1.$$

La comparaison des coefficients montre que $c(n)$ est *l'unique* suite vérifiant $c(0) = 1$ et :

$$\sum_{k=0}^{n} c(k)p(n-k) \;=\; 0 \qquad \text{pour tout } n \geqslant 1. \qquad (9)$$

Si l'on écrit le second membre de (8) sous la forme $\sum_{j=-\infty}^{\infty}(-1)^j x^{\frac{3j^2+j}{2}}$, il reste à montrer que la définition :

$$c(k) \;=\; \begin{cases} 1 & \text{si } k = \frac{3j^2+j}{2}, \text{avec } j \in \mathbb{Z} \text{ pair,} \\ -1 & \text{si } k = \frac{3j^2+j}{2}, \text{avec } j \in \mathbb{Z} \text{ impair,} \\ 0 & \text{sinon,} \end{cases}$$

est bien celle de l'unique suite qui nous intéresse. En posant $b(j) = \frac{3j^2+j}{2}$ pour $j \in \mathbb{Z}$ et en substituant cette notation dans (9), la conjecture s'écrit simplement :

$$\sum_{j \text{ pair}} p(n - b(j)) \;=\; \sum_{j \text{ impair}} p(n - b(j)) \qquad \text{pour tout } n,$$

où, bien sûr, on considère seulement les j pour lesquels $b(j) \leqslant n$. Le décor est planté, il faut maintenant exhiber une bijection :

$$\psi : \bigcup_{j \text{ pair}} P(n - b(j)) \longrightarrow \bigcup_{j \text{ impair}} P(n - b(j)).$$

Plusieurs bijections ont été proposées pour répondre à la question ; la suivante, due à David Bressoud et Doron Zeilberger est étonnament simple. Nous nous contentons d'en donner la définition et invitons le lecteur à faire la preuve de son caractère bijectif (on remarquera qu'en appliquant un procédé analogue à $\psi(\lambda)$ on revient à λ).
À $\lambda : \lambda_1 + \ldots + \lambda_t \in P(n - b(j))$, on associe :

$$\psi(\lambda) \;:=\; \begin{cases} (t + 3j - 1) + (\lambda_1 - 1) + \ldots + (\lambda_t - 1) & \text{si } \; t + 3j \geqslant \lambda_1, \\ (\lambda_2 + 1) + \ldots + (\lambda_t + 1) + \underbrace{1 + \ldots + 1}_{\lambda_1 - t - 3j - 1} & \text{si } \; t + 3j < \lambda_1, \end{cases}$$

À titre d'exemple, considérer $n = 15$, $j = 2$, donc $b(2) = 7$. La partition $3 + 2 + 2 + 1$ de $P(15 - b(2)) = P(8)$ est envoyée sur $9 + 2 + 1 + 1$, qui appartient à $P(15 - b(1)) = P(13)$.

où l'on ne mentionne pas les 0 éventuels. On constate que $\psi(\lambda)$ appartient à $P(n - b(j-1))$ dans le premier cas et à $P(n - b(j+1))$ dans le deuxième cas.

Tout cela est magnifique et l'on peut en tirer encore davantage. On sait déjà que :

$$\prod_{k \geqslant 1} (1 + x^k) \;=\; \sum_{n \geqslant 0} p_d(n)\, x^n.$$

Tels des experts de la manipulation des séries formelles, nous constatons que l'introduction de la nouvelle variable y conduit à :

$$\prod_{k \geqslant 1} (1 + y x^k) \;=\; \sum_{n,m \geqslant 0} p_{d,m}(n)\, x^n y^m,$$

où $p_{d,m}(n)$ dénombre les partitions de l'entier n en exactement m termes distincts. Pour $y = -1$, cela conduit à :

$$\prod_{k \geqslant 1} (1 - x^k) \;=\; \sum_{n \geqslant 0} (E_d(n) - O_d(n))\, x^n, \tag{10}$$

où $E_d(n)$ (respectivement $O_d(n)$) désigne le nombre de partitions de n en un nombre pair (respectivement impair) de termes distincts. Voici maintenant le coup d'éclat. En comparant (10) au développement d'Euler dans

Les partitions pour le cas $n = 10$:
$10 = 9 + 1$
$10 = 8 + 2$
$10 = 7 + 3$
$10 = 6 + 4$
$10 = 4 + 3 + 2 + 1$
et :
$10 = 10$
$10 = 7 + 2 + 1$
$10 = 6 + 3 + 1$
$10 = 5 + 4 + 1$
$10 = 5 + 3 + 2$.
On vérifie que $E_d(10) = O_d(10) = 5$.

Timbre indien à l'effigie de Srinivasa Ramanujan.

(8), on trouve le magnifique résultat suivant :

$$E_d(n) - O_d(n) = \begin{cases} 1 & \text{si } n = \frac{3j^2 \pm j}{2} \text{ avec } j \geqslant 0 \text{ pair,} \\ -1 & \text{si } n = \frac{3j^2 \pm j}{2} \text{ avec } j \geqslant 1 \text{ impair,} \\ 0 & \text{sinon.} \end{cases}$$

Ceci n'est, bien sûr, que le début d'une longue histoire qui est toujours d'actualité. La théorie des produits infinis est truffée d'identités inattendues avec leurs contreparties sous forme de bijections. Les exemples les plus célèbres sont les identités connues sous le nom d'identités de Rogers-Ramanujan dans lesquelles le nombre 5 joue un rôle mystérieux :

$$\prod_{k \geqslant 1} \frac{1}{(1 - x^{5k-4})(1 - x^{5k-1})} = \sum_{n \geqslant 0} \frac{x^{n^2}}{(1-x)(1-x^2)\cdots(1-x^n)},$$

$$\prod_{k \geqslant 1} \frac{1}{(1 - x^{5k-3})(1 - x^{5k-2})} = \sum_{n \geqslant 0} \frac{x^{n^2+n}}{(1-x)(1-x^2)\cdots(1-x^n)}.$$

Le lecteur est invité à traduire ces identités en terme de partitions selon le schéma introduit par Percy MacMahon :
- Soit $f(n)$ le nombre de partitions de n dont tous les termes sont de la forme $5k + 1$ ou $5k + 4$ et soit $g(n)$ le nombre de partitions dont les termes diffèrent au moins de 2. Alors $f(n) = g(n)$.
- Soit $r(n)$ le nombre de partitions de n dont tous les termes sont de la forme $5k+2$ ou $5k+3$ et soit $s(n)$ le nombre de partitions dont les termes diffèrent au moins de 2 et ne comportant pas 1. Alors $r(n) = s(n)$.

La plupart des preuves des identités de Rogers-Ramanujan reposent sur l'utilisation des séries formelles et l'on a longtemps attendu une preuve fondée sur une bijection. Une telle preuve a finalement été établie par Adriano Garsia et Stephen Milne ; les bijections qu'elle utilise sont cependant très compliquées et nous ne pouvons faire figurer cette démonstration dans notre ouvrage.

Bibliographie

[1] G. E. ANDREWS : *The Theory of Partitions,* Encyclopedia of Mathematics and its Applications, Vol. 2, Addison-Wesley, Reading MA 1976.

[2] D. BRESSOUD & D. ZEILBERGER : *Bijecting Euler's partitions-recurrence,* Amer. Math. Monthly **92** (1985), 54-55.

[3] A. GARSIA & S. MILNE : *A Rogers-Ramanujan bijection,* J. Combinatorial Theory, Ser. A **31** (1981), 289-339.

[4] S. RAMANUJAN : *Proof of certain identities in combinatory analysis,* Proc. Cambridge Phil. Soc. **19** (1919), 214-216.

[5] L. J. ROGERS : *Second memoir on the expansion of certain infinite products,* Proc. London Math. Soc. **25** (1894), 318-343.

Comment compléter un carré latin — Chapitre 32

Parmi les objets combinatoires les plus anciens, dont l'étude remonte apparemment à des temps reculés, se trouvent les *carrés latins*. Afin d'obtenir un carré latin, on doit remplir les n^2 cellules d'un tableau carré $(n \times n)$ avec les valeurs $1, 2, \ldots, n$ de sorte que chaque nombre apparaisse exactement une fois dans chaque ligne et dans chaque colonne. En d'autres termes, chaque ligne et chaque colonne contient une permutation de l'ensemble $\{1, \ldots, n\}$. Appelons n l'*ordre* du carré latin.

1	2	3	4
2	1	4	3
4	3	1	2
3	4	2	1

Un carré latin d'ordre 4.

Voici le problème que nous voulons traiter : on dispose d'un carré dont certaines cellules sont déjà remplies avec des valeurs de $\{1, 2, \ldots, n\}$. À quelle condition est-il possible de compléter ce carré pour en faire un carré latin ? Pour avoir une chance de réussir, on doit, bien sûr, supposer que chaque valeur apparaît au plus une fois dans chaque ligne et dans chaque colonne. On parlera de *carré latin partiel* d'ordre n si certaines cellules d'un $(n \times n)$-tableau contiennent des valeurs appartenant à l'ensemble $\{1, \ldots, n\}$ de façon telle que chaque valeur apparaisse au plus une fois dans chaque ligne et dans chaque colonne. Le problème est donc le suivant :

> *À quelle condition un carré latin partiel peut-il être complété en un carré latin de même ordre ?*

1	4	2	5	3
4	2	5	3	1
2	5	3	1	4
5	3	1	4	2
3	1	4	2	5

Un carré latin cyclique.

Examinons quelques exemples. Supposons que les $n - 1$ premières lignes soient remplies et que la dernière soit vide. On peut alors facilement remplir la dernière ligne. Chaque valeur apparaît $n-1$ fois dans le carré latin partiel et est donc absente d'une colonne exactement. Par conséquent, le carré peut être complété correctement en écrivant chaque élément dans la colonne où il manque.

Inversement, supposons que seule la première ligne soit remplie. Il est encore facile de compléter le carré en décalant d'une case les éléments de manière cyclique dans chacune des lignes suivantes.

Ainsi, alors que dans le premier cas il n'y a qu'une manière de terminer, on a beaucoup de possibilités dans le deuxième cas. En général, moins il y a de cellules remplies, plus on devrait avoir de possibilités pour compléter le carré.

Cependant l'exemple donné dans la marge montre un carré partiel qui ne peut évidemment pas être complété puisqu'il n'y a aucune possibilité de remplir le coin supérieur droit sans enfreindre les règles.

1	2	...	n-1	
				n

Un carré latin partiel qui ne peut pas être complété.

> *Si l'on a rempli moins de n cellules dans un tableau $n \times n$, peut-on toujours le compléter pour obtenir un carré latin ?*

Cette question a été soulevée par Trevor Evans en 1960 ; l'affirmation selon laquelle une réalisation est toujours possible s'est vite fait connaître sous le nom de conjecture d'Evans. C'est, comme souvent, un raisonnement par récurrence qui conduit au succès. La preuve donnée par Bohdan Smetaniuk en 1981 est un bel exemple qui montre la subtilité nécessaire d'un raisonnement par récurrence pour faire aboutir un tel travail. Et, que demander de plus, la preuve est constructive. Elle permet de compléter explicitement le carré latin à partir d'une configuration initiale partielle.

1	3	2
2	1	3
3	2	1

L : 1 1 1 2 2 2 3 3 3
C : 1 2 3 1 2 3 1 2 3
E : 1 3 2 2 1 3 3 2 1

Avant de passer à la démonstration, penchons-nous avec soin sur les carrés latins en général. On peut voir un carré latin comme un $(3 \times n^2)$-tableau, qui est appelé *tableau en ligne* associé au carré latin. La figure ci-contre montre un carré latin d'ordre 3 et son tableau en ligne associé, où L, C et E représentent les lignes, les colonnes et les éléments.

La condition imposée aux carrés latins équivaut à dire que dans chaque paire de lignes du tableau en ligne, toutes les n^2 paires ordonnées apparaissent (et par conséquent chaque paire apparaît exactement une fois). Évidemment, on peut permuter arbitrairement les symboles dans chaque ligne (ce qui correspond à des permutations de lignes, de colonnes et d'éléments) et l'on obtient encore un carré latin. Cependant, la condition sur le $(3 \times n^2)$-tableau en dit plus : aucun élément ne joue un rôle spécial. On peut aussi permuter globalement les lignes du tableau, on conserve encore les conditions sur le tableau en ligne, et l'on obtient aussi un carré latin.

Si on permute les lignes de l'exemple ci-dessus selon le cycle :

$$L \longrightarrow C \longrightarrow E \longrightarrow L$$

on obtient le carré latin et le tableau en ligne suivants :

1	2	3
3	1	2
2	3	1

L : 1 3 2 2 1 3 3 2 1
C : 1 1 1 2 2 2 3 3 3
E : 1 2 3 1 2 3 1 2 3

Les carrés latins liés par de telles permutations sont dits *conjugués*. Voici l'observation qui va éclairer la démonstration : un carré latin partiel correspond évidemment à un tableau en ligne partiel (chaque paire apparaît au plus une fois dans deux lignes quelconques), et tout conjugué d'un carré latin partiel est encore un carré latin partiel. En particulier, un carré latin partiel peut être complété si et seulement si tout conjugué peut l'être, (il suffit de compléter le conjugué et d'inverser la permutation des trois lignes).

Nous aurons besoin de deux résultats, que l'on doit à Herbert J. Ryser et Charles C. Lindner, déjà connus avant le théorème de Smetaniuk. Si un carré latin partiel est tel que ses r premières lignes sont complètement remplies et que les cellules qui restent sont vides, on dit que l'on a un $(r \times n)$-*rectangle latin*.

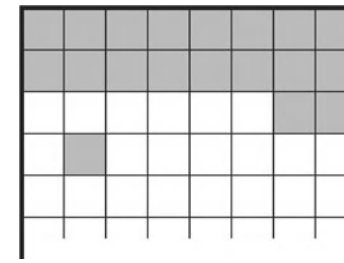
Lemme 1. *Tout $(r \times n)$-rectangle latin, $r < n$, peut être étendu en un $((r+1) \times n)$-rectangle latin et peut donc être complété en un carré latin.*

■ **Preuve.** Nous appliquons le théorème de Hall (voir chapitre 27). Soit A_j l'ensemble des nombres qui n'apparaissent *pas* dans la colonne j. Une $(r+1)$-ième ligne admissible correspond alors précisément à un système de représentants distincts pour la collection $A_1, \ldots, A_n$. Afin de prouver le lemme, nous devons d'abord vérifier la condition de Hall (H). Chaque ensemble A_j a pour cardinal $n - r$ et chaque élément est exactement dans $n - r$ ensembles A_j (puisqu'il apparaît r fois dans le rectangle). m de ces ensembles A_j contiennent $m(n - r)$ éléments et par conséquent au moins m éléments différents, ce qui est exactement la condition (H). □

Lemme 2. *Soit P un carré latin partiel d'ordre n ayant au plus $n-1$ cellules remplies et au plus $\frac{n}{2}$ éléments distincts. Alors P peut être complété en un carré latin d'ordre n.*

■ **Preuve.** Transformons d'abord le problème sous une forme plus pratique. En utilisant le principe de conjugaison abordé ci-dessus nous pouvons remplacer la condition « au plus $\frac{n}{2}$ éléments distincts » en imposant que les entrées apparaissent dans $\frac{n}{2}$ lignes au plus, et nous pouvons supposer aussi que ces lignes sont les lignes supérieures. Supposons donc que les lignes ayant des cellules remplies soient les lignes $1, 2, \ldots, r$, avec f_i cellules remplies dans la ligne i, où $r \leqslant \frac{n}{2}$ et $\sum_{i=1}^{r} f_i \leqslant n-1$. En permutant les lignes, nous pouvons supposer que $f_1 \geqslant f_2 \geqslant \ldots \geqslant f_r$. Nous complétons maintenant les lignes $1, \ldots, r$ pas à pas jusqu'à ce que nous obtenions un $(r \times n)$-rectangle qui puisse, grâce au lemme 1, être étendu en un carré latin.

Supposons que les lignes $1, 2, \ldots, \ell-1$ soient déjà remplies. Dans la ligne ℓ il y a f_ℓ cellules remplies, (nous pouvons supposer qu'elles se trouvent à la fin). Cette situation est décrite dans la figure ci-contre où les parties ombrées désignent les cellules remplies.



Une situation pour $n = 8$, avec $\ell = 3$, $f_1 = f_2 = f_3 = 2$, $f_4 = 1$. Les carrés sombres représentent les cellules pré-remplies, les carrés clairs représentent les cellules qui ont été remplies par le processus de complétion.

La complétion de la ligne ℓ est assurée par une autre application du théorème de Hall, mais elle est cette fois plus subtile. Soit X l'ensemble des éléments qui n'apparaissent *pas* dans la ligne ℓ. On a donc $|X| = n - f_\ell$. Pour $j = 1, \ldots, n - f_\ell$, notons A_j l'ensemble des éléments de X qui n'apparaissent *pas* dans la colonne j (ni au-dessus, ni au-dessous de la ligne ℓ). Afin de compléter la ligne ℓ, nous devons vérifier la condition (H) pour la famille $A_1, \ldots, A_{n-f_\ell}$.

On montre d'abord que :

$$n - f_\ell - \ell + 1 \;>\; \ell - 1 + f_{\ell+1} + \cdots + f_r. \qquad (1)$$

La cas $\ell = 1$ est clair. Sinon, $\sum_{i=1}^{r} f_i < n$, $f_1 \geqslant \cdots \geqslant f_r$ et $1 < \ell \leqslant r$ impliquent :

$$n \;>\; \sum_{i=1}^{r} f_i \;\geqslant\; (\ell-1) f_{\ell-1} + f_\ell + \cdots + f_r$$

Alors, soit $f_{\ell-1} \geqslant 2$ (dans ce cas on a (1)), soit $f_{\ell-1} = 1$. Dans le dernier cas, (1) se réduit à $n > 2(\ell-1) + r - \ell + 1 = r + \ell - 1$, qui est vérifié puisque $\ell \leqslant r \leqslant \frac{n}{2}$.

Considérons à présent m ensembles A_j, $1 \leqslant m \leqslant n - f_\ell$, et soit B leur réunion. Nous devons montrer que $|B| \geqslant m$. Considérons le nombre c de cellules des m colonnes correspondant aux A_j qui contiennent des éléments de X. Il y a au plus $(\ell-1)m$ cellules de ce type au-dessus de la ligne ℓ et au plus $f_{\ell+1} + \cdots + f_r$ au-dessous de la ligne ℓ. Par conséquent :

$$c \;\leqslant\; (\ell-1)m + f_{\ell+1} + \cdots + f_r$$

D'autre part, chaque élément $x \in X \backslash B$ apparaît dans chacune des m colonnes. Ainsi, $c \geqslant m(|X| - |B|)$ donc (puisque $|X| = n - f_\ell$) :

$$|B| \;\geqslant\; |X| - \tfrac{1}{m} c \;\geqslant\; n - f_\ell - (\ell-1) - \tfrac{1}{m}(f_{\ell+1} + \cdots + f_r)$$

Cela implique que $|B| \geqslant m$ si :

$$n - f_\ell - (\ell - 1) - \tfrac{1}{m}(f_{\ell+1} + \cdots + f_r) \;>\; m - 1$$

c'est-à-dire si :

$$m(n - f_\ell - \ell + 2 - m) \;>\; f_{\ell+1} + \cdots + f_r \qquad (2)$$

L'inégalité (2) est vraie pour $m = 1$, pour $m = n - f_\ell - \ell + 1$ d'après (1) et par conséquent pour toutes les valeurs m comprises entre 1 et $n - f_\ell - \ell + 1$, puisque le côté gauche est une fonction quadratique en m de coefficient directeur -1. Il reste à traiter le cas $m > n - f_\ell - \ell + 1$. Puisque tout élément x de X est contenu dans $\ell - 1 + f_{\ell+1} + \cdots + f_r$ lignes au plus, il peut aussi se trouver dans le même nombre de colonnes au plus. En invoquant encore une fois (1), on constate que x appartient à l'un des ensembles A_j et que dans ce cas $B = X$ donc $|B| = n - f_\ell \geqslant m$. □

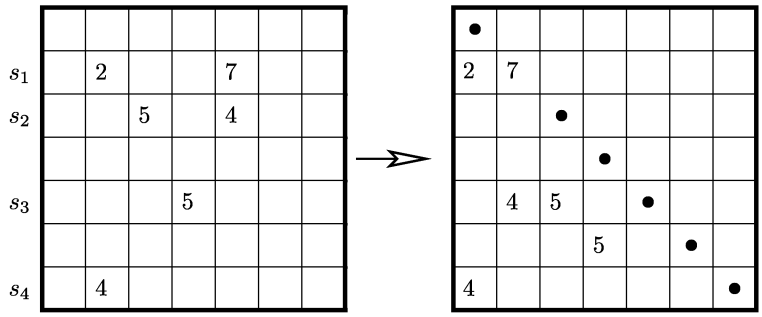
Montrons enfin le théorème de Smetaniuk.

Théorème. *Un carré latin partiel d'ordre n tel que $n - 1$ cellules au plus sont remplies peut être complété en un carré latin du même ordre.*

■ **Preuve.** Nous procédons par récurrence sur n. Les cas $n \leqslant 2$ sont triviaux. Nous étudions donc un carré latin partiel d'ordre $n \geqslant 3$ comportant au plus $n - 1$ cellules déjà remplies. Avec les notations précédentes, ces cellules se trouvent dans $r \leqslant n - 1$ lignes numérotées $s_1, \ldots, s_r$, qui présentent $f_1, \ldots, f_r > 0$ cellules remplies, avec $\sum_{i=1}^{r} f_i \leqslant n$. D'après le lemme 2, on peut supposer qu'il y a plus de $\frac{n}{2}$ éléments distincts. Par conséquent il y a un élément qui apparaît une fois seulement . Après éventuelles permutation et réindexation des lignes, on peut supposer que c'est l'élément n qui apparaît une seule fois en ligne s_1.

Dans l'étape suivante, on va permuter les lignes et les colonnes du carré partiel de sorte qu'une fois les permutations effectuées, toutes les cellules occupées figurent en dessous de la diagonale, à l'exception de la cellule qui contient n qui va finir sur la diagonale (la diagonale étant constituée des éléments (k, k) avec $1 \leqslant k \leqslant n$). On parvient à ce résultat en procédant de la manière suivante : on commence par mettre la ligne s_1 en position f_1. Par permutation des colonnes, on fait en sorte de placer toutes les cellules remplies sur la gauche, de sorte que n apparaisse en dernier sur sa ligne, sur la diagonale. On déplace ensuite la ligne s_2 vers la position $1 + f_1 + f_2$ et, de nouveau, par permutation de colonnes, on déplace les cellules remplies vers la gauche. De manière générale, pour $1 < i \leqslant r$, la ligne s_i est déplacé vers la position $1 + f_1 + f_2 + \cdots + f_i$ et les cellules pleines sont déplacés vers la gauche. Ce procédé conduit manifestement à la configuration souhaitée. La figure qui suit montre un exemple avec $n = 7$: les lignes $s_1 = 2$, $s_2 = 3$, $s_3 = 5$ et $s_4 = 7$ associées respectivement à $f_1 = f_2 = 2$ et $f3 = f_4 = 1$ sont déplacées vers les lignes de numéros respectifs 2, 5 ; 6 et 7. Dans le même temps les colonnes sont permutées « vers la gauche » si bien qu'à la

fin toutes les éléments du carré à l'exception de 7 se retrouvent en dessous de la diagonale signalée par des •.



De manière à pouvoir appliquer l'hypothèse de récurrence, on va enlever la valeur n de la diagonale et ne pas prendre en compte la première ligne et la dernière colonne (qui ne contiennent que des cellules vides) : on est donc maintenant face à un carré latin partiel d'ordre $n-1$ comportant $n-2$ cellules déjà remplies, qui d'après l'hypothèse de récurrence, peut être complété en un carré latin d'ordre $n-1$. On peut voir dans la marge l'une des nombreuses façons de compléter le carré partiel de l'exemple étudié. Les éléments originaux figurent en gras, ils sont déjà en position finale.

2	3	4	1	6	5	
5	6	1	4	2	3	
1	2	3	6	5	4	
6	**4**	**5**	2	3	1	
3	1	6	**5**	4	2	
4	5	2	3	1	6	

Dans l'étape suivante, on voudrait déplacer les éléments de la diagonale vers la dernière ligne et placer les éléments n sur la diagonale à leur place. Cependant, il n'est en général pas possible de procéder à cette opération car les éléments diagonaux doivent être distincts. C'est pourquoi on va procéder successivement pour $k = 2, 3, \ldots, n-1$ à l'opération suivante :

Mettre la valeur n dans la cellule (k, n) ce qui conduit à un carré latin partiel correct. Ensuite on échange le contenu x_k de la diagonale (k, k) avec le contenu n de la cellule (k, n) de la dernière colonne.

Si la valeur x_k ne figure pas déjà dans la dernière colonne, on en a terminé avec k. Après cette manipulation, les éléments de la k-ième colonne ne changeront plus.

2	3	4	1	6	5	7
5	6	1	4	2	3	
1	2	3	6	5	4	
6	**4**	**5**	2	3	1	
3	1	6	**5**	4	2	
4	5	2	3	1	6	

2	7	4	1	6	5	3
5	6	1	4	2	3	7
1	2	3	6	5	4	
6	**4**	**5**	2	3	1	
3	1	6	**5**	4	2	
4	5	2	3	1	6	

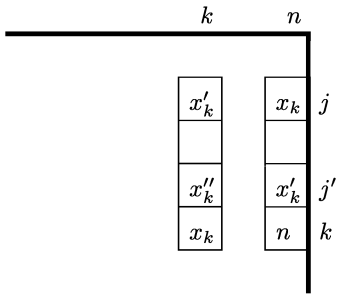
2	7	4	1	6	5	3
5	6	7	4	2	3	1
1	2	3	6	5	4	7
6	**4**	**5**	2	3	1	
3	1	6	**5**	4	2	
4	5	2	3	1	6	

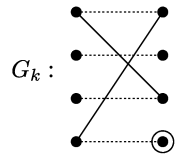
Dans l'exemple étudié cela fonctionne sans problème pour $k = 2$, 3 et 4 ; les éléments diagonaux correspondants 3, 1 et 6, se déplacent vers la dernière colonne. Les figures suivantes montrent les opérations correspon-

dantes. Les cellules ombrées contiennent des éléments qui demeurent inchangés au cours des manipulations.

Il reste à traiter le cas où il y a déjà un élément x_k dans la dernière colonne. Dans ce cas on procède de la manière suivante :

S'il y a déjà un élément x_k dans une cellule (j, n) de la dernière colonne (avec $2 \leqslant j < k$), alors on échange aussi dans la ligne j l'élément x_k avec x'_k de la k-ième colonne. Si l'élément x'_k figure lui-aussi déjà dans une cellule (j', n) de la n-ième colonne, alors on échange aussi les éléments de la j'-ième ligne qui apparaissent dans les n-ième et k-ième colonnes, et ainsi de suite.





En procédant de la sorte il n'y aura jamais deux fois le même élément dans une ligne puisque l'on ne fait que des échanges entre éléments de cette ligne. De même, le procédé assure qu'il n'y aura pas deux éléments égaux dans une même colonne. Il ne reste donc plus qu'à vérifier que ce processus d'échange entre les k-ième et n-ième colonnes ne boucle pas indéfiniment. Cela peut se voir en considérant le graphe biparti G_k dont les sommets correspondent aux cellules (i, k) et (j, n) avec $2 \leqslant i, j \leqslant k$ et dont les éléments ont pu faire l'objet d'échanges. Il y a une arête entre (i, k) et (j, n) si ces deux cellules se trouvent dans la même ligne (c'est-à-dire si $i = j$) ou si les deux cellules avant tout échange contenaient le même élément (ce qui implique $i \neq j$). Dans le schéma, les arêtes correspondant au cas $i = j$ apparaissent en pointillé. Tous les sommets de G_k sont de degré 1 ou 2. La cellule (k, n) correspond à un sommet de degré 1. Ce sommet est le début d'un chemin qui conduit à la colonne k sur une arête horizontale, et qui ramène éventuellement à la colonne n par une arête inclinée, puis qui retourne vers la colonne k horizontalement et ainsi de suite. Le chemin se termine en colonne k sur une valeur qui n'apparaît pas dans la colonne n. Ainsi, le procédé d'échange doit se terminer en un point en lequel il faut introduire un nouvel élément dans la dernière colonne. On s'arrête donc en colonne k et les contenus des cellules $(i; k)$ $(i \geqslant 2)$ de la k-ième colonne sont définitivement fixés.

2	7	4	1	6	5	3
5	6	7	4	2	3	1
1	2	3	7	5	4	6
6	4	**5**	2	3	1	7
3	1	6	**5**	4	2	
4	5	2	3	1	6	

2	7	4	1	3	5	6
5	6	7	4	2	3	1
1	2	3	7	6	4	5
6	4	**5**	2	7	1	3
3	1	6	**5**	4	2	
4	5	2	3	1	6	

Dans l'exemple étudié, le « cas d'échange » se produit pour $k = 5$: l'élément $x_5 = 3$ est déjà présent dans la dernière colonne si bien que l'élément doit être ramené en colonne 5 ; l'élément de substitution $x'_5 = 6$ n'est pas nouveau non plus, il est donc à son tour échangé avec $x''_k = 5$ qui, lui, est nouveau.

Enfin, l'échange pour $k = 6 = n - 1$ ne pose pas de problème et il n'y a alors plus qu'une seule manière de terminer le carré latin.

2	7	4	1	3	5	6
5	6	7	4	2	3	1
1	2	3	7	6	4	5
6	**4**	**5**	2	7	1	3
3	1	6	**5**	4	2	7
4	5	2	3	1	6	

2	7	4	1	3	5	6
5	6	7	4	2	3	1
1	2	3	7	6	4	5
6	**4**	**5**	2	7	1	3
3	1	6	**5**	4	7	2
4	5	2	3	1	6	

7	3	1	6	4	2	4
2	7	4	1	3	5	6
5	6	7	4	2	3	1
1	2	3	7	6	4	5
6	**4**	**5**	2	7	1	3
3	1	6	**5**	4	7	2
4	5	2	3	1	6	7

La même chose se produit dans le cas général : on place un élément n dans la cellule (n, n) et ainsi la première ligne peut être complétée par les éléments manquants des colonnes correspondantes (voir lemme 1), ce qui termine la preuve. □

Bibliographie

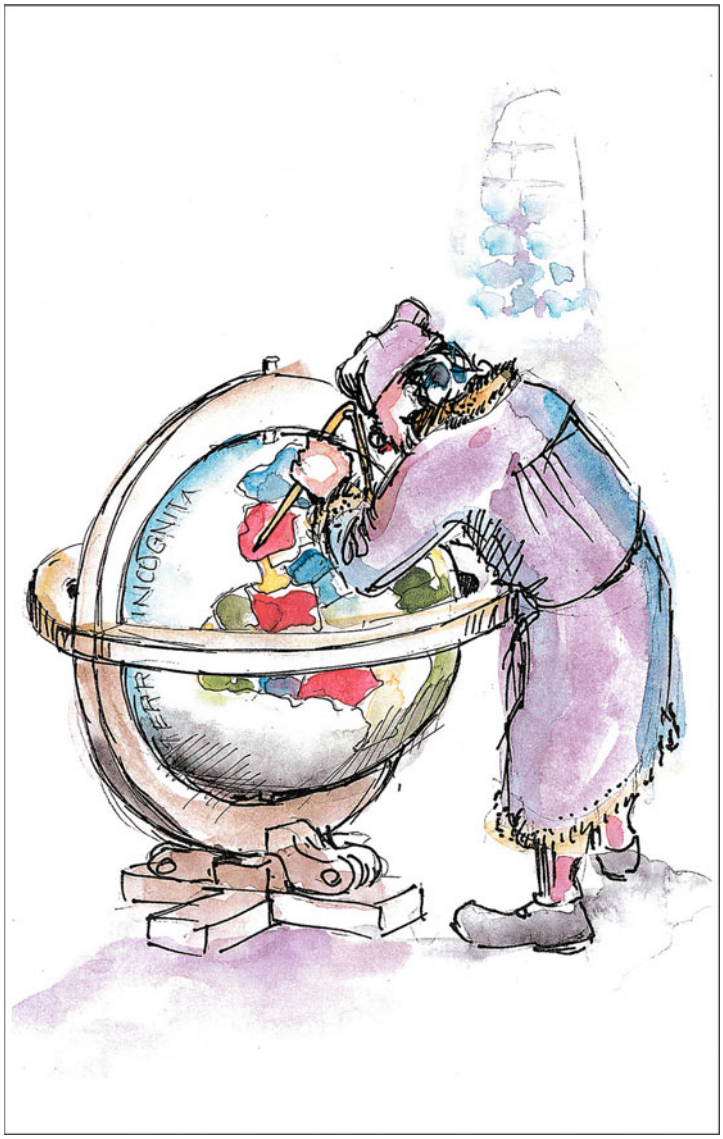
[1] T. EVANS : *Embedding incomplete Latin squares,* Amer. Math. Monthly **67** (1960), 958-961.

[2] C. C. LINDNER : *On completing Latin rectangles,* Canadian Math. Bulletin **13** (1970), 65-68.

[3] H. J. RYSER : *A combinatorial theorem with an application to Latin rectangles,* Proc. Amer. Math. Soc. **2** (1951), 550-552.

[4] B. SMETANIUK : *A new construction on Latin squares I : A proof of the Evans conjecture,* Ars Combinatoria **11** (1981), 155-172.

Théorie des graphes



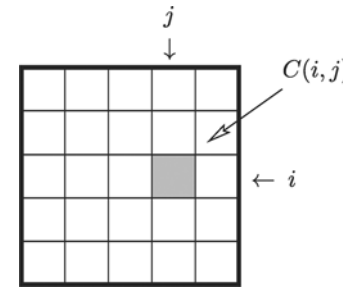
« Le géographe aux quatre couleurs ».

Le problème de Dinitz

Chapitre 33

Le problème des quatre couleurs a eu un rôle majeur dans le développement de la théorie des graphes telle que nous la connaissons aujourd'hui. La coloration de graphes est encore l'un des sujets préférés de nombreux théoriciens. Voici un problème de coloration simple, soulevé par Jeff Dinitz en 1978, qui a résisté à toutes les tentatives jusqu'à la solution étonnamment simple trouvée par Fred Galvin dix-huit ans plus tard.

> *Considérons n^2 cellules formant un $(n \times n)$-carré. On désigne par (i, j) la cellule située à la ligne i et la colonne j. Supposons qu'à chaque cellule (i, j) on associe un ensemble $C(i, j)$ de n couleurs. Est-il toujours possible de colorier la totalité du tableau en coloriant chaque cellule (i, j) d'une couleur issue de son ensemble $C(i, j)$ de façon telle que les couleurs dans chaque ligne et chaque colonne soient distinctes ?*



Pour commencer, considérons le cas où tous les ensembles de couleurs $C(i, j)$ sont identiques et égaux à $\{1, 2, \ldots, n\}$. Alors le problème de Dinitz se réduit à la question suivante : remplir le carré $(n \times n)$ avec les nombres $1, 2, \ldots, n$ de sorte que chaque ligne et chaque colonne soient composées de nombres distincts. En d'autres termes, une telle coloration correspond à un problème de carré latin, abordé au chapitre précédent. Dans ce cas, on peut répondre par l'affirmative à la question posée.

Puisque ce problème est très facile, pourquoi serait-il plus difficile dans le cas général, lorsque l'ensemble $C := \bigcup_{i,j} C(i, j)$ contient plus de n couleurs ? La difficulté provient du fait que toutes les couleurs de C ne sont pas disponibles pour chaque cellule. Dans le cas du carré latin, on peut évidemment choisir une permutation arbitraire des couleurs pour la première ligne, mais ce n'est plus possible dans le cas général. Le cas $n = 2$ illustre déjà cette difficulté.

$\{1, 2\}$	$\{2, 3\}$
$\{1, 3\}$	$\{2, 3\}$

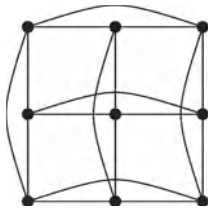
Supposons donnés les ensembles de couleurs indiqués sur la figure. Si l'on choisit les couleurs 1 et 2 pour la première ligne, on se trouve en difficulté puisque l'on ne peut prendre que la couleur 3 pour les deux cellules de la deuxième ligne.

Avant d'aborder le problème de Dinitz, formulons le problème en utilisant le langage de la théorie des graphes. Comme d'habitude, nous considérons des graphes $G = (V, E)$ sans boucles ni arêtes multiples. On note $\chi(G)$ le *nombre chromatique* du graphe, c'est-à-dire le plus petit nombre de couleurs nécessaires pour affecter une couleur à chaque sommet de façon telle que des sommets adjacents aient des couleurs différentes.

En d'autres termes, une coloration définit une partition de V en classes (de sommets de même couleur) telle qu'il n'y ait pas d'arête à l'intérieur d'une classe. Un ensemble $A \subseteq V$ est dit *indépendant* s'il n'y a pas d'arête à l'intérieur de A. Le nombre chromatique est le plus petit cardinal d'une partition constituée d'ensembles indépendants de V.

En 1976 Vizing et trois ans plus tard Erdős, Rubin et Taylor ont étudié la variante suivante qui conduit directement au problème de Dinitz. On donne $G = (V, E)$ un graphe dont chaque sommet v est associé un ensemble de couleurs $C(v)$. Une *coloration par liste* est une coloration $c : V \longrightarrow \bigcup_{v \in V} C(v)$ telle que $c(v) \in C(v)$ pour chaque $v \in V$. La définition du *nombre chromatique listé* $\chi_\ell(G)$ est claire : c'est le plus petit nombre k tel que *pour toute* liste de couleurs $C(v)$ telle que $|C(v)| = k$ pour tout $v \in V$, il existe toujours une coloration par liste. Bien sûr on a $\chi_\ell(G) \leqslant |V|$ (on ne manque jamais de couleur). Puisqu'une coloration ordinaire est un simple cas particulier de coloration par liste pour laquelle tous les ensembles $C(v)$ sont égaux, pour tout graphe G :

$$\chi(G) \leqslant \chi_\ell(G).$$

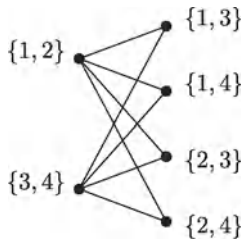


Le graphe S_3.

Pour revenir au problème de Dinitz, considérons le graphe S_n qui a pour ensemble de sommets les n^2 cellules de notre $(n \times n)$-tableau, deux cellules étant adjacentes si et seulement si elles se trouvent sur la même ligne ou sur la même colonne.

Comme n cellules quelconques d'une même ligne sont deux à deux adjacentes, on a besoin d'au moins n couleurs. En outre, toute coloration avec n couleurs correspond à un carré latin, les cellules occupées par le même nombre formant une classe de couleur. Comme nous avons vu que les carrés latins existent, on en déduit que $\chi(S_n) = n$; le problème de Dinitz peut maintenant s'énoncer succinctement comme suit :

$$\text{est-ce que } \chi_\ell(S_n) = n \,?$$



Coloration du graphe $K_{2,4}$.

On pourrait penser que $\chi(G) = \chi_\ell(G)$ pour tout graphe G, mais c'est loin d'être le cas. Considérons le graphe $G = K_{2,4}$. Le nombre chromatique est 2 puisque l'on peut utiliser une couleur pour les deux sommets de gauche et une deuxième couleur pour les sommets de droite. Supposons maintenant que l'on nous donne les ensembles de couleurs indiqués sur la figure.

Pour colorier les sommets de gauche, nous avons quatre possibilités $1|3$, $1|4$, $2|3$ et $2|4$, mais chacune de ces paires se trouve être un ensemble de couleurs sur le côté droit. Une coloration par liste n'est donc pas possible. Par conséquent $\chi_\ell(G) \geqslant 3$; le lecteur peut s'amuser à montrer que $\chi_\ell(G) = 3$ (il n'est pas nécessaire de tester tous les cas possibles !). En généralisant cet exemple, il n'est pas difficile de trouver des graphes G tels que $\chi(G) = 2$, mais tels que $\chi_\ell(G)$ soit arbitrairement grand. Le problème de la coloration par liste n'est donc pas aussi simple que l'on pourrait le croire au premier abord.

Revenons au problème de Dinitz. Un pas important vers la solution a été fait par Jeanette Janssen en 1992 lorsqu'elle a prouvé que $\chi_\ell(S_n) \leqslant n + 1$. Le

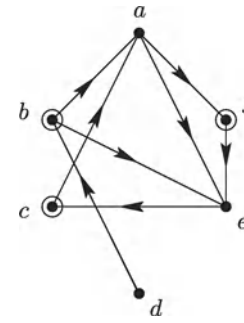
coup de grâce [1] a été donné par Fred Galvin qui a ingénieusement combiné deux résultats (connus tous les deux depuis longtemps). Nous allons étudier ces deux résultats et montrer comment ils impliquent $\chi_\ell(S_n) = n$.

Fixons d'abord quelques notations. Si v est un sommet du graphe G, on note, comme précédemment, $d(v)$ le *degré* de v. Dans le graphe carré S_n, chaque sommet a un degré égal à $2n-2$. Étant donné un sous-ensemble $A \subseteq V$, on désigne par G_A le sous-graphe qui admet A comme ensemble de sommets et qui contient toutes les arêtes de G entre les sommets de A. On appelle G_A le sous-graphe induit par A et on dit que H est un *sous-graphe induit* de G si $H = G_A$ pour un certain A.

Pour énoncer notre premier résultat, nous avons besoin de la notion de *graphe orienté* $\vec{G} = (V, E)$, c'est-à-dire de graphe tel que chaque arête e soit orientée. La notation $e = (u, v)$ signifie que l'on considère un arc e, noté aussi $u \longrightarrow v$, dont le sommet initial est u et dont le sommet final est v. On peut donc parler de *degré externe* $d^+(v)$ et de *degré interne* $d^-(v)$, où $d^+(v)$ dénombre les arêtes ayant v comme sommet initial et $d^-(v)$ dénombre les arêtes ayant v comme sommet final ; on a $d^+(v) + d^-(v) = d(v)$. Lorsque l'on écrit G, on parle du graphe $\vec{G}$ non orienté.

Le concept suivant est issu de la théorie des jeux, il joue un rôle fondamental dans la démonstration.

Définition 1. Soit $\vec{G} = (V, E)$ un graphe orienté. Un *noyau* $K \subseteq V$ est un sous-ensemble de sommets tel que :
(i) K est indépendant dans G, et
(ii) pour tout $u \notin K$ il existe un sommet $v \in K$ tel que $u \longrightarrow v$ soit une arête.



Examinons l'exemple proposé dans la marge. L'ensemble $\{b, c, f\}$ constitue un noyau, en revanche le sous-graphe induit par $\{a, c, e\}$ n'admet pas de noyau puisque les trois arêtes forment un cycle qui traverse les sommets.

Nous sommes maintenant prêts pour énoncer notre premier résultat.

Lemme 1. *Soit $\vec{G} = (V, E)$ un graphe orienté. Supposons qu'à chaque sommet $v \in V$ soit associé un ensemble de couleurs $C(v)$ de cardinal supérieur à son degré externe, $|C(v)| \geqslant d^+(v) + 1$. Si chaque sous-graphe induit de $\vec{G}$ possède un noyau, il existe une coloration par liste de G telle que chaque v ait une couleur de $C(v)$.*

■ **Preuve.** Nous procédons par récurrence sur $|V|$. Pour $|V| = 1$, il n'y a rien à prouver. Choisissons une couleur $c \in C = \bigcup_{v \in V} C(v)$ et posons :

$$A(c) \;:=\; \{v \in V : c \in C(v)\}$$

Par hypothèse, le sous-graphe induit $G_{A(c)}$ possède un noyau $K(c)$. Nous colorions maintenant chaque $v \in K(c)$ avec la couleur c (c'est possible puisque $K(c)$ est indépendant) et supprimons $K(c)$ de G et c de C. Soit G'

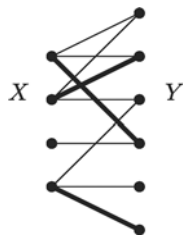
1. N.d.T. : en français dans le texte original.

le sous-graphe induit de G sur $V\backslash K(c)$ ayant $C'(v) = C(v)\backslash\{c\}$ comme nouvelle liste de couleurs. Remarquons que pour chaque $v \in A(c)\backslash K(c)$, le degré externe $d^+(v)$ diminue de 1 au plus (du fait de la condition (ii) vérifiée par un noyau). On a donc encore $d^+(v)+1 \leqslant |C'(v)|$ dans $\vec{G}'$. Les sommets qui se trouvent hors de $A(c)$ vérifient la même condition, puisque dans ce cas les ensembles de couleurs $C(v)$ demeurent inchangés. Le nouveau graphe G' contient moins de sommets que G ; le résultat s'obtient par récurrence. □

La manière de traiter le problème de Dinitz est maintenant claire : il faut trouver une orientation du graphe S_n avec des degrés externes $d^+(v) \leqslant n-1$ pour tout v, qui assure l'existence d'un noyau pour tous les sous-graphes induits. C'est possible grâce au deuxième résultat qui suit.

Nous avons encore besoin de quelques notions préliminaires. Rappelons que, (voir chapitre 10), qu'un graphe *biparti* $G = (X \cup Y, E)$ est un graphe qui présente la propriété suivante : l'ensemble des sommets V se décompose en deux parties X et Y telles que chaque arête a une extrémité dans X et l'autre dans Y. En d'autres termes, les graphes bipartis sont précisément ceux qui peuvent être coloriés avec deux couleurs l'une pour X et l'autre pour Y).

Nous arrivons maintenant à un concept important, « les couplages stables », qui ont une interprétation terre à terre. Un *couplage* M dans un graphe biparti $G = (X \cup Y, E)$ est un ensemble d'arêtes tel qu'aucun couple d'arêtes de M n'ait d'extrémités communes. Dans le graphe représenté dans la marge les arêtes tracées en gras constituent un couplage.



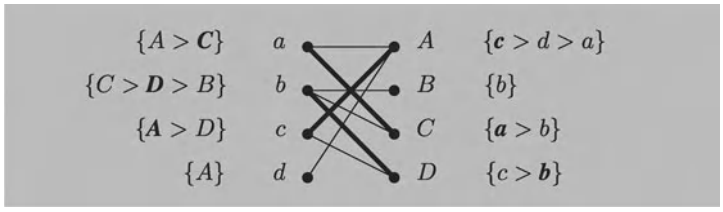
Un graphe biparti avec un couplage.

Si l'on considère X comme un ensemble d'hommes et Y comme un ensemble de femmes, et si l'on interprète $uv \in E$ en disant que u et v peuvent se marier, un couplage est alors un mariage de masse dans lequel personne ne pratique la bigamie. Dans notre contexte, nous avons besoin d'une notion plus fine (et plus réaliste ?) de couplage, telle qu'elle a été suggérée par Gale et Shapley. Dans la vie réelle, chaque personne a évidemment des préférences et c'est cet aspect que l'on veut intégrer. Dans $G = (X \cup Y, E)$ on suppose que pour chaque $v \in X \cup Y$ il y ait un ordre sur l'ensemble $N(v)$ des sommets adjacents à v, $N(v) = \{z_1 > z_2 > \ldots > z_{d(v)}\}$. Ainsi, z_1 est le meilleur choix pour v, ensuite c'est z_2, etc.

Définition 2. Un couplage M de $G = (X \cup Y, E)$ est dit *stable* si la condition suivante est réalisée : chaque fois que $uv \in E\backslash M$, $u \in X$, $v \in Y$, soit $uy \in M$ avec $y > v$ dans $N(u)$ soit $xv \in M$ avec $x > u$ dans $N(v)$, soit les deux.

Avec notre interprétation dans la vie réelle, un ensemble de mariages est stable s'il n'arrive jamais que u et v ne soient pas mariés et que u préfère v à sa partenaire (s'il en a une) et que v préfère u à son partenaire (si elle en a un), ce qui constituerait évidemment une situation instable.

Avant de montrer notre second résultat, penchons-nous sur l'exemple suivant :



Les arêtes en gras constituent un couplage stable. Dans chaque liste de priorités, le choix qui mène à un couplage stable est imprimé en gras.

Notons qu'il y a dans cet exemple un unique couplage maximal M à quatre arêtes, $M = \{aC, bB, cD, dA\}$, mais que M n'est pas stable (considérer cA).

Lemme 2. *Il existe toujours un couplage stable.*

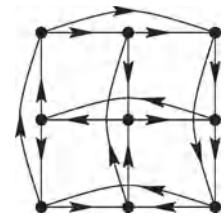
■ **Preuve.** Considérons l'algorithme suivant : dans une première étape, tous les hommes $u \in X$ demandent en mariage leur premier choix. Si une fille reçoit plusieurs proposition, elle garde le candidat qu'elle préfère auprès d'elle et si elle reçoit une seule proposition, elle garde le (seul) candidat auprès d'elle. Les hommes qui restent sont rejetés et forment le réservoir R. Dans la deuxième étape, tous les hommes de R demandent en mariage celle qu'ils ont placées en deuxième position. Les femmes comparent les propositions (avec celui qui est auprès d'elle, s'il y en a un), en choisissent un et le gardent auprès d'elles. Ceux qui restent sont rejetés et forment un nouvel ensemble R. Les hommes de R font encore une fois une demande en mariage et ainsi de suite. Si un homme fait une demande en mariage lors de son dernier choix et qu'il est encore rejeté, alors cet homme est écarté pour toute considération future, (il est aussi écarté du réservoir). Au bout d'un certain temps, le réservoir R est évidemment vide ; c'est à ce moment que s'arrête l'algorithme.

Proposition. *Lorsque l'algorithme s'arrête, les hommes qui se trouvent auprès d'une femme forment avec elles un couplage stable.*

Notons d'abord que les hommes choisis par une fille particulière se succèdent auprès d'elle dans l'ordre de ses préférences croissantes puisqu'à chaque étape, la fille compare les nouvelles propositions avec le partenaire courant et choisit son nouveau favori. Par conséquent, si $uv \in E$ mais $uv \notin M$, alors soit u n'a jamais demandé la main de v et dans ce cas il a trouvé une meilleure partenaire avant même de s'intéresser à v, ce qui implique que $uy \in M$ avec $y > v$ dans $N(u)$, soit u a demandé la main de v mais il a été rejeté, ce qui implique que $xv \in M$ avec $x > u$ dans $N(v)$. C'est exactement la condition pour que le couplage soit stable. □

En rassemblant les lemmes 1 et 2, on obtient la solution de Galvin au problème de Dinitz.

1	2	3
3	1	2
2	3	1



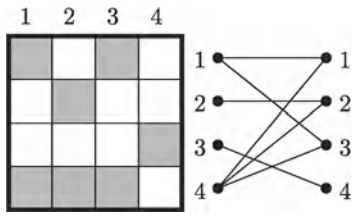
Théorème. *On a $\chi_\ell(S_n) = n$ pour tout n.*

■ **Preuve.** Comme précédemment, on note (i, j), $1 \leqslant i, j \leqslant n$, les sommets de S_n. (i, j) et (r, s) sont donc adjacents si et seulement si $i = r$ ou

$j = s$. Considérons un carré latin L dont les valeurs parcourent $\{1, 2, \ldots, n\}$ et notons $L(i,j)$ l'entrée de la cellule (i,j). Ensuite, faisons de S_n un graphe orienté $\vec{S}_n$ en orientant les arêtes horizontales dans le sens $(i,j) \longrightarrow (i,j')$ si $L(i,j) < L(i,j')$ et les arêtes verticales dans le sens $(i,j) \longrightarrow (i',j)$ si $L(i,j) > L(i',j)$. Ainsi, on oriente horizontalement du plus petit élément vers le plus grand et verticalement du plus grand vers le plus petit. (voir exemple avec $n = 3$ dans la marge.)

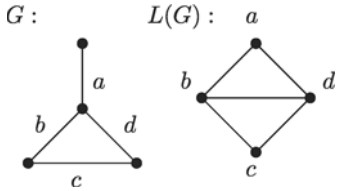
Notons que $d^+(i,j) = n-1$ pour tout (i,j). En effet, si $L(i,j) = k$, alors $n-k$ cellules de la ligne i contiennent un coefficient plus grand que k ; $k-1$ cellules de la colonne j ont donc un coefficient plus petit que k.

D'après le lemme 1 il reste à montrer que chaque sous-graphe induit de $\vec{S}_n$ possède un noyau. Considérons un sous-ensemble $A \subseteq V$, soit X l'ensemble des lignes de L et Y l'ensemble de ses colonnes. Associons à A le graphe biparti $G = (X \cup Y, A)$, où chaque $(i,j) \in A$ est représenté par l'arête ij telle que $i \in X, j \in Y$. Dans l'exemple de la marge, les cellules de A sont ombrées.



L'orientation de S_n induit naturellement un ordre sur les voisinages dans $G = (X \cup Y, A)$: on dit que $j' > j$ dans $N(i)$ si $(i,j) \longrightarrow (i,j')$ dans $\vec{S}_n$ (respectivement $i' > i$ dans $N(j)$ si $(i,j) \longrightarrow (i',j)$). Selon le lemme 2, $G = (X \cup Y, A)$ possède un couplage stable M. Cet ensemble M, considéré comme sous-ensemble de A, est le noyau recherché ! Pour comprendre pourquoi, notons d'abord que M est indépendant dans A puisqu'en tant que famille d'arêtes dans $G = (X \cup Y, A)$ celles-ci n'ont pas d'extrémités i ou j en commun. Ensuite, par définition d'un couplage stable, si $(i,j) \in A \backslash M$, il existe soit $(i,j') \in M$ tels que $j' > j$ soit $(i',j) \in M$ tels que $i' > i$, ce qui signifie pour $\vec{S}_n$ que $(i,j) \longrightarrow (i,j') \in M$ ou que $(i,j) \longrightarrow (i',j) \in M$. La preuve est terminée. □

Pour finir, allons un peu plus loin. Le lecteur a pu remarquer qu'une construction simple permet d'obtenir le graphe S_n à partir d'un graphe biparti. Considérons le graphe biparti complet, noté $K_{n,n}$, tel que $|X| = |Y| = n$, et qui contient *toutes* les arêtes entre X et Y. Si nous considérons les arêtes de $K_{n,n}$ comme des sommets d'un nouveau graphe, en joignant deux tels sommets si et seulement s'ils ont une extrémité commune, en tant qu'arêtes dans $K_{n,n}$, il est clair que nous obtenons le graphe carré S_n. Nous dirons que S_n est le *graphe des arcs* de $K_{n,n}$. On peut faire cette construction sur n'importe quel graphe G et appeler le graphe qui en résulte *graphe des arcs* $L(G)$ de G.



Construction d'un graphe des arcs.

En général, on dit que H est un *graphe d'arcs* si $H = L(G)$ pour un certain graphe G. Bien sûr, tous les graphes ne sont pas des graphes d'arcs, un exemple étant le graphe $K_{2,4}$ que nous avons déjà étudié. Nous avons vu que ce graphe vérifie $\chi(K_{2,4}) < \chi_\ell(K_{2,4})$. Mais qu'en est-il si H *est* un graphe d'arcs ? En adaptant la preuve du théorème, on peut montrer facilement que l'on a $\chi(H) = \chi_\ell(H)$ chaque fois que H est le graphe des arcs d'un graphe *biparti* et la méthode pourrait bien être utile pour vérifier la conjecture suprême dans ce domaine :

A-t-on $\chi(H) = \chi_\ell(H)$ *pour tout graphe d'arcs* H *?*

On sait peu de choses sur cette conjecture et le problème semble difficile — mais après tout, on savait peu de choses concernant le problème de Dinitz il y a vingt ans...

Bibliographie

[1] P. ERDŐS, A. L. RUBIN & H. TAYLOR : *Choosability in graphs,* Proc. West Coast Conference on Combinatorics, Graph Theory and Computing, Congressus Numerantium **26** (1979), 125-157.

[2] D. GALE & L. S. SHAPLEY : *College admissions and the stability of marriage,* Amer. Math. Monthly **69** (1962), 9-15.

[3] F. GALVIN : *The list chromatic index of a bipartite multigraph,* J. Combinatorial Theory, Ser. B **63** (1995), 153-158.

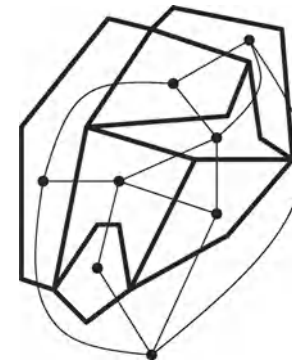
[4] J. C. M. JANSSEN : *The Dinitz problem solved for rectangles,* Bulletin Amer. Math. Soc. **29** (1993), 243-249.

[5] V. G. VIZING : *Coloring the vertices of a graph in prescribed colours (in Russian),* Metody Diskret. Analiz. **101** (1976), 3-10.

Cinq-coloration des graphes planaires

Chapitre 34

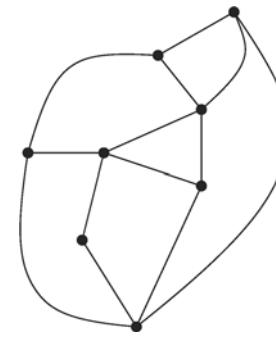
La coloration des graphes planaires a fait l'objet d'une recherche intensive depuis les débuts de la théorie des graphes à cause de son lien avec le problème des quatre couleurs. À l'origine, le problème des quatre couleurs posait la question de savoir s'il est toujours possible de colorier les régions d'une carte plane avec quatre couleurs de sorte que les régions qui ont un bord en commun (et pas seulement un point) soient de couleurs différentes. La figure montre que colorier les régions d'un plan revient à colorier les sommets d'un graphe planaire. Comme au chapitre 12 (page 83), on place un sommet à l'intérieur de chaque région (en considérant aussi la région extérieure) et l'on relie deux sommets qui appartiennent à des régions voisines par une arête qui traverse leur frontière.



Le graphe dual d'une carte.

Le graphe G qui en résulte, le *graphe dual* de la carte M, est donc un graphe planaire. Colorier les sommets de G au sens usuel revient à colorier les régions de M. On peut donc aussi bien se concentrer sur la coloration des sommets des graphes planaires. On peut supposer que G n'a ni boucles ni arêtes multiples, puisque cet aspect est sans rapport avec les problèmes de coloration.

Dans l'histoire longue et ardue des différentes tentatives pour démontrer le théorème des quatre couleurs, beaucoup se sont approchées du but, mais c'est finalement la combinaison de très vieilles idées datant du 19ème siècle et de la puissance de calcul des ordinateurs modernes (dans la preuve de Appel-Haken de 1976 et aussi dans celle plus récente de Robertson, Sanders, Seymour et Thomas de 1997) qui fut finalement couronnée de succès. Vingt-cinq ans après la preuve originale, la situation est toujours à peu près la même, aucune preuve du Grand Livre n'est en vue.



Ce graphe planaire a 8 sommets, 13 arêtes et 7 régions.

Soyons donc plus modestes et demandons-nous s'il existe une preuve claire du fait que tout graphe planaire est 5-colorable. Au tournant du vingtième siècle, Heawood avait déjà donné une démonstration de ce théorème des cinq couleurs. L'élément fondamental de sa preuve (qui est en fait aussi celle du théorème des quatre couleurs) est la formule d'Euler (voir chapitre 12). Lorsque l'on colorie un graphe G, il est clair que l'on peut supposer G connexe puisque l'on peut colorier les composantes connexes du graphe séparément. Un graphe planaire divise le plan en un ensemble R de régions (parmi lesquelles il faut faire figurer la région « extérieure »). La formule d'Euler affirme qu'étant donné un graphe planaire connexe $G = (V, E)$ on a toujours :

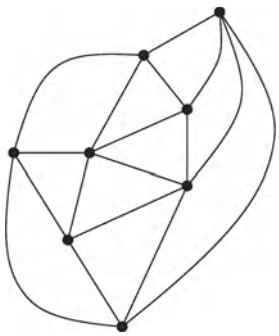
$$|V| - |E| + |R| \ = \ 2.$$

Pour nous échauffer, regardons comment la formule d'Euler pourrait être appliquée pour montrer que chaque graphe planaire G admet une 6-coloration. On procède par récurrence sur le nombre n de sommets. Pour de petites valeurs de n (en particulier, pour $n \leqslant 6$) le résultat est évident. En utilisant la partie (A) de la proposition de la page 85 on sait que G possède un sommet v de degré au plus 5. Effaçons v et toutes les arêtes incidentes à v. Le graphe $G' = G\backslash\{v\}$ qui en résulte est un graphe planaire à $n-1$ sommets. D'après l'hypothèse de récurrence, G' est 6-colorable. Puisque v a au plus 5 voisins dans G, 5 couleurs au plus sont utilisées pour ses voisins dans la coloration de G'. On peut donc étendre toute 6-coloration de G' en une 6-coloration de G en attribuant à v une couleur qui n'est utilisée pour aucun de ses voisins dans la coloration de G'. Ainsi, G admet une 6-coloration.

Examinons maintenant le *nombre chromatique listé* des graphes planaires, dont nous avons parlé dans le chapitre consacré au problème de Dinitz. Il est clair que notre méthode de 6-coloration fonctionne aussi bien pour des listes de couleurs (on ne manque jamais de couleurs) et ainsi $\chi_\ell(G) \leqslant 6$ pour tout graphe planaire G. Erdős, Rubin et Taylor ont conjecturé en 1979 que chaque graphe planaire avait un nombre chromatique listé au plus égal à 5, et qu'il existe en outre des graphes planaires G tels que $\chi_\ell(G) > 4$. Ils avaient raison sur les deux points. Margit Voigt a été la première à construire un exemple de graphe planaire G tel que $\chi_\ell(G) = 5$ (son exemple avait 238 sommets) et dans le même temps, Carsten Thomassen a donné une preuve absolument incroyable de la conjecture pour les colorations 5-listées. Sa preuve est un exemple révélateur de ce que l'on peut faire lorsque l'on trouve la bonne hypothèse de récurrence. Celle-ci n'utilise pas du tout la formule d'Euler !

Théorème. *Tous les graphes planaires G admettent des colorations 5-listées :*

$$\chi_\ell(G) \leqslant 5.$$



Un graphe planaire quasi-triangulé.

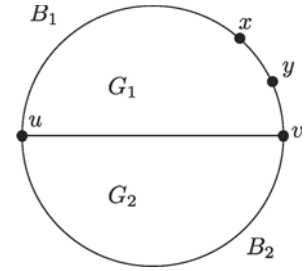
■ **Preuve.** Notons d'abord que l'addition d'arêtes augmente seulement le nombre chromatique. En d'autres termes, lorsque H est un sous-graphe de G, alors on a clairement $\chi_\ell(H) \leqslant \chi_\ell(G)$. Par conséquent, nous pouvons supposer que G est connexe et que toutes les faces bornées d'un plongement ont des triangles comme bords (et que la face non bornée est aussi fermée par un circuit simple qui peut être un triangle ou peut être formé de plus de trois arêtes). Nous disons qu'un tel graphe est *quasi-triangulé*. La validité du théorème pour les graphes quasi-triangulés établira le résultat pour tous les graphes planaires .

L'astuce de la preuve consiste à montrer l'énoncé suivant qui est plus fort et qui permet de recourir à un raisonnement par récurrence :

Soit $G = (V, E)$ un graphe quasi-triangulé et soit B le cycle qui limite la région extérieure. Faisons les hypothèses suivantes sur les

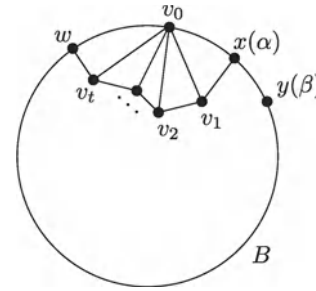
ensembles de couleurs $C(v)$, $v \in V$:
(1) *deux sommets adjacents x, y de B sont déjà coloriés avec des couleurs (différentes) α et β.*
(2) $|C(v)| \geqslant 3$ *pour tous les autres sommets v de B.*
(3) $|C(v)| \geqslant 5$ *pour tous les sommets v intérieurs.*
Alors, la coloration de x, y peut être étendue à une coloration propre de G en choisissant les couleurs dans les listes. En particulier, $\chi_\ell(G) \leqslant 5$.

Si $|V| = 3$ c'est évident, puisque le seul sommet v qui n'est pas colorié vérifie $|C(v)| \geqslant 3$; il y a donc une couleur disponible. Nous procédons maintenant par récurrence.



Cas 1: Supposons que B ait une corde, c'est-à-dire une arête qui n'est pas dans B et qui joint deux sommets $u, v \in B$. Le sous-graphe G_1 qui est borné par $B_1 \cup \{uv\}$ et qui contient x, y, u et v est quasi-triangulé ; par conséquent, par récurrence, il admet une coloration 5-listée. Supposons que dans cette coloration les sommets u et v reçoivent les couleurs γ et δ. Examinons la partie basse de G_2 qui est bornée par B_2 et uv. Comme u, v sont déjà coloriés, les hypothèses de récurrence sont également satisfaites par G_2. Par conséquent G_2 admet une coloration 5-listée avec les couleurs disponibles ; il en est donc de même pour G.

Cas 2: Supposons que B n'ait pas de corde. Soit v_0 le sommet qui se trouve sur B de l'autre côté du sommet α-colorié x et soient $x, v_1, \ldots, v_t, w$ les voisins de v_0. Puisque G est quasi-triangulé, on se trouve dans la situation montrée dans la figure.



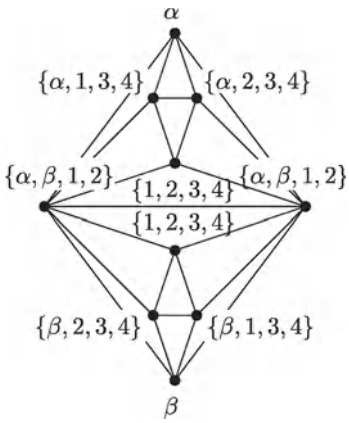
Construisons le graphe quasi-triangulé $G' = G \backslash \{v_0\}$ en supprimant de G le sommet v_0 et toutes les arêtes issues de v_0. G' admet $B' = (B \backslash \{v_0\}) \cup \{v_1, \ldots, v_t\}$ pour bord extérieur. Puisque $|C(v_0)| \geqslant 3$ d'après l'hypothèse (2), il existe deux couleurs γ et δ de $C(v_0)$ différentes de α. Remplaçons maintenant chaque ensemble de couleurs $C(v_i)$ par $C(v_i) \backslash \{\gamma, \delta\}$ en gardant les ensembles de couleurs originales pour tous les autres sommets de G'. Ainsi, G' vérifie clairement toutes les hypothèses. Il admet donc par récurrence une coloration 5-listée. En choisissant γ ou δ pour v_0, différent de la couleur de w, on peut étendre la coloration listée de G' à G tout entier. □

Ainsi, le théorème de coloration 5-listée est démontré. Toutefois, l'histoire n'est pas totalement terminée. Une conjecture plus forte affirmait que le nombre chromatique listé d'un graphe planaire G est au plus supérieur d'une unité au nombre chromatique ordinaire :

Est-ce que $\chi_\ell(G) \leqslant \chi(G) + 1$ pour tout graphe planaire G ?

Puisque $\chi(G) \leqslant 4$, d'après le théorème des quatre couleurs, nous avons trois cas :

Cas I : $\chi(G) = 2 \implies \chi_\ell(G) \leqslant 3$,
Cas II : $\chi(G) = 3 \implies \chi_\ell(G) \leqslant 4$,
Cas III : $\chi(G) = 4 \implies \chi_\ell(G) \leqslant 5$.



Le résultat de Thomassen résout le cas III ; le cas I a été démontré par un argument ingénieux (et très sophistiqué) d'Alon et Tarsi. En outre, il y a des graphes planaires G tels que $\chi(G) = 2$ et $\chi_\ell(G) = 3$, par exemple le graphe $K_{2,4}$ que nous avons étudié dans le chapitre consacré au problème de Dinitz.

Qu'en est-il du cas II ? Ici, la conjecture est fausse : cela a d'abord été montré par Margit Voigt sur un graphe qu'avait construit auparavant Shai Gutner. Son graphe sur 130 sommets peut être obtenu de la manière suivante : considérons tout d'abord le « double octaèdre » (voir figure), qui admet clairement une 3-coloration. Soient $\alpha \in \{5,6,7,8\}$ et $\beta \in \{9,10,11,12\}$; considérons les listes qui sont données dans la figure. On peut vérifier qu'aucune coloration n'est possible avec ces listes. Prenons maintenant 16 copies de ce graphe ; identifions tous les sommets du haut et tous ceux du bas. Cela donne un graphe sur $16 \cdot 8 + 2 = 130$ sommets qui est toujours planaire et qui admet une 3-coloration. Attribuons $\{5,6,7,8\}$ au sommet du haut et $\{9,10,11,12\}$ au sommet du bas, les listes intérieures correspondant aux 16 paires (α,β), $\alpha \in \{5,6,7,8\}$, $\beta \in \{9,10,11,12\}$. Pour chaque choix de α et β nous obtenons ainsi un sous-graphe (comme dans la figure). Une coloration par liste du grand graphe est donc impossible.

En modifiant un des autres exemples de Gutner, Voigt et Wirth ont trouvé un graphe planaire plus petit ayant 75 sommets, tel que $\chi = 3$ et $\chi_\ell = 5$, qui en outre utilise le nombre minimal de cinq couleurs dans les listes. Le record actuel est de 63 sommets.

Bibliographie

[1] N. ALON & M. TARSI : *Colorings and orientations of graphs,* Combinatorica **12** (1992), 125-134.

[2] P. ERDŐS, A. L. RUBIN & H. TAYLOR : *Choosability in graphs,* Proc. West Coast Conference on Combinatorics, Graph Theory and Computing, Congressus Numerantium **26** (1979), 125-157.

[3] S. GUTNER : *The complexity of planar graph choosability,* Discrete Math. **159** (1996), 119-130.

[4] N. ROBERTSON, D. P. SANDERS, P. SEYMOUR & R. THOMAS : *The four-colour theorem,* J. Combinatorial Theory, Ser. B **70** (1997), 2-44.

[5] C. THOMASSEN : *Every planar graph is* 5*-choosable,* J. Combinatorial Theory, Ser. B **62** (1994), 180-181.

[6] M. VOIGT : *List colorings of planar graphs,* Discrete Math. **120** (1993), 215-219.

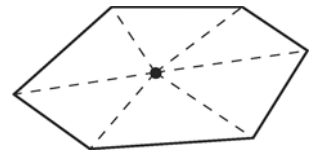
[7] M. VOIGT & B. WIRTH : *On* 3*-colorable non-*4*-choosable planar graphs,* J. Graph Theory **24** (1997), 233-235.

Comment surveiller un musée

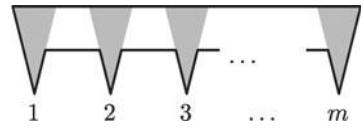
Chapitre 35

Voici un problème attrayant soulevé par Victor Klee en 1973. Le directeur d'un musée veut être sûr qu'à tout moment chaque endroit du musée est surveillé par un gardien. Les gardiens sont placés à des postes fixes, mais ils peuvent se tourner. Combien de gardiens sont nécessaires ?

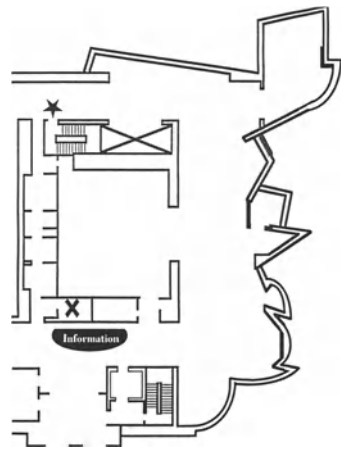
Dessinons les murs du musée comme un polygone à n côtés. Bien sûr, si le polygone est *convexe*, un gardien suffit et le gardien peut se trouver en n'importe quel point du musée. Mais, en général, les murs d'un musée peuvent avoir n'importe quelle forme de polygone fermé.



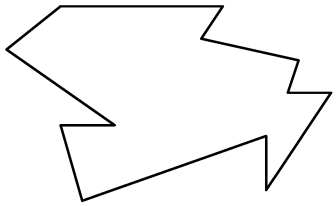
Un hall d'exposition convexe.



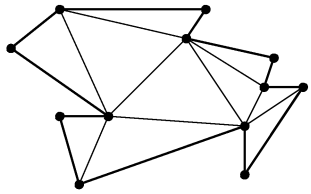
Considérons un musée en forme de peigne avec $n = 3m$ murs, comme celui représenté dans la marge. Il est facile de voir que ce dernier requiert au moins $m = \frac{n}{3}$ gardiens. En effet, il y a n murs. Remarquons que le point 1 ne peut être observé que par un gardien qui se trouve dans le triangle ombré qui contient 1 ; il en est de même pour les autres points $2, 3, \ldots, m$. Puisque tous ces triangles sont disjoints, au moins m gardiens sont nécessaires. Cependant, m gardiens suffisent également, puisqu'ils peuvent être placés sur les côtés supérieurs des triangles. En supprimant un ou deux murs à la fin, nous concluons que pour tout n il existe un musée à n murs qui requiert $\lfloor \frac{n}{3} \rfloor$ gardiens.



Une véritable galerie d'art...



Un musée avec $n = 12$ murs.



Une triangulation du musée ci-dessus.

Le résultat suivant affirme que c'est le pire des cas.

Théorème. *Pour tout musée à n murs, $\lfloor \frac{n}{3} \rfloor$ gardiens suffisent.*

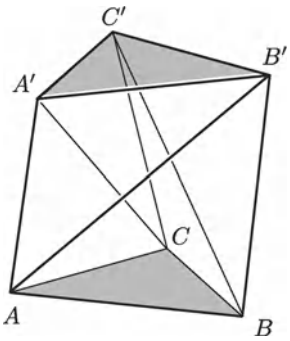
Ce « théorème de la galerie d'art » a d'abord été démontré par un ingénieux argument de Vašek Chvátal, mais voici une très belle preuve de Steve Fisk.

■ **Preuve.** Dessinons d'abord entre les coins des murs $n-3$ diagonales qui ne se coupent pas, de sorte que l'intérieur soit triangulé. Par exemple, nous pouvons tracer 9 diagonales dans le musée représenté ci-contre. La triangulation retenue n'a pas d'importance, n'importe laquelle convient. Considérons maintenant la nouvelle figure comme un graphe planaire dont les coins sont les sommets, et dont les murs et les diagonales sont les arêtes.

Proposition. *Ce graphe est 3-colorable.*

Pour $n = 3$ il n'y a rien à prouver. Pour $n > 3$, prenons deux sommets quelconques u et v reliés par une diagonale. Cette diagonale va diviser le graphe en deux graphes triangulés plus petits qui contiennent tous les deux l'arête uv. Par récurrence, nous pouvons colorer chaque partie avec 3 couleurs ; nous pouvons choisir la couleur 1 pour u et la couleur 2 pour v dans chaque coloration. En fusionnant les colorations, on obtient une 3-coloration du graphe tout entier.

Le reste est simple. Puisqu'il y a n sommets, une des classes de couleurs au moins, celle dont les sommets sont colorés par 1 par exemple, contient au plus $\lfloor \frac{n}{3} \rfloor$ sommets : c'est en ces points qu'il convient de placer les gardiens. Puisque chaque triangle contient un sommet de couleur 1, chaque triangle est gardé ; par conséquent il en est ainsi pour tout le musée. □



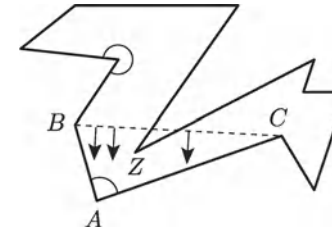
Le polyèdre de Schönhardt : Les angles diédraux intérieurs aux arêtes AB', BC' et CA' sont supérieurs à 180°.

Le lecteur attentif a pu relever un point délicat dans le raisonnement. Est-ce qu'il existe toujours une triangulation ? À première vue on peut penser que c'est le cas. Elle existe en effet, mais ce résultat n'est pas complètement évident. En fait, il ne se généralise pas en dimension trois (partition en tétraèdres) ! On peut s'en convaincre avec le *polyèdre de Schönhardt*, représenté dans la marge. Il est obtenu à partir d'un prisme triangulaire en faisant tourner le triangle du haut, de sorte que chacune des faces quadrilatérales se décompose en deux triangles avec une arête non convexe. On ne peut pas trianguler ce polyèdre ! Tout tétraèdre qui contient le triangle du bas doit contenir un des trois sommets du haut mais le tétraèdre qui en résulte n'est pas contenu dans le polyèdre de Schönhardt. Il n'y a donc pas de triangulation sans sommet supplémentaire.

Pour prouver qu'un polygone planaire non convexe admet une triangulation, on procède par récurrence sur le nombre n de sommets. Pour $n = 3$ le polygone est un triangle ; il n'y a rien à démontrer. Soit $n \geqslant 4$. Pour faire un raisonnement par récurrence, on doit seulement exhiber *une* diagonale

qui divisera le polygone P en deux parties plus petites qui pourront être fusionnées ensemble, comme on l'a fait plus haut.

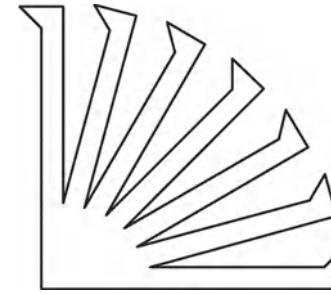
Nous disons qu'un sommet A est *convexe* si l'angle intérieur au sommet est inférieur à $180°$. Puisque la somme des angles intérieurs de P est égale à $(n-2)180°$, il doit y avoir au moins un sommet convexe A. En fait, il y en a au moins trois : c'est pour l'essentiel une application du principe des tiroirs ! On peut aussi considérer l'enveloppe convexe du polygone et remarquer que tous ses sommets sont convexes.



Examinons maintenant les deux sommets voisins B et C de A. Si le segment BC se trouve entièrement dans P, BC devient la diagonale cherchée. Sinon, le triangle ABC contient d'autres sommets. Faisons glisser BC vers A jusqu'à ce qu'il touche le dernier sommet Z dans ABC. AZ est maintenant à l'intérieur de P et constitue la diagonale cherchée.

Il existe plusieurs variantes du théorème de la galerie d'art. Nous pouvons par exemple imposer que seuls les murs soient gardés (après tout, c'est l'endroit où sont accrochés les tableaux), ou bien que les gardiens soient tous placés à des sommets. Une variante particulièrement belle du problème, à ce jour non résolue, est la suivante :

> *Supposons que chaque gardien puisse patrouiller le long d'un mur du musée, il marche donc le long de son mur et voit tout ce qui peut être vu de n'importe quel point de ce mur.*
> *De combien de « gardiens de murs » a-t-on besoin pour garder le contrôle du musée ?*



Gottfried Toussaint a construit l'exemple de musée figurant dans la marge qui montre que $\lfloor \frac{n}{4} \rfloor$ gardiens peuvent être nécessaires.

Ce polygone a 28 côtés (et en général $4m$ côtés) ; le lecteur est invité à vérifier que m gardiens placés sur les côtés sont nécessaires. On conjecture que, sauf pour quelques petites valeurs de n, ce nombre est également suffisant, mais une preuve, sans même penser à une Preuve du Grand Livre, n'est pas encore en vue.

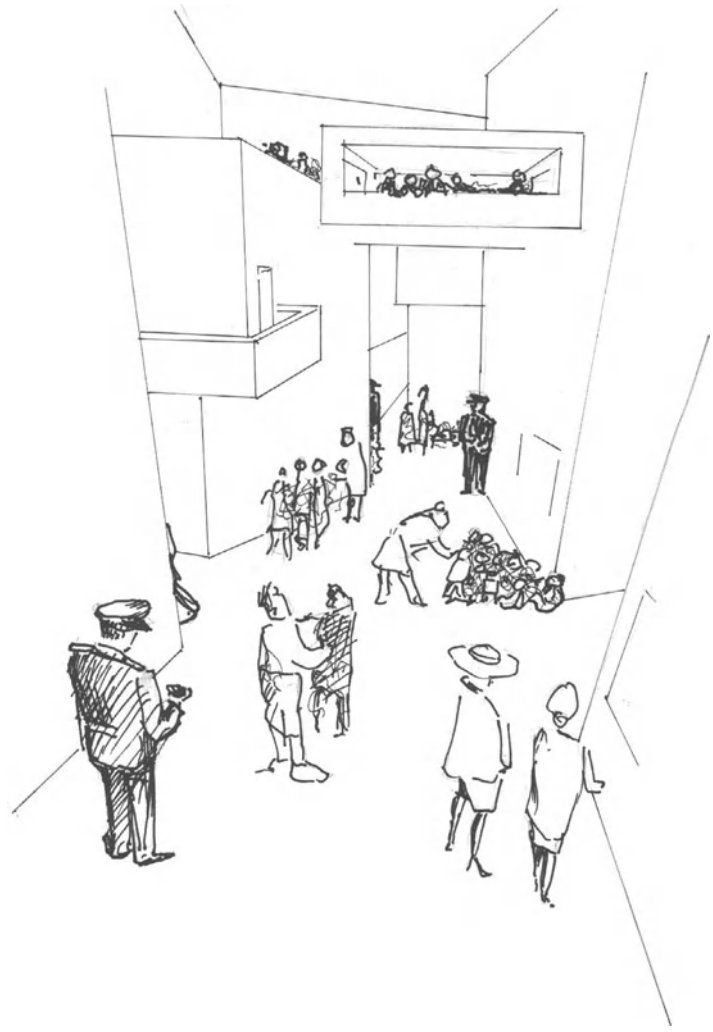
Bibliographie

[1] V. CHVÁTAL : *A combinatorial theorem in plane geometry,* J. Combinatorial Theory, Ser. B **18** (1975), 39-41.

[2] S. FISK : *A short proof of Chvátal's watchman theorem,* J. Combinatorial Theory, Ser. B **24** (1978), 374.

[3] J. O'ROURKE : *Art Gallery Theorems and Algorithms,* Oxford University Press 1987.

[4] E. SCHÖNHARDT : *Über die Zerlegung von Dreieckspolyedern in Tetraeder,* Math. Annalen **98** (1928), 309-312.



« Gardiens de musée ».
(Un problème tridimensionnel de galerie d'art).

Le théorème de Turán

Chapitre 36

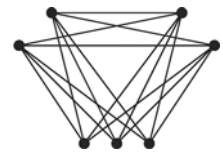
Un des résultats fondamentaux de la théorie des graphes est le théorème de Turán, datant de 1941, qui a lancé la théorie extrémale des graphes. Le théorème de Turán a été redécouvert plusieurs fois avec des preuves différentes. Nous traiterons cinq d'entre elles et laisserons le lecteur choisir celle qui selon lui doit figurer dans le Grand Livre.

Fixons quelques notations. Nous considérons des graphes simples G sur un ensemble de sommets $V = \{v_1, \ldots, v_n\}$ et un ensemble d'arêtes E. Si v_i et v_j sont voisins, on écrit $v_iv_j \in E$. Une *p-clique* de G est un sous-graphe complet de G à p sommets, noté K_p. Paul Turán a posé la question suivante :

Paul Turán

> *Supposons que G soit un graphe simple sans p-clique.*
> *Quel est le nombre maximal d'arêtes que G peut contenir ?*

On obtient facilement des exemples de tels graphes en partageant V en $p-1$ sous-ensembles deux à deux disjoints $V = V_1 \cup \ldots \cup V_{p-1}$, $|V_i| = n_i$, $n = n_1 + \ldots + n_{p-1}$ et en joignant deux sommets si et seulement s'ils se trouvent dans des ensembles distincts V_i, V_j. Notons $K_{n_1,\ldots,n_{p-1}}$ le graphe qui en résulte ; il contient $\sum_{i<j} n_in_j$ arêtes. n étant fixé, on obtient un nombre maximum d'arêtes dans de tels graphes si l'on divise les nombres n_i autant que faire se peut, c'est-à-dire si $|n_i - n_j| \leqslant 1$ pour tous i, j. En effet, supposons que $n_1 \geqslant n_2 + 2$. En déplaçant un sommet de V_1 à V_2, on obtient $K_{n_1-1,n_2+1,\ldots,n_{p-1}}$ qui contient $(n_1 - 1)(n_2 + 1) - n_1n_2 = n_1 - n_2 - 1 \geqslant 1$ arêtes de plus que $K_{n_1,n_2,\ldots,n_{p-1}}$. Appelons *graphes de Turán* les graphes $K_{n_1,\ldots,n_{p-1}}$ tels que $|n_i - n_j| \leqslant 1$. Remarquons que si $p - 1$ divise n, on peut choisir $n_i = \frac{n}{p-1}$ pour tout i, obtenant ainsi :



Le graphe $K_{2,2,3}$.

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

arêtes. Le théorème de Turán affirme alors que ce nombre est une borne supérieure du nombre d'arêtes de *tout* graphe à n sommets sans p-clique.

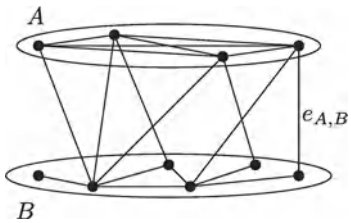
Théorème. *Si un graphe $G = (V, E)$ à n sommets n'a pas de p-clique, $p \geqslant 2$, alors :*

$$|E| \leqslant \Big(1 - \frac{1}{p-1}\Big)\frac{n^2}{2}. \qquad (1)$$

Pour $p = 2$ c'est trivial. Dans le premier cas intéressant, c'est-à-dire pour $p = 3$, le théorème affirme qu'un graphe sans triangle à n sommets contient au plus $\frac{n^2}{4}$ arêtes. On connaissait des preuves de ce cas particulier avant le résultat de Turán. Deux preuves élégantes utilisant des inégalités figurent au chapitre 18.

Penchons nous sur le cas général. Les deux premières preuves utilisent un raisonnement par récurrence ; on les doit respectivement à Turán et Erdős.

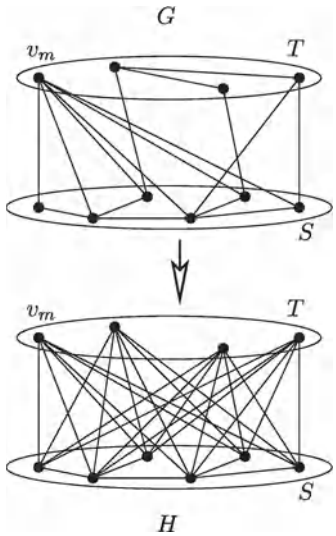
■ **Première preuve.** On procède par récurrence sur n. On vérifie facilement que (1) est vraie pour $n \leqslant p-1$. Supposons maintenant que $n \geqslant p$ et soit G un graphe sur $V = \{v_1, \ldots, v_n\}$ sans p-cliques et avec un nombre maximal d'arêtes. G contient certainement des $(p-1)$-cliques puisque sinon, on pourrait ajouter des arêtes. Soit A une $(p-1)$-clique ; posons $B := V \backslash A$.



A contient $\binom{p-1}{2}$ arêtes. Nous allons maintenant estimer le nombre d'arêtes e_B de B et le nombre d'arêtes $e_{A,B}$ entre A et B. Par récurrence, nous avons $e_B \leqslant \frac{1}{2}(1 - \frac{1}{p-1})(n-p+1)^2$. Puisque G n'a pas de p-clique, chaque $v_j \in B$ est adjacent à $p-2$ sommets de A au plus ; nous obtenons donc $e_{A,B} \leqslant (p-2)(n-p+1)$. Finalement, cela implique :

$$|E| \leqslant \binom{p-1}{2} + \frac{1}{2}\Big(1 - \frac{1}{p-1}\Big)(n-p+1)^2 + (p-2)(n-p+1)$$

qui est précisément égal à $(1 - \frac{1}{p-1})\frac{n^2}{2}$. □



■ **Deuxième preuve.** Cette preuve utilise la structure des graphes de Turán. Soit $v_m \in V$ un sommet de degré maximal $d_m = \max_{1 \leqslant j \leqslant n} d_j$. Notons S l'ensemble des voisins de v_m, $|S| = d_m$ et posons $T = V \backslash S$. Comme G ne contient pas de p-clique et que v_m est adjacent à tous les sommets de S, S ne contient pas de $(p-1)$-clique.

Construisons maintenant le graphe H sur V suivant (voir figure). H correspond à G sur S et contient toutes les arêtes entre S et T, mais pas d'arêtes à l'intérieur de T. En d'autres termes, T est un ensemble indépendant dans H donc H n'a pas de p-clique. Soit d'_j le degré de v_j dans H. Si $v_j \in S$, alors $d'_j \geqslant d_j$ par construction de H ainsi pour $v_j \in T$, on voit que $d'_j = |S| = d_m \geqslant d_j$ grâce au choix de v_m. Il en résulte que $|E(H)| \geqslant |E|$, et que parmi tous les graphes ayant un nombre maximal d'arêtes, il doit y en avoir un qui a la forme de H. Par récurrence, le graphe induit par S a au plus autant d'arêtes qu'un graphe convenable $K_{n_1,\ldots,n_{p-2}}$ sur S. Donc $|E| \leqslant |E(H)| \leqslant E(K_{n_1,\ldots,n_{p-1}})$ où $n_{p-1} = |T|$, ce qui implique (1). □

Les deux preuves suivantes sont de natures complètement différentes. Elles utilisent un argument de maximum et des idées issues de la théorie des probabilités. On les doit respectivement à Motzkin et Strauss et à Alon et Spencer.

■ **Troisième preuve.** Considérons une *distribution de probabilité* $\boldsymbol{w} = (w_1, \ldots, w_n)$ sur les sommets : attribuons aux sommets des valeurs $w_i \geqslant 0$ telles que $\sum_{i=1}^{n} w_i = 1$. Notre but est de maximiser la fonction :

$$f(\boldsymbol{w}) \;=\; \sum_{v_i v_j \in E} w_i w_j$$

Supposons que $\boldsymbol{w}$ soit une distribution quelconque ; soient v_i et v_j deux sommets non-adjacents auxquels on a attribué des valeurs positives w_i, w_j. Soit s_i la somme des valeurs de tous les sommets adjacents à v_i. Définissons s_j de la même manière pour v_j, en ayant supposé que $s_i \geqslant s_j$. Déplaçons maintenant le poids de v_j à v_i, c'est-à-dire que le nouveau poids attribué à v_i est $w_i + w_j$, tandis que le poids attribué à v_j est égal à zero . La nouvelle distribution $\boldsymbol{w}'$ vérifie :

$$f(\boldsymbol{w}') \;=\; f(\boldsymbol{w}) + w_j s_i - w_j s_j \;\geqslant\; f(\boldsymbol{w})$$

« Déplacement de poids ».

En répétant cette opération (on obtient moins de sommets à valeurs positives à chaque étape) on constate qu'il y a une distribution optimale telle que les poids non nuls soient concentrés sur une clique, par exemple une k-clique. Supposons que $w_1 > w_2 > 0$, et choisissons ε tel que $0 < \varepsilon < w_1 - w_2$. Changeons w_1 en $w_1 - \varepsilon$ et w_2 en $w_2 + \varepsilon$. La nouvelle distribution $\boldsymbol{w}'$ vérifie $f(\boldsymbol{w}') = f(\boldsymbol{w}) + \varepsilon(w_1 - w_2) - \varepsilon^2 > f(\boldsymbol{w})$. La valeur maximum de f est donc atteinte pour $w_i = \frac{1}{k}$ sur une k-clique et $w_i = 0$ ailleurs. Puisqu'une k-clique contient $\frac{k(k-1)}{2}$ arêtes :

$$f \;=\; \frac{k(k-1)}{2}\frac{1}{k^2} \;=\; \frac{1}{2}\Big(1 - \frac{1}{k}\Big)$$

Comme cette expression croît avec k, le mieux que nous puissions faire est de poser $k = p - 1$ (puisque G n'a pas de p-cliques). Ainsi :

$$f(\boldsymbol{w}) \;\leqslant\; \frac{1}{2}\Big(1 - \frac{1}{p-1}\Big)$$

pour *toute* distribution $\boldsymbol{w}$. En particulier, cette inégalité a lieu pour la distribution *uniforme* définie par $w_i = \frac{1}{n}$ pour tout i. On trouve ainsi :

$$\frac{|E|}{n^2} \;=\; f\Big(w_i = \frac{1}{n}\Big) \;\leqslant\; \frac{1}{2}\Big(1 - \frac{1}{p-1}\Big),$$

ce qui est précisément (1). □

■ **Quatrième preuve.** Nous utilisons cette fois encore quelques concepts de la théorie des probabilités. Soit G un graphe arbitraire sur l'ensemble des sommets $V = \{v_1, \ldots, v_n\}$. Notons d_i le degré de v_i et $\omega(G)$ le nombre de sommets d'une clique maximale, appelé le *nombre de clique* de G.

Proposition. *On a* $\omega(G) \;\geqslant\; \displaystyle\sum_{i=1}^{n} \frac{1}{n - d_i}$.

Choisissons au hasard, avec équiprobabilité $\frac{1}{n!}$, une permutation π de l'ensemble des sommets. À réindexation près, on peut noter $\pi = v_1 v_2 \dots v_n$ cette permutation. Construisons l'ensemble C_π suivant : on met v_i dans C_π si et seulement si v_i est adjacent à tous les v_j $(j < i)$ qui précèdent v_i. Par définition, C_π est une clique dans G. Soit $X = |C_\pi|$ la variable aléatoire correspondante. Nous avons $X = \sum_{i=1}^{n} X_i$, où X_i est la variable aléatoire indicatrice du sommet v_i, c'est-à-dire que $X_i = 1$ ou $X_i = 0$ suivant que $v_i \in C_\pi$ ou $v_i \notin C_\pi$. Notons que v_i appartient à C_π pour la permutation $v_1 v_2 \dots v_n$ si et seulement si v_i apparaît *avant* tous les $n - 1 - d_i$ sommets qui ne sont pas adjacents à v_i, en d'autres termes, si v_i est le *premier* parmi v_i et ses $n - 1 - d_i$ non-voisins. La probabilité que cela se produise est $\frac{1}{n-d_i}$; par conséquent $EX_i = \frac{1}{n-d_i}$. Par linéarité de l'espérance (voir page 107) on obtient alors :

$$E(|C_\pi|) \;=\; EX \;=\; \sum_{i=1}^{n} EX_i \;=\; \sum_{i=1}^{n} \frac{1}{n - d_i}$$

Il doit donc y avoir une clique de cette taille au moins, ce que nous affirmions. Pour en déduire le théorème de Turán, on va utiliser l'inégalité de Cauchy-Schwarz évoquée au chapitre 18,

$$\Big(\sum_{i=1}^{n} a_i b_i\Big)^2 \;\leqslant\; \Big(\sum_{i=1}^{n} a_i^2\Big)\Big(\sum_{n=1}^{n} b_i^2\Big)$$

On pose $a_i = \sqrt{n - d_i}$, $b_i = \frac{1}{\sqrt{n-d_i}}$, ainsi $a_i b_i = 1$ donc :

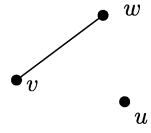
$$n^2 \;\leqslant\; (\sum_{i=1}^{n}(n - d_i))(\sum_{i=1}^{n} \frac{1}{n - d_i}) \;\leqslant\; \omega(G) \sum_{i=1}^{n}(n - d_i) \qquad (2)$$

Appliquons à présent l'hypothèse $\omega(G) \leqslant p - 1$ du théorème de Turán. En utilisant aussi le fait que $\sum_{i=1}^{n} d_i = 2|E|$ d'après le chapitre sur le double décompte, l'inégalité (2) implique :

$$n^2 \;\leqslant\; (p - 1)(n^2 - 2|E|)$$

ce qui est équivalent à l'inégalité de Turán. □

Nous sommes maintenant fin prêts pour la dernière preuve, qui est peut-être la plus belle de toutes. Son origine n'est pas claire : elle nous a été communiquée par Stephan Brandt, qui l'a entendue à Oberwolfach. Elle doit venir du folklore de la théorie des graphes. Elle montre simultanément que le graphe de Turán est en fait l'unique exemple qui possède un nombre maximal d'arêtes (remarquons que les deux preuves 1 et 2 impliquent aussi ce résultat plus fort).

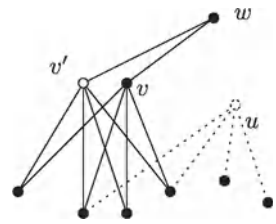


■ **Cinquième preuve.** Soit G un graphe sur n sommets sans p-clique, ayant un nombre maximal d'arêtes.

> ***Proposition.*** *G ne peut pas contenir trois sommets u, v, w tels que $vw \in E$, mais tels que ni $uv \notin E$, ni $uw \notin E$.*

Supposons le contraire et considérons les cas suivants.

Cas 1: $d(u) < d(v)$ ou $d(u) < d(w)$.
On suppose par exemple que $d(u) < d(v)$. On duplique v, c'est-à-dire que l'on crée un nouveau sommet v' qui a exactement les mêmes voisins que v (mais vv' n'est pas une arête), on efface u et l'on garde le reste tel quel.

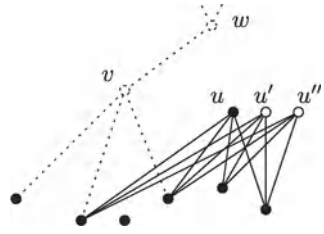


Le nouveau graphe G' n'a toujours pas de p-clique et son nombre d'arêtes vérifie :

$$|E(G')| \;=\; |E(G)| + d(v) - d(u) \;>\; |E(G)|$$

ce qui est contradictoire.

Cas 2: $d(u) \geqslant d(v)$ et $d(u) \geqslant d(w)$.
On duplique u deux fois et l'on efface v et w (comme indiqué dans la marge). Une fois de plus, le graphe G' n'a pas de p-clique et l'on a (le -1 résulte de l'arête vw) :



$$|E(G')| \;=\; |E(G)| + 2d(u) - (d(v) + d(w) - 1) \;>\; |E(G)|$$

On obtient encore une contradiction.

Un peu de réflexion montre que l'assertion démontrée est équivalente au fait que la relation :

$$u \sim v \;\;:\Longleftrightarrow\;\; uv \notin E(G)$$

est une relation d'équivalence. Ainsi, G est un graphe multiparti complet, $G = K_{n_1,\ldots,n_{p-1}}$, ce qui termine la démonstration. □

Bibliographie

[1] M. AIGNER : *Turán's graph theorem,* Amer. Math. Monthly **102** (1995), 808-816.

[2] N. ALON & J. SPENCER : *The Probabilistic Method,* Wiley Interscience 1992.

[3] P. ERDŐS : *On the graph theorem of Turán (in Hungarian),* Math. Fiz. Lapok **21** (1970), 249-251.

[4] T. S. MOTZKIN & E. G. STRAUSS : *Maxima for graphs and a new proof of a theorem of Turán,* Canad. J. Math. **17** (1965), 533-540.

[5] P. TURÁN : *On an extremal problem in graph theory,* Math. Fiz. Lapok **48** (1941), 436-452.

« Des poids encore plus lourds à déplacer ».

Communiquer sans erreur

Chapitre 37

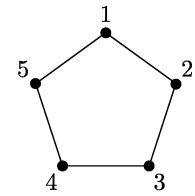
En 1956, Claude Shannon, le fondateur de la théorie de l'information, posa l'intéressante question suivante :

> *Supposons que l'on veuille transmettre des messages à un récepteur à travers un canal (dans lequel certains symboles peuvent être modifiés au cours de la transmission). Quel est le taux de transmission maximum qui permette au récepteur de retrouver le message original sans erreur ?*

Claude Shannon

Examinons ce que Shannon a voulu dire par « canal » et « taux de transmission ». On se donne un ensemble V de symboles. Un message est simplement une chaîne de symboles de V. Modélisons le canal comme un graphe $G = (V, E)$, où V est l'ensemble des symboles et E l'ensemble des arêtes entre des paires de symboles incertains, c'est-à-dire des symboles qui peuvent être confondus pendant la transmission. Par exemple, dans le cadre d'une communication par téléphone on relie les symboles B et P par une arête puisqu'il y a de bonnes chances que le récepteur les confonde. Appelons G le *graphe de confusion*.

Le 5-cycle C_5 joue un rôle primordial dans le développement qui suit. Ce graphe signifie qu'il y a un risque de confusion entre 1 et 2 mais pas entre 1 et 3 etc. Dans l'idéal, on aimerait utiliser les 5 symboles pour la transmission, mais, comme on veut communiquer sans erreur, on peut, à condition de n'envoyer que des symboles seuls, utiliser seulement une lettre pour chaque paire qui pourrait être confondue. Ainsi, en ce qui concerne le 5-cycle, on peut se restreindre à deux lettres différentes pourvu qu'elles ne soient pas reliées par une arête. Dans le langage de la théorie de l'information, cela signifie que l'on réalise pour le 5-cycle un taux d'information de $\log_2 2 = 1$ (au lieu du maximum $\log_2 5 \approx 2.32$). Dans ce modèle, il est clair que pour un graphe arbitraire $G = (V, E)$, le mieux que l'on puisse faire est de transmettre des symboles à partir d'un ensemble indépendant de taille maximale. Ainsi, en envoyant des symboles seuls, on atteint le taux d'information $\log_2 \alpha(G)$, où $\alpha(G)$ est le *nombre de stabilité* de G.



Voyons si l'on peut accroître le taux d'information en utilisant des chaînes plus grandes à la place de symboles seuls. Supposons que l'on veuille transmettre des chaînes de longueur 2. Les chaînes u_1u_2 et v_1v_2 peuvent être confondues dans l'un des trois cas suivants seulement :

- $u_1 = v_1$ et u_2 peut être confondu avec v_2,
- $u_2 = v_2$ et u_1 peut être confondu avec v_1,
- $u_1 \neq v_1$ peuvent être confondus et $u_2 \neq v_2$ peuvent être confondus.

En termes de théorie des graphes cela revient à considérer le *produit* $G_1 \times G_2$ des deux graphes $G_1 = (V_1, E_1)$ et $G_2 = (V_2, E_2)$. $G_1 \times G_2$ a pour ensemble de sommets $V_1 \times V_2 = \{(u_1, u_2) : u_1 \in V_1, u_2 \in V_2\}$, où $(u_1, u_2) \neq (v_1, v_2)$ sont reliés par une arête si et seulement si $u_i = v_i$ ou $u_i v_i \in E_i$ pour $i = 1, 2$. Le graphe de confusion des chaînes de longueur 2 est ainsi $G^2 = G \times G$, le produit du graphe avec lui-même. Le taux d'information *par symbole* déterminé par les chaînes de longueur 2 est donc :

$$\frac{\log_2 \alpha(G^2)}{2} \quad = \quad \log_2 \sqrt{\alpha(G^2)}$$

On peut, bien sûr, utiliser des chaînes de longueur n quelconque. Le n-ième graphe de confusion $G^n = G \times G \times \ldots \times G$ a pour ensemble de sommets $V^n = \{(u_1, \ldots, u_n) : u_i \in V\}$ où $(u_1, \ldots, u_n) \neq (v_1, \ldots v_n)$ sont reliés par une arête si $u_i = v_i$ ou si $u_i v_i \in E$ pour tout i. Le taux d'information par symbole déterminé par les chaînes de longueur n est :

$$\frac{\log_2 \alpha(G^n)}{n} \quad = \quad \log_2 \sqrt[n]{\alpha(G^n)}$$

Que dire sur $\alpha(G^n)$? Voici une première observation. Soit $U \subseteq V$ un ensemble indépendant maximal dans G, avec $|U| = \alpha$. Les α^n sommets de G^n de la forme $(u_1, \ldots, u_n)$, où $u_i \in U$ pour tout i, forment clairement un ensemble indépendant dans G^n. Par conséquent :

$$\alpha(G^n) \quad \geqslant \quad \alpha(G)^n$$

et donc :

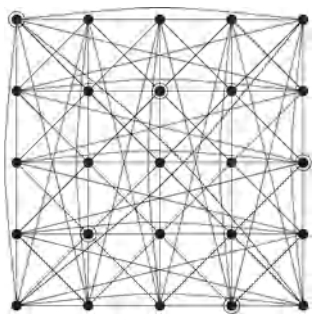
$$\sqrt[n]{\alpha(G^n)} \quad \geqslant \quad \alpha(G)$$

ce qui signifie que l'on ne diminue jamais le taux d'information en utilisant des chaînes de plusieurs symboles au lieu de chaînes de symboles seuls. C'est d'ailleurs une idée fondamentale de la théorie des codes : en codant des symboles dans des chaînes plus longues on peut établir une communication sans erreur plus efficace.

En ne tenant pas compte du logarithme on arrive à la définition fondamentale de Shannon : la *capacité d'erreur nulle* d'un graphe G est donnée par :

$$\Theta(G) \quad := \quad \sup_{n \geqslant 1} \sqrt[n]{\alpha(G^n)}$$

Le problème de Shannon a été de calculer $\Theta(G)$, en particulier $\Theta(C_5)$.



Le graphe $C_5 \times C_5$.

Examinons C_5. Nous savons jusqu'ici que $\alpha(C_5) = 2 \leqslant \Theta(C_5)$. En étudiant le 5-cycle représenté plus haut, ou le produit $C_5 \times C_5$ représenté ci-contre, on voit que l'ensemble $\{(1,1), (2,3), (3,5), (4,2), (5,4)\}$ est indépendant dans ${C_5}^2$. Ainsi, nous avons $\alpha({C_5}^2) \geqslant 5$. Puisqu'un ensemble indépendant ne peut contenir que deux sommets de deux lignes consécutives, $\alpha({C_5}^2) = 5$. Par conséquent, en utilisant des chaînes de longueur 2 nous avons augmenté la borne inférieure de la capacité jusqu'à $\Theta(C_5) \geqslant \sqrt{5}$.

Nous n'avons jusqu'ici aucune borne supérieure de la capacité. Pour en obtenir une, suivons encore les idées originales de Shannon. Nous avons d'abord besoin de la définition duale d'un ensemble indépendant. Nous rappelons qu'un sous-ensemble $C \subseteq V$ est une *clique* si deux sommets quelconques de C sont joints par une arête. Ainsi, les sommets forment des cliques triviales de taille 1, les arêtes les cliques de taille 2, les triangles les cliques de taille 3, et ainsi de suite. Soit $\mathcal{C}$ l'ensemble des cliques dans G. Considérons une distribution de probabilité arbitraire $\boldsymbol{x} = (x_v : v \in V)$ sur l'ensemble des sommets, c'est-à-dire que $x_v \geqslant 0$ et $\sum_{v \in V} x_v = 1$. Associons à chaque distribution $\boldsymbol{x}$ la « valeur maximale d'une clique » :

$$\lambda(\boldsymbol{x}) \;=\; \max_{C \in \mathcal{C}} \sum_{v \in C} x_v$$

et posons enfin :

$$\lambda(G) \;=\; \min_{\boldsymbol{x}} \lambda(\boldsymbol{x}) \;=\; \min_{\boldsymbol{x}} \max_{C \in \mathcal{C}} \sum_{v \in C} x_v$$

Pour être rigoureux, on devrait utiliser $\inf$ au lieu de $\min$, mais le minimum existe car λ est continue sur l'ensemble compact de toutes les distributions.

Considérons maintenant un ensemble indépendant $U \subseteq V$ de taille maximale $\alpha(G) = \alpha$ et définissons la distribution associée à U, $\boldsymbol{x}_U = (x_v : v \in V)$, en posant $x_v = \frac{1}{\alpha}$ si $v \in U$ et $x_v = 0$ sinon. Puisque toute clique contient au plus un sommet de U, $\lambda(\boldsymbol{x}_U) = \frac{1}{\alpha}$ donc, par définition de $\lambda(G)$:

$$\lambda(G) \leqslant \frac{1}{\alpha(G)} \qquad \text{ou} \qquad \alpha(G) \leqslant \lambda(G)^{-1}$$

Shannon a observé que $\lambda(G)^{-1}$ est en fait une borne supérieure de tous les $\sqrt[n]{\alpha(G^n)}$, donc aussi de $\Theta(G)$. Pour le prouver, il suffit de montrer qu'étant donnés des graphes G et H :

$$\lambda(G \times H) \;=\; \lambda(G)\lambda(H) \tag{1}$$

puis que cela implique $\lambda(G^n) = \lambda(G)^n$ et donc :

$$\begin{aligned} \alpha(G^n) &\leqslant \lambda(G^n)^{-1} = \lambda(G)^{-n} \\ \sqrt[n]{\alpha(G^n)} &\leqslant \lambda(G)^{-1} \end{aligned}$$

Pour prouver (1) on applique le théorème de dualité en programmation linéaire (voir [1]) et l'on obtient :

$$\lambda(G) \;=\; \min_{\boldsymbol{x}} \max_{C \in \mathcal{C}} \sum_{v \in C} x_v \;=\; \max_{\boldsymbol{y}} \min_{v \in V} \sum_{C \ni v} y_C \tag{2}$$

où le terme de droite parcourt toutes les distributions de probabilité $\boldsymbol{y} = (y_C : C \in \mathcal{C})$ sur $\mathcal{C}$.

Considérons $G \times H$. Soit $\boldsymbol{x}$ et $\boldsymbol{x}'$ des distributions qui atteignent les minima, $\lambda(\boldsymbol{x}) = \lambda(G)$, $\lambda(\boldsymbol{x}') = \lambda(H)$. Dans l'ensemble des sommets de

$G \times H$ attribuons la valeur $z_{(u,v)} = x_u x'_v$ au sommet (u, v). Puisque $\sum_{(u,v)} z_{(u,v)} = \sum_u x_u \sum_v x'_v = 1$, nous obtenons une distribution. Nous observons ensuite que les cliques maximales de $G \times H$ sont de la forme $C \times D = \{(u, v) : u \in C, v \in D\}$ où C et D sont des cliques de G et H respectivement. Par conséquent :

$$\begin{aligned} \lambda(G \times H) \leqslant \lambda(\boldsymbol{z}) &= \max_{C \times D} \sum_{(u,v) \in C \times D} z_{(u,v)} \\ &= \max_{C \times D} \sum_{u \in C} x_u \sum_{v \in D} x'_v = \lambda(G)\lambda(H) \end{aligned}$$

par définition de $\lambda(G \times H)$. De la même manière on montre l'inégalité inverse $\lambda(G \times H) \geqslant \lambda(G)\lambda(H)$ en utilisant l'expression duale de $\lambda(G)$ dans (2). En résumé :

$$\Theta(G) \leqslant \lambda(G)^{-1}$$

pour tout graphe G.

Appliquons ce que nous avons trouvé au 5-cycle et, plus généralement, au m-cycle C_m. En utilisant la distribution uniforme $(\frac{1}{m}, \ldots, \frac{1}{m})$ sur les sommets, nous obtenons $\lambda(C_m) \leqslant \frac{2}{m}$, puisque toute clique contient deux sommets au plus. De la même manière, en choisissant $\frac{1}{m}$ pour les arêtes et 0 pour les sommets, nous obtenons $\lambda(C_m) \geqslant \frac{2}{m}$ d'après l'expression duale de (2). On en conclut que $\lambda(C_m) = \frac{2}{m}$ et par conséquent :

$$\Theta(C_m) \leqslant \frac{m}{2}$$

pour tout m. Si m est pair, on a évidemment $\alpha(C_m) = \frac{m}{2}$ et donc aussi $\Theta(C_m) = \frac{m}{2}$. Si m est impair on a $\alpha(C_m) = \frac{m-1}{2}$. Si $m = 3$, C_3 est une clique et chaque produit C_3^n est aussi une clique, ce qui implique $\alpha(C_3) = \Theta(C_3) = 1$. Donc le premier cas intéressant est le 5-cycle, pour lequel nous savons jusqu'à présent que :

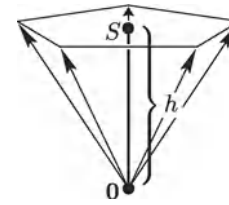
$$\sqrt{5} \leqslant \Theta(C_5) \leqslant \frac{5}{2}. \tag{3}$$

Avec des méthodes issues de la programmation linéaire (et d'autres idées) Shannon a été en mesure de calculer la capacité de nombreux graphes, en particulier de tous les graphes à cinq sommets ou moins, à la seule exception de C_5, pour lequel il n'a pas pu améliorer les inégalités (3). C'est à ce stade que les choses ont stagné pendant plus de vingt ans jusqu'à ce que László Lovász montre par un argument étonnamment simple qu'en fait, $\Theta(C_5) = \sqrt{5}$. Un problème combinatoire apparemment très difficile fut résolu par un argument élégant et inattendu.

La principale idée neuve de Lovász a consisté à représenter les sommets v du graphe par des vecteurs réels de longueur 1 de sorte que deux vecteurs quelconques associés à des sommets non adjacents dans G soient orthogonaux. Appelons *représentation orthonormale* de G un tel ensemble de vecteurs. Il est clair qu'une telle représentation existe toujours : il suffit

simplement de prendre les vecteurs unitaires $(1,0,\ldots,0)$, $(0,1,0,\ldots,0)$, $\ldots$, $(0,0,\ldots,1)$ en dimension $m = |V|$.

On peut obtenir une représentation orthonormale du graphe C_5 dans $\mathbb{R}^3$ en considérant un « parapluie » à cinq baleines $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_5$ de longueur unité. Ouvrons maintenant le parapluie (les extrémités restant à l'origine) jusqu'à ce que les angles entre des baleines alternées soient égaux à 90°.



Le parapluie de Lovász.

Lovász a alors montré que la hauteur h du parapluie, c'est à dire la distance entre $\mathbf{0}$ et S, conduit à la majoration :

$$\Theta(C_5) \leqslant \frac{1}{h^2} \tag{4}$$

Un calcul simple donne $h^2 = \frac{1}{\sqrt{5}}$ (voir encadré) On en déduit $\Theta(C_5) \leqslant \sqrt{5}$; par conséquent $\Theta(C_5) = \sqrt{5}$.

Voyons maintenant comment Lovász a procédé pour montrer l'inégalité (4) (ses résultats étaient en fait beaucoup plus généraux). Considérons le produit scalaire habituel :

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = x_1 y_1 + \ldots + x_s y_s$$

de deux vecteurs $\boldsymbol{x} = (x_1, \ldots, x_s)$ et $\boldsymbol{y} = (y_1, \ldots, y_s)$ dans $\mathbb{R}^s$. $|\boldsymbol{x}|^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle = x_1^2 + \ldots + x_s^2$ est alors le carré de la longueur $|\boldsymbol{x}|$ de $\boldsymbol{x}$. L'angle γ entre $\boldsymbol{x}$ et $\boldsymbol{y}$ est défini par :

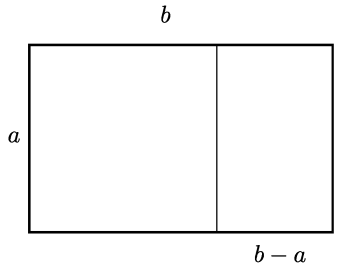
$$\cos\gamma = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{|\boldsymbol{x}||\boldsymbol{y}|}$$

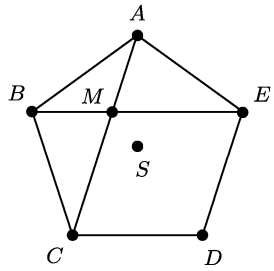
Ainsi, $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$ si et seulement si $\boldsymbol{x}$ et $\boldsymbol{y}$ sont orthogonaux.

Recherchons maintenant une borne supérieure pour la capacité de Shannon de tout graphe G possédant une représentation orthonormale particulièrement « belle ». Pour cela, soit $T = \{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)}\}$ une représentation orthonormale de G dans $\mathbb{R}^s$, où $\boldsymbol{v}^{(i)}$ correspond au sommet v_i. Supposons de plus que tous les vecteurs $\boldsymbol{v}^{(i)}$ font le *même* angle ($\neq 90°$) avec le vecteur $\boldsymbol{u} := \frac{1}{m}(\boldsymbol{v}^{(1)} + \ldots + \boldsymbol{v}^{(m)})$ ou, ce qui est équivalent, que le produit scalaire :

$$\langle \boldsymbol{v}^{(i)}, \boldsymbol{u} \rangle = \sigma_T$$

ait la même valeur $\sigma_T \neq 0$ pour tout i. Appelons cette valeur σ_T la *constante* de la représentation T. En ce qui concerne le parapluie de Lovász qui représente C_5, la condition $\langle \boldsymbol{v}^{(i)}, \boldsymbol{u} \rangle = \sigma_T$ est certainement réalisée si $\boldsymbol{u} := \vec{OS}$.





Les pentagones et le nombre d'or

Traditionnellement, on considère qu'un rectangle est esthétiquement agréable si, après en avoir découpé un carré de longueur a, le rectangle restant a la même forme que l'original. Les longueurs a, b des côtés d'un tel rectangle doivent vérifier $\frac{b}{a} = \frac{a}{b-a}$. En posant $\tau := \frac{b}{a}$, on obtient : $\tau = \frac{1}{\tau - 1}$ c'est-à-dire $\tau^2 - \tau - 1 = 0$. La résolution de cette équation du second degré conduit au *nombre d'or* $\tau = \frac{1+\sqrt{5}}{2} \approx 1.6180$.

Considérons maintenant un pentagone régulier de côté a. Soit d la longueur de ses diagonales. Euclide savait déjà (*Livre XIII*, 8) que $\frac{d}{a} = \tau$ et que le point d'intersection de deux diagonales divise les diagonales dans une proportion égale au nombre d'or.
Voici la Preuve du Grand Livre d'Euclide : puisque la somme totale des angles du pentagone est 3π, l'angle en tout sommet est égal à $\frac{3\pi}{5}$. Cela implique $\sphericalangle ABE = \frac{\pi}{5}$, puisque ABE est un triangle isocèle. Cela, à son tour, implique $\sphericalangle AMB = \frac{3\pi}{5}$; on en conclut que les triangles ABC et AMB sont semblables. Le quadrilatère $CMED$ est un losange puisque ses côtés opposés sont parallèles (examiner les angles), donc $|MC| = a$ et par suite $|AM| = d - a$. Comme ABC et AMB sont semblables,

$$\frac{d}{a} = \frac{|AC|}{|AB|} = \frac{|AB|}{|AM|} = \frac{a}{d-a} = \frac{|MC|}{|MA|} = \tau$$

En outre, la distance s d'un sommet au centre de gravité S vérifie la relation suivante, que le lecteur est invité à démontrer : $s^2 = \frac{d^2}{\tau+2}$ (noter que BS coupe la diagonale AC à angle droit et la divise en deux segments égaux). Pour terminer cette excursion en géométrie, considérons maintenant un parapluie formant un pentagone régulier. Puisque des baleines alternées (de longueur 1) forment un angle droit, le théorème de Pythagore implique que $d = \sqrt{2}$; par conséquent $s^2 = \frac{2}{\tau+2} = \frac{4}{\sqrt{5}+5}$. Donc, en utilisant à nouveau la relation de Pythagore, la hauteur $h = |OS|$ vérifie le résultat espéré

$$h^2 = 1 - s^2 = \frac{1+\sqrt{5}}{\sqrt{5}+5} = \frac{1}{\sqrt{5}}$$

Nous procédons maintenant en trois étapes.

(A) Considérons une distribution de probabilité $\boldsymbol{x} = (x_1, \ldots, x_m)$ sur V, posons :

$$\mu(\boldsymbol{x}) \;:=\; |x_1 \boldsymbol{v}^{(1)} + \ldots + x_m \boldsymbol{v}^{(m)}|^2$$

et :

$$\mu_T(G) \;:=\; \inf_{\boldsymbol{x}} \mu(\boldsymbol{x})$$

Soit U un ensemble indépendant maximal dans G tel que $|U| = \alpha$. Définissons $\boldsymbol{x}_U = (x_1, \ldots, x_m)$ où $x_i = \frac{1}{\alpha}$ si $v_i \in U$ et $x_i = 0$ sinon. Puisque tous les vecteurs $\boldsymbol{v}^{(i)}$ sont unitaires et que $\langle \boldsymbol{v}^{(i)}, \boldsymbol{v}^{(j)} \rangle = 0$ pour tout couple de sommets non-adjacents :

$$\mu_{_T}(G) \;\leqslant\; \mu(\boldsymbol{x}_U) \;=\; \Big| \sum_{i=1}^{m} x_i \boldsymbol{v}^{(i)} \Big|^2 \;=\; \sum_{i=1}^{m} x_i^2 = \alpha \frac{1}{\alpha^2} = \frac{1}{\alpha}$$

Ainsi, nous avons $\mu_{_T}(G) \leqslant \alpha^{-1}$; par conséquent :

$$\alpha(G) \;\leqslant\; \frac{1}{\mu_{_T}(G)}$$

(B) Calculons ensuite $\mu_{_T}(G)$. Rappelons l'inégalité de Cauchy-Schwarz :

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle^2 \;\leqslant\; |\boldsymbol{a}|^2 \, |\boldsymbol{b}|^2$$

pour des vecteurs $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^s$. Appliquée à $\boldsymbol{a} = x_1 \boldsymbol{v}^{(1)} + \ldots + x_m \boldsymbol{v}^{(m)}$ et $\boldsymbol{b} = \boldsymbol{u}$, l'inégalité implique :

$$\langle x_1 \boldsymbol{v}^{(1)} + \ldots + x_m \boldsymbol{v}^{(m)}, \boldsymbol{u} \rangle^2 \;\leqslant\; \mu(\boldsymbol{x}) \, |\boldsymbol{u}|^2 \tag{5}$$

Par hypothèse nous avons $\langle \boldsymbol{v}^{(i)}, \boldsymbol{u} \rangle = \sigma_{_T}$ pour tout i donc :

$$\langle x_1 \boldsymbol{v}^{(1)} + \ldots + x_m \boldsymbol{v}^{(m)}, \boldsymbol{u} \rangle \;=\; (x_1 + \ldots + x_m)\, \sigma_{_T} \;=\; \sigma_{_T}$$

pour *toute* distribution $\boldsymbol{x}$. En particulier, cela se produit pour la distribution uniforme $(\frac{1}{m}, \ldots, \frac{1}{m})$, ce qui implique $|\boldsymbol{u}|^2 = \sigma_{_T}$. Par conséquent (5) s'écrit :

$$\sigma_{_T}^2 \;\leqslant\; \mu(\boldsymbol{x})\, \sigma_{_T} \qquad \text{où} \qquad \mu_{_T}(G) \;\geqslant\; \sigma_{_T}$$

D'autre part, si $\boldsymbol{x} = (\frac{1}{m}, \ldots, \frac{1}{m})$, on obtient :

$$\mu_{_T}(G) \;\leqslant\; \mu(\boldsymbol{x}) \;=\; |\tfrac{1}{m}(\boldsymbol{v}^{(1)} + \ldots + \boldsymbol{v}^{(m)})|^2 \;=\; |\boldsymbol{u}|^2 \;=\; \sigma_{_T}$$

et nous avons donc prouvé que :

$$\mu_{_T}(G) \;=\; \sigma_{_T} \tag{6}$$

En résumé, nous avons établi l'inégalité :

$$\alpha(G) \;\leqslant\; \frac{1}{\sigma_{_T}} \tag{7}$$

pour *toute* représentation orthonormale T de constante $\sigma_{_T}$.

(C) Pour étendre cette inégalité à $\Theta(G)$, procédons comme précédemment. Considérons encore le produit $G \times H$ de deux graphes. Attribuons à G et H des représentations orthonormales R et S respectivement dans $\mathbb{R}^r$ et $\mathbb{R}^s$, de constantes $\sigma_{_R}$ et $\sigma_{_S}$. Soit $\boldsymbol{v} = (v_1, \ldots, v_r)$ un vecteur de R et

$\boldsymbol{w} = (w_1, \ldots, w_s)$ un vecteur de S. Au sommet de $G \times H$ correspondant à la paire $(\boldsymbol{v}, \boldsymbol{w})$ nous associons le vecteur :

$$\boldsymbol{v}\boldsymbol{w}^T := (v_1 w_1, \ldots, v_1 w_s, v_2 w_1, \ldots, v_2 w_s, \ldots, v_r w_1, \ldots, v_r w_s) \in \mathbb{R}^{rs}$$

Il est immédiat de vérifier que $R \times S := \{\boldsymbol{v}\boldsymbol{w}^T : \boldsymbol{v} \in R, \boldsymbol{w} \in S\}$ est une représentation orthonormale de $G \times H$ de constante $\sigma_R \sigma_S$. Par conséquent, nous obtenons grâce à (6) :

$$\mu_{R\times S}(G \times H) \;=\; \mu_R(G)\mu_S(H)$$

Pour $G^n = G \times \ldots \times G$ et la représentation T de constante σ_T, cela signifie que :

$$\mu_{T^n}(G^n) \;=\; \mu_T(G)^n \;=\; \sigma_T^n$$

et nous obtenons d'après (7) :

$$\alpha(G^n) \leqslant \sigma_T^{-n} \qquad \sqrt[n]{\alpha(G^n)} \leqslant \sigma_T^{-1}$$

En réunissant ces éléments, on obtient le résultat de Lovász :

Théorème. *Pour toute représentation orthonormale* $T = \{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)}\}$ *de* G *de constante* σ_T*, on a :*

$$\Theta(G) \leqslant \frac{1}{\sigma_T}. \tag{8}$$

« Parapluies à cinq baleines ».

En examinant le parapluie de Lovász, nous constatons que $\boldsymbol{u} = (0, 0, h{=}\frac{1}{\sqrt[4]{5}})$ et par conséquent : $\sigma = \langle \boldsymbol{v}^{(i)}, \boldsymbol{u} \rangle = h^2 = \frac{1}{\sqrt{5}}$, ce qui implique $\Theta(C_5) \leqslant \sqrt{5}$. Le problème de Shannon est donc résolu.

Poursuivons encore un peu notre discussion. On déduit de (8) que la majoration obtenue pour $\Theta(G)$ est d'autant meilleure que σ_T est grand pour une représentation de G. Voici une méthode qui nous donne une représentation orthonormale pour *n'importe quel* graphe G. Associons à $G = (V, E)$, avec $V = \{v_1, \ldots, v_m\}$, la *matrice d'adjacence* $A = (a_{ij})$, définie par :

$$a_{ij} \;:=\; \begin{cases} 1 & \text{si } v_i v_j \in E \\ 0 & \text{sinon} \end{cases}$$

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

La matrice d'adjacence du 5-cycle C_5.

A est une matrice symétrique réelle avec des 0 sur la diagonale principale.

Nous avons maintenant besoin de deux résultats d'algèbre linéaire. D'abord, en tant que matrice symétrique, A possède m valeurs propres réelles $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_m$ (certaines d'entre elles pouvant être égales), et la somme des valeurs propres est égale à la somme des éléments de la diagonale de A, c'est-à-dire 0. Par conséquent, la plus petite valeur propre doit être négative (sauf dans le cas trivial où G n'a pas d'arêtes). Soit $p = |\lambda_m| = -\lambda_m$ la

valeur absolue de la plus petite valeur propre. Considérons la matrice :

$$M \;:=\; I + \frac{1}{p}A,$$

où I désigne la $(m \times m)$-matrice identité. Cette matrice M a les valeurs propres suivantes : $1 + \frac{\lambda_1}{p} \geqslant 1 + \frac{\lambda_2}{p} \geqslant \ldots \geqslant 1 + \frac{\lambda_m}{p} = 0$. Énonçons maintenant le second résultat nécessaire pour la suite : si $M = (m_{ij})$ est une matrice symétrique réelle dont toutes les valeurs propres sont $\geqslant 0$, alors il existe des vecteurs $\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)} \in \mathbb{R}^s$ où $s = \mathrm{rg}(M)$, tels que :

$$m_{ij} \;=\; \langle \boldsymbol{v}^{(i)}, \boldsymbol{v}^{(j)} \rangle \qquad (1 \leqslant i, j \leqslant m)$$

En particulier, si $M = I + \frac{1}{p}A$ on obtient :

$$\langle \boldsymbol{v}^{(i)}, \boldsymbol{v}^{(i)} \rangle \;=\; m_{ii} = 1 \qquad \text{pour tout } i$$

Les valeurs propres de C_m

Examinons la matrice d'adjacence A du cycle C_m. Pour trouver ses valeurs propres (et ses vecteurs propres) on utilise les racines m-ièmes de l'unité. Ce sont $1, \zeta, \zeta^2, \ldots, \zeta^{m-1}$ avec $\zeta = e^{\frac{2\pi i}{m}}$ (voir encadré en page 37).

Considérons $\lambda = \zeta^k$ l'une quelconque de ces racines. Nous affirmons que $(1, \lambda, \lambda^2, \ldots, \lambda^{m-1})$ est un vecteur propre de A relatif à la valeur propre $\lambda + \lambda^{-1}$. En effet, la forme de A implique que :

$$A \begin{pmatrix} 1 \\ \lambda \\ \lambda^2 \\ \vdots \\ \lambda^{m-1} \end{pmatrix} = \begin{pmatrix} \lambda + \lambda^{m-1} \\ \lambda^2 + 1 \\ \lambda^3 + \lambda \\ \vdots \\ 1 + \lambda^{m-2} \end{pmatrix} = (\lambda + \lambda^{-1}) \begin{pmatrix} 1 \\ \lambda \\ \lambda^2 \\ \vdots \\ \lambda^{m-1} \end{pmatrix}$$

Comme les vecteurs $(1, \lambda, \ldots, \lambda^{m-1})$ sont indépendants (ils forment une matrice de Vandermonde), si m est impair :

$$\begin{aligned} \zeta^k + \zeta^{-k} \;&=\; [(\cos(2k\pi/m) + i\sin(2k\pi/m)] \\ &\quad + [\cos(2k\pi/m) - i\sin(2k\pi/m)] \\ &=\; 2\cos(2k\pi/m) \qquad (0 \leqslant k \leqslant \tfrac{m-1}{2}) \end{aligned}$$

sont toutes les valeurs propres de A. La fonction cosinus étant décroissante, la plus petite valeur propre de A est :

$$2\cos\Big(\frac{(m-1)\pi}{m}\Big) \;=\; -2\cos\frac{\pi}{m}$$

et :

$$\langle \boldsymbol{v}^{(i)}, \boldsymbol{v}^{(j)} \rangle \;=\; \frac{1}{p} a_{ij} \qquad \text{pour } i \neq j$$

Comme $a_{ij} = 0$ dès que $v_i v_j \notin E$, on voit facilement que les vecteurs $\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)}$ forment une représentation orthonormale de G.

Appliquons enfin cette construction aux m-cycles C_m lorsque $m \geqslant 5$ est impair. On trouve facilement dans ce cas que $p = |\lambda_{\min}| = 2\cos\frac{\pi}{m}$ (voir encadré).

Chaque ligne de la matrice d'adjacence contient deux 1, ce qui implique que chaque ligne de la matrice M a pour somme $1 + \frac{2}{p}$. En ce qui concerne la représentation $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)}\}$ cela signifie que :

$$\langle \boldsymbol{v}^{(i)}, \boldsymbol{v}^{(1)} + \ldots + \boldsymbol{v}^{(m)} \rangle \;=\; 1 + \frac{2}{p} \;=\; 1 + \frac{1}{\cos\frac{\pi}{m}}$$

et par conséquent :

$$\langle \boldsymbol{v}^{(i)}, \boldsymbol{u} \rangle \;=\; \frac{1}{m}\bigl(1 + (\cos\tfrac{\pi}{m})^{-1}\bigr) \;=\; \sigma$$

pour tout i. Nous pouvons ainsi appliquer le résultat principal (8) et conclure que :

$$\Theta(C_m) \;\leqslant\; \frac{m}{1 + (\cos\frac{\pi}{m})^{-1}} \qquad (\text{pour } m \geqslant 5 \text{ impair}) \qquad (9)$$

Comme $\cos\frac{\pi}{m} < 1$, la majoration (9) est meilleure que la majoration $\Theta(C_m) \leqslant \frac{m}{2}$ trouvée auparavant. Notons de plus que $\cos\frac{\pi}{5} = \frac{\tau}{2}$, où $\tau = \frac{\sqrt{5}+1}{2}$ est le nombre d'or. Par conséquent pour $m = 5$ nous obtenons :

$$\Theta(C_5) \leqslant \frac{5}{1 + \frac{4}{\sqrt{5}+1}} = \frac{5(\sqrt{5}+1)}{5+\sqrt{5}} = \sqrt{5}$$

La représentation orthonormale donnée par cette construction est précisément le « parapluie de Lovász ».

Qu'en est-il de C_7, C_9 et des autres cycles impairs ? En considérant $\alpha(C_m^2)$, $\alpha(C_m^3)$ et d'autres petites puissances, on peut certainement améliorer la minoration $\frac{m-1}{2} \leqslant \Theta(C_m)$ mais pour aucun nombre impair $m \geqslant 7$, les meilleures minorations connues ne correspondent à la borne supérieure donnée dans (8). Ainsi, vingt ans après que Lovász eut merveilleusement démontré que $\Theta(C_5) = \sqrt{5}$, ces problèmes restent ouverts et sont considérés comme très difficiles. Mais après tout, c'était déjà le cas avant l'intervention de Lovász...

Par exemple, pour $m = 7$ nous savons seulement que :

$$\sqrt[4]{108} \leqslant \Theta(C_7) \leqslant \frac{7}{1 + (\cos\frac{\pi}{7})^{-1}}$$

c'est-à-dire que :
$3.2237 \leqslant \Theta(C_7) \leqslant 3.3177$.

Bibliographie

[1] V. CHVÁTAL : *Linear Programming,* Freeman, New York 1983.

[2] W. HAEMERS : *Eigenvalue methods,* in : "Packing and Covering in Combinatorics" (A. Schrijver, ed.), Math. Centre Tracts **106** (1979), 15-38.

[3] L. LOVÁSZ : *On the Shannon capacity of a graph,* IEEE Trans. Information Theory **25** (1979), 1-7.

[4] C. E. SHANNON : *The zero-error capacity of a noisy channel,* IRE Trans. Information Theory **3** (1956), 3-15.

Le nombre chromatique des graphes de Kneser

Chapitre 38

En 1955, Martin Kneser, spécialiste de théorie des nombres, formula une conjecture d'apparence anodine. Elle constitua l'un des plus grands défis de la théorie des graphes jusqu'à ce que László Lovász lui trouve, vingt-trois ans plus tard, une solution brillante et complètement inattendue, utilisant la topologie et le théorème de Borsuk-Ulam.

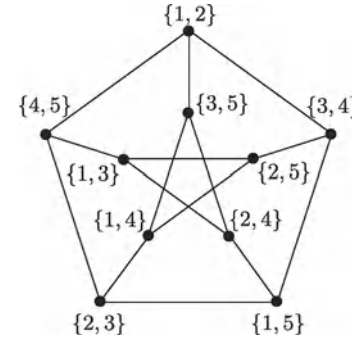
En mathématiques, il arrive souvent, lorsqu'on trouve la démonstration d'un problème ouvert depuis longtemps, que cette démonstration soit rapidement améliorée. Ce fut le cas pour le problème posé par Kneser. Quelques semaines après la démonstration proposée par Lovász, Imre Bárány montra comment combiner le théorème de Borsuk-Ulam avec un autre résultat connu afin d'établir une démonstration élégante de la conjecture de Kneser. Enfin, en 2002, Joshua Greene, alors étudiant de premier cycle, simplifia la démonstration de Bárány et c'est sa version que nous présentons ici.

Commençons par le début. Considérons le *graphe de Kneser* $K(n,k)$ défini comme suit pour les entiers $n \geqslant k \geqslant 1$. L'ensemble $V(n,k)$ de ses sommets est la famille des sous-ensembles à k éléments (appelés k-parties dans la suite) de $\{1,\dots,n\}$. Le nombre de sommets est donc $|V(n,k)| = \binom{n}{k}$. Les sommets associés à deux telles k-parties A et B sont adjacents dans le graphe si A et B sont disjoints, c'est-à-dire si $A \cap B = \varnothing$.

Si $n < 2k$, il n'est pas possible de trouver deux k-parties disjointes de $\{1,\dots,n\}$, si bien que le graphe $K(n,k)$ correspondant n'a aucune arête. Ainsi, on suppose désormais que $n \geqslant 2k$.

Les graphes de Kneser fournissent un lien intéressant entre la théorie des graphes et les ensembles finis. Par exemple, on peut déterminer le *nombre de stabilité*[1] $\alpha(K(n,k))$ de $K(n,k)$, c'est-à-dire que l'on peut se demander quelle est la taille maximale d'une famille intersectante de k-parties (c'est-à-dire une famille de k-parties de $\{1,\dots,n\}$ telles que deux d'entre elles ont toujours au moins un élément en commun). La réponse est donnée par le théorème d'Erdős-Ko-Rado, qui est présenté en détail au chapitre 27 : $\alpha(K(n,k)) = \binom{n-1}{k-1}$.

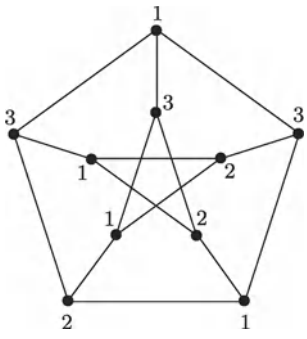
On peut de la sorte s'intéresser au calcul de divers paramètres pour cette famille de graphes. Kneser choisit le plus intéressant : le *nombre chromatique* $\chi(K(n,k))$. On rappelle (se reporter aux chapitres précédents) qu'une coloration (des sommets) d'un graphe G consiste en une application $c : V \to \{1,\dots,m\}$ telle que deux sommets adjacents soient affectés de couleurs distinctes. Le nombre chromatique $\chi(G)$ du graphe G est le nombre minimum de couleurs suffisant pour établir une coloration de V.



Le graphe de Kneser $K(5,2)$ n'est autre que le célèbre graphe de Petersen.

Cela implique en particulier :
$\chi(K(n,k)) \geqslant \frac{|V(n,k)|}{\alpha(K(n,k))} = \frac{\binom{n}{k}}{\binom{n-1}{k-1}} = \frac{n}{k}$.

1. N.d.T. : la terminologie anglo-saxonne préfère le terme de *nombre d'indépendance*.



Une 3-coloration du graphe de Petersen.

En d'autres termes, on veut que l'ensemble V des sommets soit partitionné en une union disjointe de *classes de couleurs* aussi peu nombreuses que possible, $V = V_1 \dot\cup \cdots \dot\cup V_{\chi(G)}$, où chaque classe V_i est dépourvue d'arête.

Pour les graphes $K(n,k)$, cela consiste à trouver une partition $V(n,k)$ telle que $V(n,k) = V_1 \dot\cup \cdots \dot\cup V_\chi$ où chaque V_i est une *famille intersectante* de k-parties. Comme on suppose que $n \geqslant 2k$, on écrit désormais n sous la forme $n = 2k + d$ avec $k \geqslant 1$ et $d \geqslant 0$.

Voici comment on peut trouver une coloration de $K(n,k)$ qui n'utilise que $d+2$ couleurs : pour $i \in \{1, 2, \ldots, d+1\}$, définissons V_i comme la famille réunissant les k-parties possédant i comme plus petit élément. Les k-parties restantes sont incluses dans l'ensemble $\{d+2, d+3, \ldots, 2k+d\}$ qui n'a que $2k-1$ éléments. Ainsi, ces parties sont intersectantes et l'on peut attribuer la couleur $d+2$ à chacune d'elle.

Il en résulte que $\chi(K(2k+d,k)) \leqslant d+2$. Le défi lancé par Kneser consiste à montrer qu'il y a égalité.

La conjecture de Kneser.

$$\chi(K(2k+d,k)) \;=\; d+2.$$

Pour $d = 0$, $K(2k,k)$ est constitué d'arêtes disjointes, une pour chaque couple de k-parties complémentaires. Ainsi $\chi(K(2k,k)) = 2$, en accord avec la conjecture.

La première idée qui vient à l'esprit pour démontrer ce résultat est de procéder par récurrence sur k et d. Les cas initiaux $k = 1$ et $d = 0$ ou $d = 1$ sont faciles à établir mais le passage de k à $k+1$ (ou de d à $d+1$) résiste. C'est pourquoi on en vient plutôt à reformuler la conjecture sous forme d'un problème d'existence :

Si la famille de k-parties de $\{1, 2, \ldots, 2k+d\}$ est partitionnée en $d+1$ classes, $V(n,k) = V_1 \dot\cup \cdots \dot\cup V_{d+1}$, alors, pour un certain i, V_i contient un couple (A, B) de k-parties disjointes.

L'intuition brillante de Lovász a été de comprendre qu'au cœur (topologique) du problème se trouve un théorème célèbre qui concerne la sphère unité d-dimensionnelle S^d de $\mathbb{R}^{d+1}$, $S^d = \{x \in \mathbb{R}^{d+1} : |x| = 1\}$.

Le théorème de Borsuk-Ulam.
Toute application $f : S^d \to \mathbb{R}^d$ continue de la sphère S^d dans l'espace euclidien de dimension d est telle qu'il existe deux points antipodaux x^* et $-x^*$ ayant même image par f : $f(x^*) = f(-x^*)$.

Ce résultat est l'une des pierres angulaires de la topologie. Il apparaît pour la première fois dans le célèbre article de Borsuk daté de 1933. Nous donnons les grandes lignes de sa démonstration en appendice ; on en trouve la preuve détaillée dans la partie 2.2 du merveilleux ouvrage de Matoušek intitulé *Using the Borsuk-Ulam theorem* [*Utilisation du théorème de Borsuk-Ulam*] dont le titre montre à lui seul la puissance et l'étendue du résultat. Ce

résultat admet en effet de nombreux énoncés équivalents, ce qui souligne la position centrale de ce théorème en topologie.

Nous allons nous appuyer sur une version que l'on peut déjà trouver dans un ouvrage de Lyusternik et Shnirel'man datant de 1930, qui précède donc l'énoncé de Borsuk.

Théorème. *Si la d-sphère S^d est recouverte par $d+1$ ensembles*

$$S^d \;=\; U_1 \cup \cdots \cup U_d \cup U_{d+1}$$

tels que chacun des d premiers ensembles $U_1, \ldots, U_d$ est soit ouvert soit fermé, alors l'un des $d+1$ ensembles contient un couple de points antipodaux $(x^*, -x^*)$.

Le cas où les $d+1$ ensembles sont fermés est dû à Lyusternik et Shnirel'man. Le cas où les $d+1$ ensembles sont ouverts est aussi répandu et est connu sous le nom de théorème de Lyusternik-Shnirel'man. L'apport de Greene consiste à remarquer que le théorème est encore vrai si les $d+1$ ensembles sont *soit ouverts, soit fermés*. Comme on va le voir, on n'a même pas besoin d'hypothèse sur U_{d+1}. Pour démontrer la conjecture de Kneser, il suffit que $U_1, \ldots, U_d$ soient ouverts.

■ **Démonstration du théorème de Lyusternik-Shnirel'man à l'aide du théorème de Borsuk-Ulam.** Soit $U_1 \cup \cdots \cup U_d \cup U_{d+1}$ un recouvrement de S^d conforme aux hypothèses évoquées, c'est-à-dire tel que les ensembles $U_1, \ldots, U_d$ sont soit ouverts soit fermés. Raisonnons par l'absurde : supposons qu'aucun des ensembles U_i ne comporte un couple de points antipodaux. On définit l'application $f : S^d \to \mathbb{R}^d$ par :

$$f(x) \;:=\; \big(\delta(x, U_1), \delta(x, U_2), \ldots, \delta(x, U_d)\big)$$

où $\delta(x, U_i)$ désigne la distance de x à U_i. Les applications $x \mapsto \delta(x, U_i)$ étant continues, il en résulte que f est continue. Le théorème de Borsuk-Ulam permet alors d'affirmer qu'il existe un couple de points antipodaux $(x^*, -x^*)$ éléments de S^d tels que $f(x^*) = f(-x^*)$. Comme U_{d+1} ne contient pas de couple de points antipodaux, au moins l'un des points x^* ou $-x^*$ doit se trouver dans l'un des ensembles U_i ($i \leqslant d$). Soit U_k ($k \leqslant d$) cet ensemble. À échange près de x^* avec $-x^*$, on peut supposer $x^* \in U_k$. Cela conduit en particulier à $\delta(x^*, U_k) = 0$ et, comme $f(x^*) = f(-x^*)$, on en déduit aussitôt $\delta(-x^*, U_k) = 0$.

Si U_k est fermé, alors $\delta(-x^*, U_k) = 0$ implique que $-x^* \in U_k$, ce qui signifie que U_k contient un couple de points antipodaux. Contradiction.

Si U_k est ouvert, alors $\delta(-x^*, U_k) = 0$ implique que $-x^*$ se trouve dans $\overline{U_k}$, l'adhérence de U_k. L'ensemble $\overline{U_k}$ est inclus dans $S^d \backslash (-U_k)$ car ce dernier est un sous-ensemble fermé de S^d contenant U_k. Cela signifie alors que $-x^*$ est élément de $S^d \backslash (-U_k)$, donc il n'appartient pas à $-U_k$ et donc x^* n'appartient pas à U_k. Contradiction. □

L'*adhérence* (ou *fermeture*) de U_k est le plus petit ensemble fermé contenant U_k (c'est-à-dire l'intersection de tous les ensembles fermés contenant U_k).

Un autre résultat d'existence concernant la sphère unité S^d est utilisé par Imre Báráni comme deuxième argument de sa démonstration.

Théorème de Gale. *Il existe sur S^d un ensemble de $2k+d$ points tel que toute demi-sphère ouverte (de S^d) contienne au moins k de ces points.*

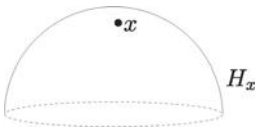
David Gale établit ce résultat en 1956 dans le contexte des polytopes présentant de nombreuses faces. Il en proposa une démonstration par récurrence assez compliquée mais aujourd'hui, avec le recul, on peut très facilement exhiber un tel ensemble de points et en vérifier les propriétés.

À l'aide de ces résultats, il n'y a plus qu'une petite étape à franchir pour établir la conjecture de Kneser. Toutefois, comme l'a montré Greene, on ne va même pas avoir besoin du résultat de Gale. Il suffit de prendre n'importe quel ensemble de $2k+d$ points sur S^{d+1} *disposés de manière générale* (c'est-à-dire que l'on ne peut pas trouver $d+2$ points situés sur un hyperplan passant par le centre de la sphère).

■ **Démonstration de la conjecture de Kneser.** Pour l'ensemble de départ, on prend $2k+d$ points en position générale sur la sphère S^{d+1}. Supposons l'ensemble $V(n,k)$ de toutes les k-parties de cet ensemble partitionné en $d+1$ classes, $V(n,k) = V_1 \dot\cup \cdots \dot\cup V_{d+1}$. Il nous faut trouver un couple de k-parties disjointes A et B appartenant à la même classe V_i.

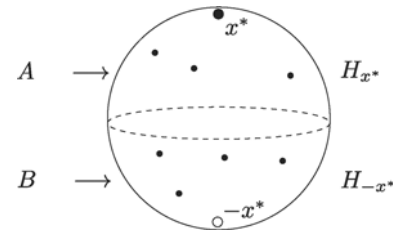
Pour $i \in \{1,\dots,d+1\}$ on pose :

$$O_i = \{x \in S^{d+1} / \text{ la demi-sphère ouverte } H_x \text{ de pôle } x \text{ contient une } k\text{-partie de } V_i\}.$$



Une demi-sphère ouverte incluse dans S^2.

Les O_i sont manifestement des ouverts. La réunion des ouverts O_i et du fermé $C = S^{d+1} \setminus (O_1 \cup \cdots \cup O_{d+1})$ recouvrent S^{d+1}. Selon le théorème de Lyusternik-Shnirel'man, nous savons que l'un de ces ensembles contient des points antipodaux x^* et $-x^*$. Cet ensemble ne peut être C. En effet, si x^* et $-x^*$ étaient éléments de C, alors par définition des O_i, les demi-sphères H_{x^*} et H_{-x^*} contiendraient moins de k points de la répartition initiale. Cela signifierait alors qu'au moins $d+2$ des points se trouveraient situés sur l'équateur $\overline{H}_{x^*} \cap \overline{H}_{-x^*}$ (équateur vis-à-vis du pôle x^*), c'est-à-dire qu'ils seraient situés sur un hyperplan passant par l'origine. Cela ne peut se produire puisque les points sont en position générale. Ainsi, l'un des O_i contient un couple $(x^*,-x^*)$; il existe donc des k-parties A et B appartenant toutes les deux à la classe V_i, avec $A \subseteq H_{x^*}$ et $B \subseteq H_{-x^*}$.



Comme les demi-sphères considérées sont *ouvertes*, H_{x^*} et H_{-x^*} sont disjointes, donc A et B sont disjointes aussi, ce qui termine la démonstration. □

Le lecteur peut se demander si des résultats aussi sophistiqués que le théorème de Borsuk-Ulam sont véritablement nécessaires pour établir un résultat sur des ensembles finis. Une belle démonstration combinatoire a certes été trouvée récemment par Jiří Matoušek ; toutefois, après lecture attentive, il apparaît que cette démonstration fait, elle aussi, appel à des arguments d'ordre topologique, topologie discrète en l'occurrence.

Appendice - Esquisse de la démonstration du théorème de Borsuk-Ulam

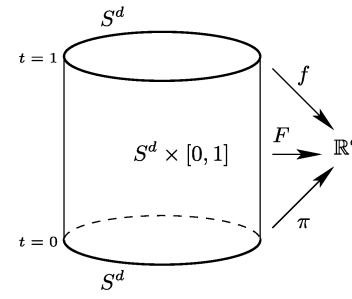
Pour toute application *générique* (encore dite application en *position générale*) d'un espace compact de dimension d dans un espace de dimension d, tout élément de l'image admet seulement un nombre fini d'antécédents. Pour une application générique d'un espace de dimension $(d+1)$ dans un espace de dimension d, on s'attend à ce que la préimage d'un élément de l'image soit un ensemble de dimension 1, c'est-à-dire que ce soit un ensemble de courbes. Que l'on soit dans le cas d'une application linéaire par morceaux ou dans le cas d'une application suffisamment régulière, on peut prouver assez facilement que l'on peut déformer l'application en une application presque générique.

Dans le cas du théorème de Borsuk-Ulam, l'idée consiste à montrer que toute fonction (générique) $f : S^d \to \mathbb{R}^d$ est associée à un nombre impair (donc fini et non nul) de couples antipodaux. Si f n'était pas associée à un couple antipodal, alors f serait arbitrairement proche d'une fonction générique $\widetilde{f}$ non associée à un couple antipodal.

On considère alors la projection $\pi : S^d \to \mathbb{R}^d$ qui a juste pour effet d'effacer la dernière coordonnée. Cette fonction confond le « pôle nord » e_{d+1} de la d-sphère avec le « pôle sud » $-e_{d+1}$. Pour toute fonction $f : S^d \to \mathbb{R}^d$ on construit une déformation continue de π vers f, c'est-à-dire que l'on interpole (par exemple linéairement) afin d'obtenir une application continue :

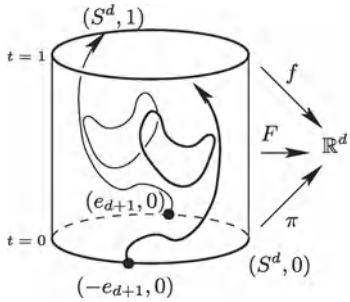
$$F : S^d \times [0,1] \longrightarrow \mathbb{R}^d,$$

avec $F(x,0) = \pi(x)$ et $F(x,1) = f(x)$ pour tout $x \in S^d$ (une telle application s'appelle une *homotopie*).



À présent, on perturbe F avec précaution pour en faire une application (générique) $\widetilde{F} : S^d \times [0,1] \to \mathbb{R}^d$, que l'on peut supposer suffisamment régulière ou linéaire par morceaux sur une triangulation bien choisie de $S^d \times [0,1]$. Si cette perturbation est « suffisamment petite » et qu'on la réalise avec précaution, alors la version perturbée de la projection $\widetilde{\pi}(x) := \widetilde{F}(x,0)$ devrait encore reconnaître les points antipodaux $\pm e_{d+1}$ à l'exclusion de tout autre couple de points antipodaux. Si $\widetilde{F}$ est suffisamment générique, alors les points de $S^d \times [0,1]$ définis par

$$M := \left\{(x,t) \in S^d \times [0,1] : \widetilde{F}(-x,t) = \widetilde{F}(x,t)\right\}$$



selon le théorème des fonctions implicites (pour le cas des fonctions régulières ou pour le cas des fonctions linéaires par morceaux) constituent un ensemble de chemins et de courbes fermés. Manifestement, cet ensemble est *symétrique*, c'est-à-dire vérifie $(-x,t) \in M$ si et seulement si $(x,t) \in M$.

Les extrémités des chemins situés dans M ne peuvent se trouver qu'au bord de $S^d \times [0,1]$, soit donc en $t = 0$ ou en $t = 1$. Toutefois, les seules extrémités pour $t = 0$ sont situées en $(\pm e_{d+1}, 0)$ et les deux chemins qui commencent en ce point sont symétriques l'un de l'autre, si bien qu'ils sont disjoints ; ils ne peuvent donc que se terminer en $t = 1$. Cela montre qu'il y a des solutions à l'équation $\widetilde{F}(-x,t) = \widetilde{F}(x,t)$ en $t = 1$ et donc à l'équation $f(-x) = f(x)$. □

Bibliographie

[1] I. BÁRÁNY : *A short proof of Kneser's conjecture,* J. Combinatorial Theory, Ser. B **25** (1978), 325-326.

[2] K. BORSUK : *Drei Sätze über die n-dimensionale Sphäre,* Fundamenta Math. **20** (1933), 177-190.

[3] D. GALE : *Neighboring vertices on a convex polyhedron,* in : "Linear Inequalities and Related Systems" (H. W. Kuhn, A. W. Tucker, eds.), Princeton University Press, Princeton 1956, 255-263.

[4] J. E. GREENE : *A new short proof of Kneser's conjecture,* American Math. Monthly **109** (2002), 918-920.

[5] M. KNESER : *Aufgabe 360,* Jahresbericht der Deutschen Mathematiker-Vereinigung **58** (1955), 27.

[6] L. LOVÁSZ : *Kneser's conjecture, chromatic number, and homotopy,* J. Combinatorial Theory, Ser. B **25** (1978), 319-324.

[7] L. LYUSTERNIK & S. SHNIREL'MAN : *Topological Methods in Variational Problems (en russe),* Issledowatelskiĭ Institute Matematiki i Mechaniki pri O. M. G. U., Moscou, 1930.

[8] J. MATOUŠEK : *Using the Borsuk-Ulam Theorem. Lectures on Topological Methods in Combinatorics and Geometry,* Universitext, Springer-Verlag, Berlin 2003.

[9] J. MATOUŠEK : *A combinatorial proof of Kneser's conjecture,* Combinatorica **24** (2004), 163-170.

Amis et politiciens

Chapitre 39

Personne ne sait qui le premier a soulevé le problème suivant. Personne ne sait non plus qui lui a donné une connotation humaine. Le voici :

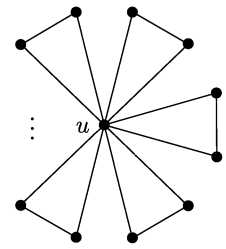
> *Dans un groupe d'individus, supposons que toute paire de personnes ait exactement un ami commun. Alors, il y a toujours un individu, le « politicien », qui est ami avec tout le monde.*

« Un sourire de politicien ».

Dans la littérature mathématique, ce résultat est connu sous le nom de *théorème de l'amitié*.

Avant d'en aborder la preuve, formulons le problème en termes de théorie des graphes. Interprétons les individus comme un ensemble de sommets V et joignons deux sommets par une arête si les personnes correspondantes sont des amis. Nous supposons tacitement que l'amitié est réciproque, c'est-à-dire que si u est l'ami de v, v est aussi l'ami de u. Nous supposons aussi que personne n'est son propre ami. Ainsi, le théorème prend la forme suivante :

Théorème. *Soit G soit un graphe fini dans lequel deux sommets ont exactement un voisin commun. Il existe alors un sommet adjacent à tous les autres sommets.*



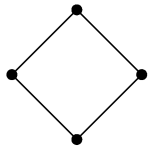
Un graphe en moulin à vent.

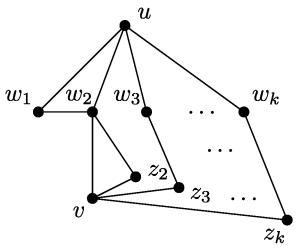
Remarquons qu'il existe bien des graphes qui possèdent cette propriété, comme dans la figure, où u est le politicien ; en fait, nous montrerons que ces « graphes en moulin à vent » sont les seuls graphes qui ont cette propriété. En effet, il n'est pas difficile de vérifier qu'en présence d'un politicien, seuls sont possibles les graphes en moulin à vent.

Cet énoncé ne se généralise pas aux graphes infinis. On peut, en effet, construire un contre-exemple par récurrence. On part d'un 5-cycle et l'on ajoute de manière répétitive un voisin commun à toutes les paires de sommets du graphe qui n'ont pas déjà un voisin commun. Cela conduit à un graphe d'amitié comportant un nombre infini dénombrable de sommets mais sans politicien.

Plusieurs preuves du théorème de l'amitié existent, mais la première, de Paul Erdős, Alfred Rényi et Vera Sós, est encore la meilleure.

■ **Preuve.** Supposons que l'affirmation soit fausse et que G soit un contre-exemple, c'est-à-dire qu'aucun sommet de G ne soit adjacent à tous les





autres sommets. Pour aboutir à une contradiction nous procédons en deux étapes. La première est combinatoire, la deuxième utilise l'algèbre linéaire.

(1) Nous affirmons que G est un graphe régulier donc $d(u) = d(v)$ pour tout $u, v \in V$.

Notons d'abord que la condition du théorème implique qu'il n'y a pas de cycle de longueur 4 dans G. Appelons cela la *C_4-condition*.

Prouvons dans un premier temps que deux sommets *non-adjacents* u et v ont des degrés égaux $d(u) = d(v)$. Supposons que $d(u) = k$, où $w_1, \ldots, w_k$ sont les voisins de u. L'un exactement des w_i, appelons le w_2, est adjacent à v, et w_2 est adjacent à l'un des autres w_i exactement, appelons le w_1, de sorte que nous obtenons la situation de la figure de gauche. Le sommet v a, avec w_1, le voisin commun w_2, et avec w_i $(i \geqslant 2)$ un voisin commun z_i $(i \geqslant 2)$. D'après la condition C_4, tous ces z_i doivent être distincts. Ainsi, $d(v) \geqslant k = d(u)$ et donc $d(u) = d(v) = k$ par symétrie.

Pour terminer la preuve de **(1)**, observons que tout sommet différent de w_2 n'est adjacent ni à u ni à v. Il a donc un degré égal à k, d'après ce que nous avons déjà montré. Cependant, puisque w_2 a aussi un non-voisin, son degré est également k et donc G est k-régulier.

La somme des degrés des k voisins de u est égale à k^2. Puisque chaque sommet (sauf u) a exactement un voisin commun avec u, nous avons compté chaque sommet une fois, sauf pour u, qui a été compté k fois. Le nombre total de sommets de G est donc :

$$n \;=\; k^2 - k + 1 \tag{1}$$

(2) Le reste de la preuve est une magnifique application de résultats standards d'algèbre linéaire. Notons d'abord que k doit être supérieur à 2, puisque pour $k \leqslant 2$ seuls $G = K_1$ et $G = K_3$ sont possibles d'après (1), les deux étant des graphes en moulin à vent triviaux. Considérons la matrice d'adjacence $A = (a_{ij})$, définie en page 278. D'après la partie **(1)**, toute ligne contient 1 exactement k fois. Par ailleurs, d'après la condition du théorème, étant données deux lignes, il y a exactement une colonne sur laquelle elles ont toutes deux un 1. Notons en outre que la diagonale principale est formée de 0. Nous avons par conséquent :

$$A^2 \;=\; \begin{pmatrix} k & 1 & \ldots & 1 \\ 1 & k & & 1 \\ \vdots & & \ddots & \vdots \\ 1 & \ldots & 1 & k \end{pmatrix} \;=\; (k-1)\,I + J$$

où I est la matrice identité et J la matrice composée de 1. On vérifie immédiatement que J a pour valeurs propres n (de multiplicité 1) et 0 (de multiplicité $n-1$). A^2 admet donc les valeurs propres $k - 1 + n = k^2$ (de multiplicité 1) et $k - 1$ (de multiplicité $n - 1$).

Puisque A est symétrique et donc diagonalisable, A a pour valeurs propres soit k soit $-k$ (de multiplicité 1), $\sqrt{k-1}$ (de multiplicité r) et $-\sqrt{k-1}$

(de multiplicité $n-1-r$). Supposons que r des valeurs propres soient égales à $\sqrt{k-1}$ et que s d'entre elles soient égales à $-\sqrt{k-1}$, avec $r+s=n-1$ (on peut remarquer que l'on a déjà calculé les valeurs propres d'une matrice de ce type au chapitre 25). Nous sommes presque arrivés au bout. Puisque la somme des valeurs propres de A est égale à la trace de A (qui est nulle), on obtient :

$$\pm k + r\sqrt{k-1} - s\sqrt{k-1} \;=\; 0$$

en particulier $r \neq s$ et :

$$\sqrt{k-1} \;=\; \pm\frac{k}{s-r}\,.$$

Or, si la racine carrée $\sqrt{m}$ d'un nombre entier est rationnelle, alors $\sqrt{m}$ un entier ! Dedekind a proposé une démonstration élégante de ce résultat en 1858 ; la voici. Soit n_0 le plus petit entier naturel tel que $n_0\sqrt{m} \in \mathbb{N}$. Si $\sqrt{m} \notin \mathbb{N}$, alors il existe $\ell \in \mathbb{N}$ tel que $0 < \sqrt{m} - \ell < 1$. En posant $n_1 := n_0(\sqrt{m} - \ell)$, on trouve finalement que $n_1 \in \mathbb{N}$ et que $n_1\sqrt{m} = n_0(\sqrt{m} - \ell)\sqrt{m} = n_0 m - \ell(n_0\sqrt{m}) \in \mathbb{N}$. Comme $n_1 < n_0$, on est conduit à une contradiction.

Revenons au résultat précédent et posons $h = \sqrt{k-1}$, si bien que :

$$h(s-r) \;=\; k \;=\; h^2+1$$

Puisque h divise h^2+1 et h^2, h doit être égal à 1, donc $k=2$, ce que nous avons déjà exclu. Nous sommes arrivés à une contradiction ; la preuve est terminée. □

Cependant, l'histoire n'est pas totalement finie. Formulons le théorème de la manière suivante. Soit G un graphe possédant la propriété suivante : entre deux sommets, il y a exactement un chemin de longueur 2. Il est clair que c'est une formulation équivalente de la condition d'amitié. Le théorème dit donc que de tels graphes ne peuvent être que des graphes en moulin à vent. Qu'en est-il si nous considérons des chemins qui ont une longueur supérieure à 2 ? Une conjecture d'Anton Kotzig affirme que la situation analogue est impossible.

La conjecture de Kotzig. *Soit $\ell > 2$. Il n'existe pas de graphe fini possédant la propriété suivante : entre deux sommets il existe précisément un chemin de longueur ℓ.*

Kotzig a vérifié lui-même sa conjecture pour $\ell \leqslant 8$. Dans [3] la conjecture a été démontrée jusqu'à $\ell = 20$ et A. Kostochka nous a dit récemment que la conjecture venait d'être prouvée pour tous les $\ell \leqslant 33$. Une démonstration générale semble pour le moment hors de portée.

Bibliographie

[1] P. ERDŐS, A. RÉNYI & V. SÓS : *On a problem of graph theory,* Studia Sci. Math. **1** (1966), 215-235.

[2] A. KOTZIG : *Regularly k-path connected graphs,* Congressus Numerantium **40** (1983), 137-141.

[3] A. KOSTOCHKA : *The nonexistence of certain generalized friendship graphs,* in : "Combinatorics" (Eger, 1987), Colloq. Math. Soc. János Bolyai **52**, North Holland, Amsterdam 1988, 341-356.

Les probabilités facilitent (parfois) le dénombrement

Chapitre 40

Comme nous avons commencé ce livre avec les premiers articles de Paul Erdős en théorie des nombres, nous le terminons en discutant de ce qui sera probablement considéré comme son héritage le plus durable, l'introduction, avec Alfred Rényi, de la *méthode probabiliste*. La voici énoncée de la manière la plus simple :

> *Si, dans un ensemble donné d'objets, la probabilité pour qu'un objet n'admette pas une certaine propriété est strictement inférieure à* 1, *alors il doit exister un objet admettant cette propriété.*

Nous avons là un résultat d'*existence*. Il peut être très difficile (et c'est souvent le cas) de trouver l'objet en question, mais on sait qu'il existe. Nous présentons ici trois exemples de plus en plus sophistiqués de la méthode probabiliste d'Erdős en terminant par une application très récente et particulièrement élégante.

Pour nous échauffer, considérons une famille $\mathcal{F}$ de sous-ensembles A_i, tous de cardinal $d \geqslant 2$, d'un ensemble fini X. On dit que $\mathcal{F}$ est 2-*colorable* s'il existe une coloration de X à deux couleurs telle que dans chaque ensemble A_i les deux couleurs apparaissent. Il est évident que l'on ne peut pas colorier n'importe quelle famille de cette manière. À titre d'exemple, prenons *tous* les sous-ensembles de taille d d'un $(2d-1)$-ensemble X. Dans ce cas, peu importe comment on 2-colorie X, il y a toujours d éléments qui sont coloriés de la même façon. D'autre part, il est également clair que chaque sous-famille d'une famille 2-colorable de d-ensembles est elle-même 2-colorable. Par conséquent, nous nous intéressons au *plus petit* nombre $m = m(d)$ pour lequel il existe une famille de m ensembles qui n'est pas 2-colorable. Autrement dit, $m(d)$ est le plus petit nombre qui garantit que *chaque* famille ayant moins de $m(d)$ ensembles est 2-colorable.

Théorème 1. *Toute famille ayant au plus* 2^{d-1} d*-ensembles est* 2*-colorable, c'est-à-dire* $m(d) > 2^{d-1}$.

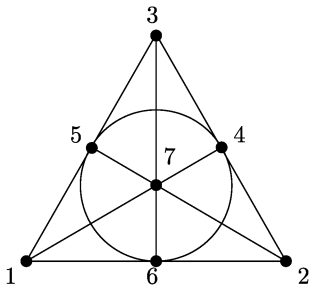
■ **Preuve.** Supposons que $\mathcal{F}$ soit une famille de d-ensembles comprenant au plus 2^{d-1} ensembles. Colorions X de manière aléatoire avec deux couleurs, toutes les colorations étant équiprobables. Pour chaque ensemble $A \in \mathcal{F}$, soit E_A l'événement suivant : « tous les éléments de A sont coloriés de la même manière ». Puisqu'il y a exactement deux colorations de ce type :

$$\text{Prob}(E_A) \;=\; \left(\tfrac{1}{2}\right)^{d-1}$$

et par conséquent si $m = |\mathcal{F}| \leqslant 2^{d-1}$ (notons que les événements E_A ne sont pas disjoints) :

$$\text{Prob}(\bigcup_{A\in\mathcal{F}} E_A) \;<\; \sum_{A\in\mathcal{F}} \text{Prob}(E_A) \;=\; m\,(\tfrac{1}{2})^{d-1} \;\leqslant\; 1$$

Nous concluons qu'il existe une 2-coloration de X sans ensemble unicolore, ce qui est exactement la condition voulue. □

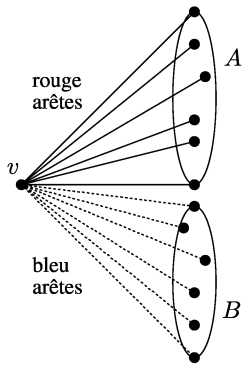


Erdős a aussi établi une borne supérieure pour $m(d)$, grossièrement égale à $d^2 2^d$, en utilisant encore une méthode probabiliste, considérant cette fois des ensembles aléatoires et une coloration fixe. On connaît seulement les valeurs exactes des deux premiers $m(2) = 3$, $m(3) = 7$. Bien sûr, $m(2) = 3$ est réalisé par le graphe K_3, alors que la configuration de Fano implique que $m(3) \leqslant 7$. Ici $\mathcal{F}$ se compose des sept 3-ensembles de la figure (incluant le cercle $\{4,5,6\}$). Le lecteur peut s'amuser à montrer que $\mathcal{F}$ requiert 3 couleurs. Pour prouver que toutes les familles de six 3-ensembles sont 2-colorables, et par conséquent $m(3) = 7$, il faut un petit peu plus d'attention.

L'exemple suivant est un classique dans ce domaine : les nombres de Ramsey. Considérons le graphe complet K_N sur N sommets. On dit que K_N présente la propriété (m,n) si, quelle que soit la façon dont on colorie les arêtes de K_N en rouge et bleu, il y a toujours un sous-graphe complet sur m sommets dont toutes les arêtes sont coloriées en rouge ou un sous-graphe complet sur n sommets dont toutes les arêtes sont coloriées en bleu. Il est clair que si K_N a la propriété (m,n), il en est de même pour tout K_s tel que $s \geqslant N$. Ainsi, comme dans le premier exemple, nous cherchons le *plus petit* nombre N (s'il existe) possédant cette propriété : c'est le *nombre de Ramsey* $R(m,n)$.

Pour commencer, nous avons évidemment $R(m,2) = m$ parce que, soit toutes les arêtes de K_m sont rouges, soit il y a une arête bleue et il en résulte un K_2 bleu. Par symétrie, $R(2,n) = n$. Supposons maintenant que $R(m-1,n)$ et $R(m,n-1)$ existent. Nous allons prouver alors que $R(m,n)$ existe et que :

$$R(m,n) \;\leqslant\; R(m-1,n) \;+\; R(m,n-1) \qquad (1)$$



Notons $N = R(m-1,n) + R(m,n-1)$ et considérons une coloration arbitraire rouge et bleu de K_N. Étant donné un sommet v, considérons l'ensemble A des sommets reliés à v par une arête rouge et l'ensemble B des sommets reliés par une arête bleue. Puisque $|A| + |B| = N - 1$, soit $|A| \geqslant R(m-1,n)$, soit $|B| \geqslant R(m,n-1)$. Supposons que $|A| \geqslant R(m-1,n)$, l'autre cas se traitant de manière analogue. D'après la définition de $R(m-1,n)$, soit il existe dans A un sous-ensemble A_R de taille $m-1$ dont toutes les arêtes sont coloriées en rouge, ce qui donne avec v un K_m rouge, soit il y a un sous-ensemble A_B de taille n dont toutes les arêtes sont coloriées en bleu. Il en résulte que K_N vérifie la propriété (m,n) et l'affirmation (1) en découle.

En combinant (1) avec les valeurs de départ $R(m,2)=m$ et $R(2,n)=n$, on obtient d'après la formule classique de récurrence sur les coefficients binomiaux :

$$R(m,n) \leqslant \binom{m+n-2}{m-1}$$

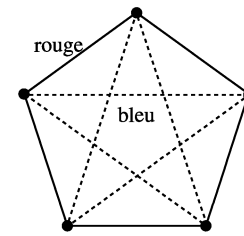
et en particulier :

$$R(k,k) \leqslant \binom{2k-2}{k-1} = \binom{2k-3}{k-1} + \binom{2k-3}{k-2} \leqslant 2^{2k-3} \quad (2)$$

On est particulièrement intéressé par un minorant de $R(k,k)$. Cela signifie trouver un $N < R(k,k)$ aussi grand que possible tel qu'il *existe* une coloration des arêtes sans pour autant qu'il en résulte un K_k rouge ou bleu. C'est ici que la méthode probabiliste entre en jeu .

Théorème 2. *Pour tout $k \geqslant 2$, les nombres de Ramsey vérifient l'inégalité suivante :*

$$R(k,k) \geqslant 2^{\frac{k}{2}}.$$



■ **Preuve.** Nous savons déjà que $R(2,2)=2$. En utilisant (2), nous pouvons écrire $R(3,3) \leqslant 6$; le pentagone colorié de la figure ci-contre montre que $R(3,3)=6$.

Supposons maintenant que $k \geqslant 4$ et que $N < 2^{\frac{k}{2}}$; considérons toutes les colorations rouge-bleu, en coloriant chaque arête indépendamment en rouge ou en bleu avec une probabilité $\frac{1}{2}$. Alors, toutes les colorations sont équiprobables, de probabilité $2^{-\binom{n}{2}}$. Soit A un ensemble de sommets de taille k. La probabilité de l'événement A_R « les arêtes de A sont toutes coloriées en rouge » est donc $2^{-\binom{k}{2}}$. Par conséquent, la probabilité p_R qu'un k-ensemble *quelconque* soit colorié entièrement en rouge est majorée par :

$$p_R = \text{Prob}\Big(\bigcup_{|A|=k} A_R\Big) \leqslant \sum_{|A|=k} \text{Prob}(A_R) = \binom{N}{k} 2^{-\binom{k}{2}}$$

Maintenant si $N < 2^{\frac{k}{2}}$ et $k \geqslant 4$, en utilisant $\binom{N}{k} \leqslant \frac{N^k}{2^{k-1}}$ si $k \geqslant 2$ (voir page 13), on trouve :

$$\binom{N}{k} 2^{-\binom{k}{2}} \leqslant \frac{N^k}{2^{k-1}} 2^{-\binom{k}{2}} < 2^{\frac{k^2}{2}-\binom{k}{2}-k+1} = 2^{-\frac{k}{2}+1} \leqslant \frac{1}{2}$$

Par conséquent $p_R < \frac{1}{2}$ et par symétrie $p_B < \frac{1}{2}$, (probabilité que k sommets aient toutes les arêtes qui les relient de couleur bleu). On en conclut que $p_R + p_B < 1$ pour $N < 2^{\frac{k}{2}}$. Il *doit* donc y avoir une coloration sans K_k rouge ou bleu, ce qui signifie que K_N n'a pas la propriété (k,k). □

Bien sûr, il y a un grand écart entre les bornes supérieure et inférieure de $R(k,k)$. Néanmoins, aucune borne inférieure avec un meilleur exposant n'a été trouvée pour k quelconque, pendant les cinquante ans qui suivirent

le résultat d'Erdős. En fait, personne n'a été capable d'établir une borne inférieure de la forme $R(k,k) > 2^{(\frac{1}{2}+\varepsilon)k}$ ni une borne supérieure de la forme $R(k,k) < 2^{(2-\varepsilon)k}$ avec un $\varepsilon > 0$ fixé.

Notre troisième résultat est une autre belle illustration de la méthode probabiliste. Considérons un graphe G sur n sommets et son nombre chromatique $\chi(G)$. Si $\chi(G)$ est élevé, c'est-à-dire si l'on a besoin d'un grand nombre de couleurs, on pourrait penser que G contient un grand sous-graphe complet. C'est cependant loin d'être vrai. Déjà, dans les années quarante, Blanche Descartes avait construit des graphes ayant des nombres chromatiques arbitrairement élevés et sans triangle, c'est-à-dire dont chaque cycle a une longueur supérieure ou égale à 4 ; d'autres auteurs ont fait de même (voir encadré).

Cependant, dans ces exemples, il y a beaucoup de cycles de longueur 4. Peut-on faire mieux ? Peut-on stipuler qu'il n'y a pas de cycle de petite longueur et que, malgré tout, le nombre chromatique est arbitrairement élevé ? Oui ! Pour préciser les choses, appelons *circonférence* $\gamma(G)$ de G la longueur d'un cycle minimal de G ; on a le théorème suivant, que Paul Erdős fut le premier à démontrer.

Théorème 3. *Pour tout $k \geqslant 2$, il existe un graphe G à nombre chromatique $\chi(G) > k$ et de circonférence $\gamma(G) > k$.*

La stratégie est semblable à celle des preuves précédentes : on considère un certain espace probabilisé sur des graphes, on montre encore que la probabilité d'avoir $\chi(G) \leqslant k$ est inférieure à $\frac{1}{2}$ et de même que la probabilité d'avoir $\gamma(G) \leqslant k$ est inférieure à $\frac{1}{2}$. Par conséquent, il doit exister un graphe avec les propriétés désirées.

■ **Preuve.** Soit $V = \{v_1, v_2, \ldots, v_n\}$ l'ensemble des sommets et soit p un nombre fixe entre 0 et 1(qui devra être soigneusement choisi plus tard). L'espace probabilisé $\mathcal{G}(n,p)$ se compose de tous les graphes sur V, les arêtes individuelles apparaissant avec une probabilité p, indépendamment les unes des autres. En d'autres termes, on considère une expérience de Bernoulli où chaque arête est tirée avec une probabilité p. Par exemple, la probabilité $\text{Prob}(K_n)$ d'un graphe complet est $\text{Prob}(K_n) = p^{\binom{n}{2}}$. En général, nous avons $\text{Prob}(H) = p^m(1-p)^{\binom{n}{2}-m}$ si le graphe H sur V a exactement m arêtes.

Considérons d'abord le nombre chromatique $\chi(G)$. On désigne par $\alpha = \alpha(G)$ le *nombre d'indépendance*, c'est-à-dire la taille maximale d'un ensemble indépendant de G. Puisque dans une coloration ayant $\chi = \chi(G)$ couleurs, toutes les classes de couleurs sont indépendantes (et par conséquent de taille $\leqslant \alpha$), $\chi\alpha \geqslant n$. Ainsi, si α est petit par rapport à n, χ doit être grand, ce que nous voulons.

Soit r vérifiant $2 \leqslant r \leqslant n$. La probabilité qu'un r-ensemble fixé de V soit indépendant est $(1-p)^{\binom{r}{2}}$; nous concluons, par le même argument que

pour la démonstration du théorème 2, que :

$$\begin{aligned}\text{Prob}(\alpha \geqslant r) &\leqslant \binom{n}{r}(1-p)^{\binom{r}{2}} \\ &\leqslant n^r(1-p)^{\binom{r}{2}} = (n(1-p)^{\frac{r-1}{2}})^r \leqslant (ne^{-p(r-1)/2})^r\end{aligned}$$

puisque $1-p \leqslant e^{-p}$ pour tout p.

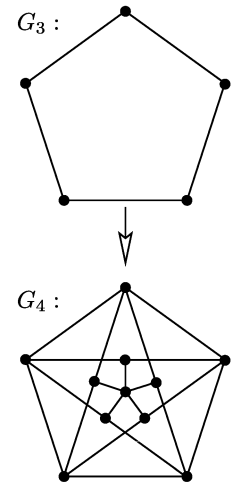
Graphes sans triangle présentant un nombre chromatique élevé

Voici une suite de graphes sans triangle $G_3, G_4, \ldots$ tels que :

$$\chi(G_n) = n.$$

On commence par le 5-cycle $G_3 = C_5$; on a $\chi(G_3) = 3$. Supposons que nous ayons déjà construit G_n sur l'ensemble de sommets V. Le nouveau graphe G_{n+1} admet $V \cup V' \cup \{z\}$ pour ensemble de sommets, les sommets $v' \in V'$ étant en bijection avec les $v \in V$, z étant encore un autre sommet. Les arêtes de G_{n+1} se classent en 3 catégories : d'abord, nous prenons toutes les arêtes de G_n ; ensuite, chaque sommet v' est relié à chaque voisin de son image v dans G_n ; enfin z est relié à tous les $v' \in V'$. Ainsi, à partir de $G_3 = C_5$ on obtient pour G_4 le *graphe de Mycielski*.

Il est clair que G_{n+1} n'a pas de triangles. Pour montrer que $\chi(G_{n+1}) = n+1$ procédons par récurrence sur n. Prenons une n-coloration quelconque de G_n et considérons une classe de couleur C. Il doit exister un sommet $v \in C$ qui est adjacent à au moins un sommet de chacune des autres classes de couleur ; sinon, on pourrait distribuer les autres sommets de C dans les $n-1$ autres classes de couleur, ce qui impliquerait $\chi(G_n) \leqslant n-1$. Toutefois, il est clair que v' (le sommet de V' correspondant à v) doit recevoir la même couleur que v dans cette n-coloration. Donc, toutes les n couleurs apparaissent dans V' : on a besoin d'une nouvelle couleur pour z.



Construction du graphe de Mycielski.

Étant donné un $k > 0$ fixé, choisissons maintenant $p = n^{-\frac{k}{k+1}}$. Nous allons montrer que pour n assez grand :

$$\text{Prob}\Big(\alpha \geqslant \frac{n}{2k}\Big) < \frac{1}{2} \tag{3}$$

En effet, puisque $n^{\frac{1}{k+1}}$ augmente plus vite que $\ln n$, on a $n^{\frac{1}{k+1}} \geqslant 6k \ln n$ pour n assez grand donc $p \geqslant 6k\frac{\ln n}{n}$. Avec $r = \lceil \frac{n}{2k} \rceil$ cela implique que $pr \geqslant 3 \ln n$, ainsi :

$$ne^{-p(r-1)/2} = ne^{-\frac{pr}{2}}e^{\frac{p}{2}} \leqslant ne^{-\frac{3}{2}\ln n}e^{\frac{1}{2}} = n^{-\frac{1}{2}}e^{\frac{1}{2}} = (\tfrac{e}{n})^{\frac{1}{2}}$$

qui converge vers 0 lorsque n tend vers l'infini. Par conséquent (3) est vérifiée pour tout $n \geqslant n_1$.

Examinons maintenant le second paramètre $\gamma(G)$. Pour k fixé, montrons qu'il n'y a pas trop de cycles de longueur $\leqslant k$. Soit i un entier compris entre 3 et k, et soit $A \subseteq V$ un i-ensemble fixé. Le nombre de i-cycles possibles sur A est clairement le nombre de permutations cycliques de A divisé par 2 (puisque l'on peut parcourir le cycle dans les deux sens) ; il est donc égal à $\frac{(i-1)!}{2}$. Le nombre total de i-cycles possibles est donc $\binom{n}{i}\frac{(i-1)!}{2}$; en outre chaque cycle C apparaît avec une probabilité p^i. Soit X la variable aléatoire qui compte le nombre de cycles de longueur $\leqslant k$. Pour estimer X nous avons à notre disposition deux outils simples mais beaux. Le premier est la linéarité de l'espérance et le second est l'inégalité de Markov pour les variables aléatoires non-négatives, qui s'écrit :

$$\text{Prob}(X \geqslant a) \leqslant \frac{EX}{a}$$

où EX est l'espérance de X (se reporter à l'appendice du chapitre 15 en ce qui concerne ces deux outils).

Soit X_C la variable aléatoire indicatrice du cycle C de longueur i par exemple. En d'autres termes, posons $X_C = 1$ ou 0 selon que C apparaît ou non dans le graphe ; on a $EX_C = p^i$. Puisque X dénombre tous les cycles de longueur $\leqslant k$, nous avons $X = \sum X_C$ et, par linéarité :

$$EX = \sum_{i=3}^{k} \binom{n}{i} \frac{(i-1)!}{2} p^i \leqslant \frac{1}{2} \sum_{i=3}^{k} n^i p^i \leqslant \frac{1}{2}(k-2)n^k p^k$$

la dernière inégalité étant une conséquence du fait que $np = n^{\frac{1}{k+1}} \geqslant 1$. En appliquant maintenant l'inégalité de Markov avec $a = \frac{n}{2}$, on obtient :

$$\text{Prob}(X \geqslant \tfrac{n}{2}) \leqslant \frac{EX}{n/2} \leqslant (k-2)\frac{(np)^k}{n} = (k-2)n^{-\frac{1}{k+1}}$$

Puisque le membre droit tend vers 0 lorsque n tend vers l'infini, on voit que $p(X \geqslant \frac{n}{2}) < \frac{1}{2}$ si $n \geqslant n_2$.

Nous sommes presque arrivés au bout de nos peines. Notre analyse implique que pour $n \geqslant \max(n_1, n_2)$ il existe un graphe H sur n sommets avec $\alpha(H) < \frac{n}{2k}$ et moins de $\frac{n}{2}$ cycles de longueur inférieure ou égale à k. Supprimons un sommet de chacun de ces cycles et considérons le graphe G obtenu. On a donc $\gamma(G) > k$. Puisque G contient plus de $\frac{n}{2}$ sommets et vérifie $\alpha(G) \leqslant \alpha(H) < \frac{n}{2k}$, on obtient finalement :

$$\chi(G) \geqslant \frac{n/2}{\alpha(G)} \geqslant \frac{n}{2\alpha(H)} > \frac{n}{n/k} = k$$

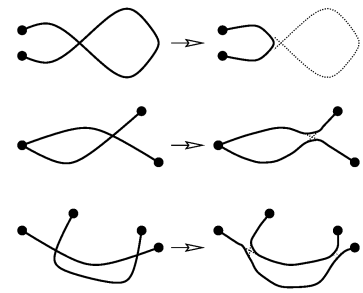
□

On connaît des constructions explicites de graphes (de tailles énormes) à circonférence et nombre chromatique élevés. En revanche, on ne sait pas

fabriquer de colorations rouges/bleus n'ayant pas de grandes cliques monochromatiques, dont l'existence est pourtant assurée par le théorème 2. Ce qui reste frappant à propos de la démonstration d'Erdős, c'est qu'elle prouve l'existence de graphes relativement petits à nombre chromatique et à circonférence élevés.

Pour terminer notre excursion dans le monde des probabilités, traitons un résultat important de la théorie géométrique des graphes (qui remonte encore une fois à Paul Erdős) dont la Preuve étonnante du Grand Livre est d'un cru récent.

Considérons un graphe simple $G(V, E)$ à n sommets et m arêtes. Nous voulons plonger G dans le plan comme nous l'avons fait pour des graphes planaires. Nous savons depuis le chapitre 12, comme conséquence de la formule d'Euler, qu'un graphe planaire simple G à n sommets possède au plus $3n - 6$ arêtes. Par conséquent, si m est supérieur à $3n - 6$, il doit y avoir des intersections d'arêtes. Le *nombre d'intersections* $\mathrm{cr}(G)$ est donc défini de façon naturelle : c'est le plus petit nombre d'intersections parmi toutes les plongements de G. Ainsi, $\mathrm{cr}(G) = 0$ si et seulement si G est planaire.

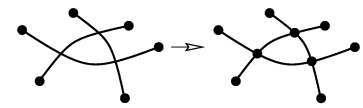


Étant donné un tel plongement minimal, les trois situations suivantes sont écartées :

- aucune arête ne peut se croiser elle-même ;
- les arêtes qui ont un sommet commun ne peuvent pas se croiser ;
- deux arêtes ne peuvent se croiser deux fois.

En effet, dans chacun de ces cas, nous pouvons dessiner le même graphe avec moins d'intersections, en utilisant les opérations qui sont indiquées dans notre figure. À partir de maintenant, nous considérons que tout plongement suit ces règles.

Supposons que G soit plongé dans $\mathbb{R}^2$ avec $\mathrm{cr}(G)$ intersections. Nous pouvons immédiatement en déduire une borne inférieure du nombre d'intersections. Considérons le graphe H suivant : les sommets de H sont ceux de G auxquels on ajoute tous les points d'intersection, et les arêtes sont tous les morceaux des arêtes originales en parcourant le graphe de point d'intersection en point d'intersection.

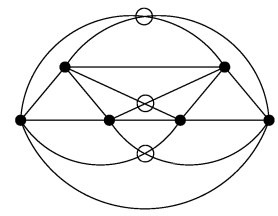


Le nouveau graphe H est désormais planaire et simple, (cela se déduit des trois hypothèses précédentes). Le nombre de sommets dans H est $n+\mathrm{cr}(G)$ et le nombre d'arêtes est $m+2\mathrm{cr}(G)$, puisque chaque nouveau sommet a un degré 4. En utilisant la borne sur le nombre d'arêtes des graphes planaires, on trouve :

$$m + 2\,\mathrm{cr}(G) \;\leqslant\; 3(n + \mathrm{cr}(G)) - 6$$

c'est-à-dire,

$$\mathrm{cr}(G) \;\geqslant\; m - 3n + 6 \qquad (4)$$



Par exemple, pour le graphe complet K_6 on obtient :

$$\mathrm{cr}(K_6) \geqslant 15 - 18 + 6 = 3$$

et en fait, il existe un plongement avec seulement 3 intersections.

La borne fournie par (4) est relativement bonne lorsque m est linéaire en n, mais lorsque m est plus grand, la situation change et fait l'objet du théorème qui suit.

Théorème 4. *Soit G un graphe simple à n sommets et m arêtes, tel que $m \geqslant 4n$. Alors :*

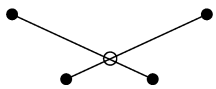
$$\mathrm{cr}(G) \geqslant \frac{1}{64}\frac{m^3}{n^2}$$

L'histoire de ce résultat, appelé *lemme d'intersection*, est assez intéressante. Il a été conjecturé par Erdős et Guy en 1973 ($\frac{1}{64}$ étant remplacé par une constante c). les premières preuves furent données par Leighton en 1982 (avec $\frac{1}{100}$ à la place de $\frac{1}{64}$) et indépendamment par Ajtai, Chvátal, Newborn et Szemerédi. Le lemme d'intersection n'était guère connu (en fait, beaucoup de mathématiciens le connaissaient toujours en tant que conjecture bien après l'apparition des premières preuves), jusqu'à ce que László Székely montre son utilité dans un bel article, en l'appliquant à différents problèmes géométriques extrémaux difficiles. La preuve que nous présentons maintenant est le fruit de conversations par courrier électronique entre Bernard Chazelle, Micha Sharir et Emo Welzl. Elle appartient sans aucun doute au Grand Livre.

■ **Preuve.** Considérons un plongement de G, c'est-à-dire une représentation plane minimale de G. Soit p un nombre réel compris entre 0 et 1 (qui sera choisi plus tard). On génère un sous-ensemble des sommets de G en effectuant un tirage indépendant de chaque sommet de G avec une probabilité p. Le graphe induit ainsi obtenu est appelé G_p.

Soient n_p, m_p et X_p les variables aléatoires qui comptent le nombre de sommets, d'arêtes et d'intersections dans G_p. Puisque l'on a $\mathrm{cr}(G) - m + 3n \geqslant 0$ d'après (4) pour *tout* graphe, l'espérance doit satisfaire :

$$E(X_p - m_p + 3n_p) \geqslant 0$$



Calculons maintenant chaque espérance $E(n_p)$, $E(m_p)$ et $E(X_p)$. Il est clair que $E(n_p) = pn$ et $E(m_p) = p^2m$, puisqu'une arête est dans G_p si et seulement si ses deux extrémités y sont. Finalement, $E(X_p) = p^4\mathrm{cr}(G)$, puisqu'une intersection est dans G_p si et seulement si les quatre sommets (distincts !) y figurent.

Par linéarité de l'espérance, nous obtenons donc que :

$$0 \leqslant E(X_p) - E(m_p) + 3E(n_p) = p^4\mathrm{cr}(G) - p^2m + 3pn$$

c'est-à-dire :

$$\mathrm{cr}(G) \geqslant \frac{p^2m - 3pn}{p^4} = \frac{m}{p^2} - \frac{3n}{p^3} \qquad (5)$$

Voici maintenant l'argument final. En posant $p = \frac{4n}{m}$ (qui vaut 1 au plus par hypothèse), (5) devient :

$$\mathrm{cr}(G) \geqslant \frac{1}{64}\left[\frac{4m}{(n/m)^2} - \frac{3n}{(n/m)^3}\right] = \frac{1}{64}\frac{m^3}{n^2}$$

ce que nous voulions établir. □

Paul Erdős aurait beaucoup aimé cette preuve.

Bibliographie

[1] M. AJTAI, V. CHVÁTAL, M. NEWBORN & E. SZEMERÉDI : *Crossing-free subgraphs,* Annals of Discrete Mathematics **12** (1982), 9-12.

[2] N. ALON & J. SPENCER : *The Probabilistic Method,* Wiley-Interscience 1992.

[3] P. ERDŐS : *Some remarks on the theory of graphs,* Bulletin Amer. Math. Soc. **53** (1947), 292-294.

[4] P. ERDŐS : *Graph theory and probability,* Canadian J. Math. **11** (1959), 34-38.

[5] P. ERDŐS : *On a combinatorial problem I,* Nordisk Math. Tidskrift **11** (1963), 5-10.

[6] P. ERDŐS & R. K. GUY : *Crossing number problems,* Amer. Math. Monthly **80** (1973), 52-58.

[7] P. ERDŐS & A. RÉNYI : *On the evolution of random graphs,* Magyar Tud. Akad. Mat. Kut. Int. Közl. **5** (1960), 17-61.

[8] T. LEIGHTON : *Complexity Issues in VLSI,* MIT Press, Cambridge MA 1983.

[9] L. A. SZÉKELY : *Crossing numbers and hard Erdős problems in discrete geometry,* Combinatorics, Probability, and Computing **6** (1997), 353-358.

À propos des illustrations

Nous sommes heureux d'avoir eu le privilège d'illustrer ce volume par de magnifiques dessins originaux de Karl Heinrich Hofmann (Darmstadt). Qu'il en soit remercié !

Le polyèdre régulier en page 84 et le schéma de la sphère flexible dépliée en page 94 sont de WAF Ruppert (Vienne).
Jürgen Richter-Gebert a fourni les deux illustrations de la page 86 et Ronald Wotzlaw a fourni les fichiers postscript de la page 152.

La page 261 montre une vue extérieure et un plan du Musée d'art Weisman de Minneapolis conçu par Frank Gehry. La photo de la façade ouest a été fournie par Chris Faust. Le plan est celui de la Galerie Dolly Fiterman Riverview, qui se trouve derrière la façade ouest.

Les portraits de Bertrand, Cantor, Erdős, Euler, Fermat, Herglotz, Hermite, Hilbert, Pólya, Littlewood et Sylvester proviennent des archives photographiques de l'Institut de Recherche Mathématiques d'Oberwolfach, avec son autorisation. (Merci à Annette Disch !)

Le portrait de Gauss en page 27 est une lithographie de Siegfried Detlev Bendixen publiée dans les Astronomische Nachrichten de 1828, telle qu'on la trouve sur Wikipedia.

Le portrait de Hermite est issu du premier volume de ses œuvres complètes.

Le portrait d'Eisenstein a été obtenue avec l'aimable concours du Professeur Karine Reich ; c'est une reproduction issue d'une collection de portraits de la Société de mathématiques de Hambourg.

Les reproductions des timbres à l'effigie de Buffon, Chebyshev, Euler et Ramanujan proviennent du site web de Jeff Miller consacré aux timbres mathématiques (http ://jeff560.tripod.com). Nous le remercions de nous avoir accordé la permission de reproduire ces timbres.

La photo de Claude Shannon a été fournie par le musée du MIT et est reproduite avec son aimable autorisation.

Le portrait de Cayley est tiré du *Photoalbum für Weierstraß* (édité par Reinhard Bölling, Vieweg 1994), avec l'autorisation de la Kunstbibliothek, Staatliche Museen zu Berlin, Preussischer Kulturbesitz.

Le portrait de Cauchy est reproduit avec l'autorisation des Collections de l'École Polytechnique, Paris.

Le portrait de Fermat est tiré de l'ouvrage de Stefan Hildebrandt et Anthony Tromba : *The Parsimonious Universe. Shape and Form in the Natural World*, Springer-Verlag, New York, 1996.

Le portrait d'Ernst Witt est tiré du volume 426 (1992) du *Journal für die*

Reine und Angewandte Mathematik, avec l'autorisation de Walter de Gruyter Publishers. Le cliché a été pris vers 1941.

La photo de Karol Borsuk a été prise en 1967 par Isaac Namioka ; elle est reproduite avec son aimable autorisation .

Nous remercions le Dr. Peter Sperner (Braunschweig) pour le portrait de son père et Vera Sós pour la photographie de Paul Turán.

Merci à Noga Alon pour le portrait de A. Nilli !

Index