Université Denis Diderot Paris 7, Semestre Février-juin 2005

Cours d'arithmétique (M1)

(Marc Hindry : hindry@math.jussieu.fr)

PLAN

BIBLIOGRAPHIE (commentée subjectivement)

D. Perrin, *Cours d'algèbre*, Ellipses. (*excellent livre particulièrement recommandé pour la préparation à l'agrégation*)

M. Demazure, *Cours d'algèbre*, Cassini, Paris, 1997. (*surtout pour la partie "structures finies" et algorithmes*)

K. Ireland, M. Rosen, *A classical introduction to modern number theory*, Graduate texts in math. 84, Springer, 1982. (*comme le titre l'indique ...*)

P. Samuel, *théorie algébrique des nombres*, Hermann. (*traite les anneaux de Dedekind à un niveau un peu plus élevé que ce cours – deuxième partie – mais si bien écrit. Un classique*)

A. Baker, *Transcendental Number Theory*, Cambridge University Press, 1975. (*les toutes premières pages démontrent la transcendance de e et π, le reste du livre est plus spécialisé*)

... et mes livres préférés :

G. H. Hardy and E. M.Wright, *An introduction to the theory of numbers*, Oxford University Press, 4th ed.,1960. (*présentation de la plupart des sujets de théorie des nombres à un niveau élémentaire, très attrayant!*)

J.-P. Serre, *Cours d'arithmétique*, Presses Universitaires de France, 1970. (*Un classique insurpassable. J'utiliserai le début sur les corps finis, la loi de réciprocité et la partie théorie analytique pour les séries et théorème de Dirichlet*)

Borevich, Shafarevich, *Théorie des nombres* (traduit du russe), Gauthier-Villars. (*Un très joli livre qui, même s'il démarre à un niveau élémentaire, est d'un niveau plus élevé que ce cours*)

Université Denis Diderot Paris 7, Semestre Février-juin 2004

Cours de maîtrise : arithmétique

Première partie : Structures finies $\mathbf{Z}/n\mathbf{Z}$, $(\mathbf{Z}/n\mathbf{Z})^*$, $\mathbf{F}_q$, $\mathbf{F}_q^*$.

A. Rappels sur $\mathbf{Z}/n\mathbf{Z}$, $(\mathbf{Z}/n\mathbf{Z})^*$, $\mathbf{F}_q$, $\mathbf{F}_q^*$.
B. La structure des groupes $(\mathbf{Z}/n\mathbf{Z})^*$ et $\mathbf{F}_q^*$.
C. Symboles de Legendre et Jacobi.
D. Sommes de Gauss.
E. Applications au nombre de solutions d'équations.

A. Rappels sur $\mathbf{Z}/n\mathbf{Z}$, $(\mathbf{Z}/n\mathbf{Z})^*$, $\mathbf{F}_q$, $\mathbf{F}_q^*$.

La théorie des congruences amène, pour chaque entiers $n \geq 2$, à considérer l'anneau $\mathbf{Z}/n\mathbf{Z}$ ainsi que le groupe de ses éléments inversibles (pour la multiplication) $(\mathbf{Z}/n\mathbf{Z})^*$. Pour chaque puissance d'un nombre premier $q = p^f$, il existe un corps fini de cardinal q, unique à isomorphisme près, noté $\mathbf{F}_q$. Nous rappelons la construction de ces objets et précisons leurs principales propriétés.

Le groupe $\mathbf{Z}$ est l'unique groupe (à isomorphisme près) qui est cyclique (engendré par un élément) et infini. Tous ses sous-groupes sont du type $m\mathbf{Z}$ pour $m \geq 0$. L'ensemble $\mathbf{Z}$ est également muni d'une multiplication qui en fait un anneau commutatif. Dans cet anneau on a la notion de divisibilité et de PGCD et PPCM. Dans le cas de $\mathbf{Z}$ la notion d'*idéal* coïncide avec celle de sous-groupe. On peut en déduire facilement le théorème suivant

Théorème. (Bézout) *Soit $m, n \in \mathbf{Z}$ et soit d leur PGCD, alors il existe $u, v \in \mathbf{Z}$ tels que*

$$d = um + vn.$$

Preuve. L'ensemble $H := m\mathbf{Z} + n\mathbf{Z} = \{um + vn \mid u, v \in \mathbf{Z}\}$ est clairement un sous-groupe; il est donc de la forme $d'\mathbf{Z}$ et il existe u, v tels que $d' = um + vn$. Comme d divise m et n, on voit que d divise $um + vn = d'$ mais m, n appartiennent à H donc d' divise m et n donc d' divise également d et on conclut que $d\mathbf{Z} = d'\mathbf{Z}$ (on aura même $d = d'$ si l'on a pris soin de les prendre tous les deux positifs). □

Le groupe $\mathbf{Z}/n\mathbf{Z}$ est l'unique groupe cyclique à n éléments (à isomorphisme près) i.e. engendré par un élément d'ordre n. On peut déjà étudier ses générateurs

Proposition. *Soit $m \in \mathbf{Z}$ et $\bar{m}$ sa classe dans $\mathbf{Z}/n\mathbf{Z}$, les trois propriétés suivantes sont équivalentes*
(i) L'élément $\bar{m}$ est un générateur de $\mathbf{Z}/n\mathbf{Z}$.
(ii) Les éléments m et n sont premiers entre eux.
(iii) L'élément $\bar{m}$ est inversible modulo n, c'est-à-dire qu'il existe $m' \in \mathbf{Z}$ tel que $mm' \equiv 1 \bmod n$ ou encore $\bar{m}\bar{m}' = 1 \in \mathbf{Z}/n\mathbf{Z}$.

Preuve. Supposons que $\bar{m}$ engendre $\mathbf{Z}/n\mathbf{Z}$ alors il existe $m' \in \mathbf{Z}$ tel que $m'\bar{m} = 1 \in \mathbf{Z}/n\mathbf{Z}$; ainsi $mm' \equiv 1 \bmod n$ ce qui signifie que m est inversible modulo n. Si $mm' \equiv 1 \bmod n$ alors $mm' = 1 + an$ et donc m est premier avec n. Si m est premier avec n alors, d'après le théorème de Bézout, il existe a, b tels que $am + bn = 1$ donc $a\bar{m} = 1 \in \mathbf{Z}/n\mathbf{Z}$ et donc $\bar{m}$ engendre $\mathbf{Z}/n\mathbf{Z}$. □

Exercice. Montrer que si, dans un groupe commutatif, l'ordre de x_1 est d_1, l'ordre de x_2 est d_2 avec d_1 et d_2 premiers entre eux, alors l'ordre de x_1x_2 est d_1d_2. Montrer également que si, dans un groupe cyclique, l'ordre de x_1 est d_1, l'ordre de x_2 est d_2, alors l'ordre du sous-groupe engendré par x_1 et x_2 est égal au PPCM de d_1 et d_2.

Le groupe des éléments inversibles de l'anneau $\mathbf{Z}/n\mathbf{Z}$ est égal à

$$(\mathbf{Z}/n\mathbf{Z})^* = \{\bar{m} \in \mathbf{Z}/n\mathbf{Z} \mid m \text{ est premier avec } n\} = \{ \text{ générateurs de } \mathbf{Z}/n\mathbf{Z}\}.$$

Définition. On note $\phi(n) := \mathrm{card}(\mathbf{Z}/n\mathbf{Z})^*$ l'*indicatrice d'Euler*.

On en déduit facilement que, si p est premier, $\phi(p^r) = p^r - p^{r-1} = (p-1)p^{r-1}$. Le calcul en général de $\phi(n)$ se fait grâce au lemme classique suivant.

Proposition. (Lemme chinois) *Soit $m, n \in \mathbf{Z}$, supposons m et n premiers entre eux, alors les groupes $\mathbf{Z}/mn\mathbf{Z}$ et $\mathbf{Z}/m\mathbf{Z}\times\mathbf{Z}/n\mathbf{Z}$ sont naturellement isomorphes. De plus cet isomorphisme est aussi un isomorphisme d'anneaux et, par conséquent induit un isomorphisme entre $(\mathbf{Z}/mn\mathbf{Z})^*$ et $(\mathbf{Z}/m\mathbf{Z})^*\times(\mathbf{Z}/n\mathbf{Z})^*$. En particulier $\phi(mn) = \phi(m)\phi(n)$.*

Preuve. Considérons l'application $f : \mathbf{Z} \to \mathbf{Z}/m\mathbf{Z} \times \mathbf{Z}/n\mathbf{Z}$ donnée par $x \mapsto (x \bmod m, x \bmod n)$. C'est un homomorphisme de groupe de noyau $\operatorname{ppcm}(m,n)\mathbf{Z}$, d'où une injection

$$\hat{f} : \mathbf{Z}/\operatorname{ppcm}(m,n)\mathbf{Z} \hookrightarrow \mathbf{Z}/m\mathbf{Z} \times \mathbf{Z}/n\mathbf{Z}.$$

Comme m et n sont supposés premiers entre eux, on a $\operatorname{ppcm}(m,n) = mn$ et, pour des raisons de cardinalité, l'homomorphisme $\hat{f}$ doit être un isomorphisme. De manière générale, si A et B sont des anneaux, on a $(A \times B)^* = A^* \times B^*$ d'où la deuxième assertion. □

Rappelons que, d'après le théorème de Lagrange l'ordre d'un sous-groupe divise toujours l'ordre du groupe. La description des sous-groupes de $\mathbf{Z}/n\mathbf{Z}$ est assez simple.

Proposition. *Pour chaque entier $d \geq 1$ divisant n, il existe un unique sous-groupe de $\mathbf{Z}/n\mathbf{Z}$ d'ordre d, c'est le sous-groupe cyclique engendré par la classe de n/d dans $\mathbf{Z}/n\mathbf{Z}$.*

Preuve. Supposons $n = dd'$ alors l'élément $x = \bar{d}' \in \mathbf{Z}/n\mathbf{Z}$ est d'ordre d car clairement $dx = 0$ et, si $cx = 0$ alors n divise cd' donc d divise c. Soit maintenant H un sous-groupe de $\mathbf{Z}/n\mathbf{Z}$ d'ordre d. Notons $s : \mathbf{Z} \to \mathbf{Z}/n\mathbf{Z}$ la surjection canonique. On sait que $s^{-1}(H) = m\mathbf{Z}$ est engendré par m donc H est engendré par $\bar{m} \in \mathbf{Z}/n\mathbf{Z}$. On a $d\bar{m} = 0$ donc n divise dm donc d' divise m donc le sous-groupe H est contenu dans le sous-groupe engendré par $\bar{d}'$ et donc égal à ce sous-groupe. □

Comme application, on peut en tirer la formule (que nous utiliserons plus bas)

$$n = \sum_{d \mid n} \phi(d).$$

En effet on écrit $\mathbf{Z}/n\mathbf{Z}$ comme union (disjointe) des ensembles d'éléments d'ordre d pour d divisant n. Le nombre de ces éléments est le nombre de générateurs de l'unique sous-groupe de cardinal d, et comme ce dernier est isomorphe à $\mathbf{Z}/d\mathbf{Z}$, le nombre de générateurs est $\phi(d)$.

Un corps fini k est nécessairement de caractéristique finie égale à p un nombre premier et contient donc $\mathbf{Z}/p\mathbf{Z} = \mathbf{F}_p$ (l'homomorphisme $\mathbf{Z} \to k$ a pour noyau $n\mathbf{Z}$ avec $n > 0$ et comme $\mathbf{Z}/n\mathbf{Z} \hookrightarrow k$ on doit avoir n premier). La dimension de k sur $\mathbf{F}_p$, comme $\mathbf{F}_p$-espace vectoriel est finie, égale disons à f et donc $\operatorname{card}(k) = p^f$. On observe que $\operatorname{card}(k^*) = p^f - 1$ donc tous les éléments de k^* vérifient $x^{p^f-1} = 1$ et donc tous les éléments de k vérifient $x^{p^f} = x$. Inversement on peut en déduire une construction d'un corps à p^f éléments ainsi : on considère une extension K de $\mathbf{F}_p = \mathbf{Z}/p\mathbf{Z}$ dans laquelle le polynôme $P = X^{p^f} - X$ est scindé et on pose $k := \{x \in K \mid P(x) = 0\}$. Comme $P'(X) = -1$, les racines de P sont simples et $\operatorname{card}(k) = \deg(P) = p^f$; de plus k est un sous-corps de K car, en caractéristique p, l'application "Frobenius" définie par $\phi : x \mapsto x^p$ est un homomorphisme de corps de même que ϕ^f. C'est-à-dire que l'on a :

$$(xy)^p = x^p y^p \qquad \text{et} \qquad (x+y)^p = x^p + y^p.$$

D'après les théorèmes généraux de théorie des corps le corps k de cardinal p^f est donc unique à isomorphisme près, on le note $\mathbf{F}_{p^f}$. Résumons cela dans un énoncé.

Théorème. *Soit p premier et $f \geq 1$, notons $q = p^f$. Il existe un corps fini de cardinal q, unique à isomorphisme près. Les éléments de $\mathbf{F}_q$ sont les racines du polynôme $X^q - X \in \mathbf{Z}/p\mathbf{Z}[X]$.*

Corollaire. *Soit $q = p^f$ et $\mathbf{F}_q$ comme ci-dessus. Les sous-corps de $\mathbf{F}_q$ sont isomorphes à un $\mathbf{F}_{p^d}$ avec d divisant f. Inversement si d divise f il existe un unique sous-corps de $\mathbf{F}_q$ isomorphe à $\mathbf{F}_{p^d}$: c'est l'ensemble des éléments vérifiant $x^{p^d} = x$.*

Preuve. Si l'on a $\mathbf{F}_p \subset k \subset \mathbf{F}_q$ alors on a $\mathrm{card}(k) = p^d$ avec $d = [k : \mathbf{F}_p]$ et $k \cong \mathbf{F}_{p^d}$; de plus $f = [\mathbf{F}_q : \mathbf{F}_p] = [\mathbf{F}_q : k][k : \mathbf{F}_p]$ donc d divise f. Inversement si d divise f (disons $f = ed$), tout élément (dans une extension de $\mathbf{F}_p$) vérifiant $x^{p^d} = x$ vérifie $x^{p^f} = x^{p^{ed}} = x$ donc est dans $\mathbf{F}_q$ et ces éléments forment un sous-corps isomorphe à $\mathbf{F}_{p^d}$. □

En pratique on construit les corps $\mathbf{F}_{p^f}$ ainsi : on choisit un polynôme unitaire, de degré f, irréductible $P \in \mathbf{F}_p[X]$ (pourquoi existe-t-il?) et on décrit $\mathbf{F}_{p^f}$ comme $\mathbf{F}_p[X]/P\mathbf{F}_p[X]$; un élément de $\mathbf{F}_{p^f}$ peut être vu comme un polynôme de degré $\leq f-1$ à coefficients dans $\mathbf{Z}/p\mathbf{Z}$. L'addition est simple à effectuer, la multiplication est simplement la multiplication de polynômes suivie par l'opération consistant à prendre le reste dans la division euclidienne par P. Par exemple :

$$\mathbf{F}_4 = \mathbf{F}_2[X]/(X^2+X+1)\mathbf{F}_2[X],\ \mathbf{F}_8 = \mathbf{F}_2[X]/(X^3+X+1)\mathbf{F}_2[X],\ \mathbf{F}_{16} = \mathbf{F}_2[X]/(X^4+X^3+X^2+X+1)\mathbf{F}_2[X].$$

B. La structure des groupes $(\mathbf{Z}/n\mathbf{Z})^*$ et $\mathbf{F}_q^*$.

Commençons par montrer le résultat suivant.

Lemme. *Soit k un corps commutatif et G un sous-groupe fini de k^*, alors G est cyclique. En particulier $(\mathbf{Z}/p\mathbf{Z})^*$ ou plus généralement $\mathbf{F}_q^*$ est cyclique.*

Preuve. Notons $n := \mathrm{card}(G)$ et $\psi(d)$ le nombre d'éléments d'ordre d dans G. On a clairement $n = \sum_{d \mid n} \psi(d)$. Soit d divisant n, ou bien il n'y a pas d'élément d'ordre d dans G auquel cas $\psi(d) = 0$, ou bien il en existe un qui engendre alors un sous-groupe cyclique H d'ordre d. Tous les éléments de H sont solutions de l'équation $X^d = 1$, mais, comme k est un corps commutatif, une telle équation possède au plus d racines dans k; tous les éléments d'ordre d sont donc dans H et il en a $\phi(d)$ puisque $H \cong \mathbf{Z}/d\mathbf{Z}$. Ainsi $\psi(d)$ vaut zéro ou $\phi(d)$, mais comme $n = \sum_{d \mid n} \psi(d) = \sum_{d \mid n} \phi(d)$, on voit que $\psi(d) = \phi(d)$ pour tout d divisant n. En particulier $\psi(n) = \phi(n) \geq 1$, ce qui implique bien que G est cyclique. □

D'après ce que nous avons vu, si $n = p_1^{\alpha_1} \dots p_s^{\alpha_s}$ alors

$$(\mathbf{Z}/n\mathbf{Z})^* \cong (\mathbf{Z}/p_1^{\alpha_1}\mathbf{Z})^* \times \dots \times (\mathbf{Z}/p_s^{\alpha_s}\mathbf{Z})^*$$

et en particulier

$$\phi(n) = \phi(p_1^{\alpha_1}) \dots \phi(p_s^{\alpha_s}) = \prod_{i=1}^{s} \left(p_i^{\alpha_i} - p_i^{\alpha_i - 1}\right) = n \prod_{i=1}^{s} \left(1 - \frac{1}{p_i}\right)$$

Il reste à décrire la structure des groupes $(\mathbf{Z}/p^\alpha\mathbf{Z})^*$.

Proposition. *Soit p premier et $\alpha \geq 1$ alors*
(i) Si p est impair $(\mathbf{Z}/p^\alpha\mathbf{Z})^*$ est cyclique.
(ii) Si $p = 2$ et $\alpha \geq 3$ alors $(\mathbf{Z}/2^\alpha\mathbf{Z})^* \cong \mathbf{Z}/2^{\alpha-2}\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$ n'est pas cylique. Par contre $(\mathbf{Z}/2\mathbf{Z})^* = \{1\}$ et $(\mathbf{Z}/4\mathbf{Z})^* \cong \mathbf{Z}/2\mathbf{Z}$ sont cycliques.

Preuve. Si $\alpha = 1$ on a vu que $(\mathbf{Z}/p\mathbf{Z})^* = \mathbf{F}_p^*$ était cyclique. Lorsque $\alpha > 1$ nous allons utiliser l'élément $p+1$.

Lemme. *Soit p premier impair, la classe de $p+1$ dans $(\mathbf{Z}/p^\alpha\mathbf{Z})^*$ est d'ordre $p^{\alpha-1}$.*

Preuve du lemme. Montrons d'abord par récurrence la congruence

$$(p+1)^{p^k} \equiv 1 + p^{k+1} \bmod p^{k+2}.$$

Pour $k=0$, la congruence est triviale. Pour $k=1$, on a $(p+1)^p \equiv 1+C_p^1 p+C_p^2 p^2 \equiv 1+p^2+p^3(p-1)/2 \bmod p^3$ et ce dernier est bien sûr congru à $1+p^2$ si p est impair (remarquer cependant que $3^2 \not\equiv 1+2^2 \bmod 2^3$). Supposons donc $k \geq 1$ et $(p+1)^{p^{k-1}} = 1+p^k+ap^{k+1}$ alors $(p+1)^{p^k} = \left(1+p^k+ap^{k+1}\right)^p \equiv 1+p(p^k+ap^{k+1}) \equiv 1+p^{k+1} \bmod p^{k+2}$ puisque $1+2k \geq k+2$. En particulier, on voit que $(p+1)^{p^{\alpha-1}} \equiv 1 \bmod p^\alpha$ mais $(p+1)^{p^{\alpha-2}} \equiv 1+p^{\alpha-1} \not\equiv 1 \bmod p^\alpha$, ce qui implique bien que $p+1$ est d'ordre $p^{\alpha-1}$ dans $(\mathbf{Z}/p^\alpha\mathbf{Z})^*$.

On peut maintenant terminer la preuve de la proposition pour p impair. Soit $x \in \mathbf{Z}$ tel que x modulo p engendre $(\mathbf{Z}/p\mathbf{Z})^*$ i.e. est d'ordre $p-1$ dans $(\mathbf{Z}/p\mathbf{Z})^*$; alors $\bar{x}$ est d'ordre $m(p-1)$ dans $(\mathbf{Z}/p^\alpha\mathbf{Z})^*$ et donc $y=\bar{x}^m$ est d'ordre exactement $p-1$ dans $(\mathbf{Z}/p^\alpha\mathbf{Z})^*$. L'élément $y(p+1)$ est donc d'ordre $p^{\alpha-1}(p-1)$ donc est un générateur de $(\mathbf{Z}/p^\alpha\mathbf{Z})^*$ (car $p^{\alpha-1}$ et $p-1$ sont premiers entre eux).

Lemme. *Soit $\alpha \geq 3$, la classe de 5 dans $(\mathbf{Z}/2^\alpha\mathbf{Z})^*$ est d'ordre $2^{\alpha-2}$. De plus la classe de -1 n'appartient pas au sous-groupe engendré par la classe de 5.*

Preuve du lemme. On montre d'abord par récurrence que

$$5^{2^k} \equiv 1+2^{k+2} \bmod 2^{k+3}.$$

La congruence est triviale pour $k=0$, pour $k=1$ on vérifie $25=5^2 \equiv 1+2^3=9 \bmod 2^4$. Supposons donc que $5^{2^{k-1}} = 1+2^{k+1}+a2^{k+2}$ alors $5^{2^k} = (1+2^{k+1}+a2^{k+2})^2 = 1+2(2^{k+1}+a2^{k+2})+2^{2(k+1)}(1+2a)^2 \equiv 1+2^{k+2} \bmod 2^{k+3}$. En particulier $5^{2^{\alpha-2}} \equiv 1 \bmod 2^\alpha$ mais $5^{2^{\alpha-3}} \equiv 1+2^{\alpha-1} \not\equiv 1 \bmod 2^\alpha$ donc 5 est bien d'ordre $2^{\alpha-2}$. Supposons que $5^\beta \equiv -1 \bmod 2^\alpha$ alors $5^{2\beta} \equiv 1 \bmod 2^\alpha$ donc $2^{\alpha-2}$ divise 2β donc $2^{\alpha-3}$ divise β ou encore $\beta = \gamma 2^{\alpha-3}$. Comme 5 est d'ordre $2^{\alpha-2}$, on peut considérer β comme un entier modulo $2^{\alpha-2}$ et donc γ modulo 2. L'entier γ doit être impair donc on peut le supposer égal à 1, c'est-à-dire $5^{2^{\alpha-3}} \equiv 1 \bmod 2^\alpha$, mais $5^{2^{\alpha-3}} \equiv 1+2^{\alpha-1} \bmod 2^\alpha$ donc $-1 \equiv 1+2^{\alpha-1} \bmod 2^\alpha$ ou encore $2+2^{\alpha-1} \equiv 0 \bmod 2^\alpha$ soit $1+2^{\alpha-2} \equiv 0 \bmod 2^{\alpha-1}$, ce qui n'est pas possible. □

Pour la démonstration de la deuxième partie de la proposition, on peut supposer $\alpha \geq 3$ (en effet le calcul de $(\mathbf{Z}/2\mathbf{Z})^*$ et $(\mathbf{Z}/4\mathbf{Z})^*$ est immédiat). La classe de 5 engendre donc un sous-groupe isomorphe à $\mathbf{Z}/2^{\alpha-2}\mathbf{Z}$ et -1 engendre un sous-groupe d'ordre 2 non contenu dans le précédent donc $(\mathbf{Z}/2^\alpha\mathbf{Z})^* = \langle 5\rangle \oplus \langle -1\rangle \cong \mathbf{Z}/2^{\alpha-2}\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$. □

Exercice. Montrer que si la classe de $x \in \mathbf{Z}$ engendre $(\mathbf{Z}/p^2\mathbf{Z})^*$ alors elle engendre aussi $(\mathbf{Z}/p^\alpha\mathbf{Z})^*$ (pour p impair).

Remarque. Le sous-groupe quaternionique $H_8 = \{\pm 1, \pm i, \pm j, \pm k\}$ est un sous-groupe fini du groupe multiplicatif du corps des quaternions $\mathbf{H}$ mais n'est pas cyclique (cela ne contredit pas le lemme vu car $\mathbf{H}$ n'est pas commutatif).

Applications. On peut déduire des énoncés précédents le nombre de solutions de l'équation $x^m=1$ dans $\mathbf{F}_q^*$ ou $(\mathbf{Z}/N\mathbf{Z})^*$, ainsi que le nombre de puissance m-ièmes. En effet, dans un groupe cyclique de cardinal n, disons $G=\mathbf{Z}/n\mathbf{Z}$ le nombre d'éléments vérifiant $mx=0$ est égal à $d := \operatorname{pgcd}(m,n)$: en observant que si m et n sont premiers entre eux alors la multiplication par n est un isomorphisme, on voit que $\{x \in \mathbf{Z}/n\mathbf{Z} \mid mx=0\}$ est égal à $\{x \in \mathbf{Z}/n\mathbf{Z} \mid dx=0\}$ et, puisque d divise n, ce dernier ensemble est le sous-groupe cyclique de cardinal d dans $\mathbf{Z}/n\mathbf{Z}$. Résumons cela dans le lemme suivant :

Lemme. *Soit G un groupe cyclique de cardinal n et f l'homomorphisme de G dans G défini par $x \to x^m$. Le noyau de f est cyclique de cardinal $\operatorname{pgcd}(m,n)$ et l'image de f (l'ensemble des puissances m-ièmes est cyclique de cardinal $n/\operatorname{pgcd}(m,n)$.*

En appliquant cela à $G=\mathbf{F}_q^*$ ou $G=(\mathbf{Z}/p^\alpha\mathbf{Z})^*$ on obtient la première partie de la proposition suivante

Proposition. *Soit m un entier ≥ 1 alors*
(1) On a les formules suivantes (pour p impair)

$$\operatorname{card}\{x \in \mathbf{F}_q^* \mid x^m=1\} = \operatorname{pgcd}(m,q-1) \quad \textit{et} \quad \operatorname{card}\{x \in (\mathbf{Z}/p^\alpha\mathbf{Z})^* \mid x^m=1\} = \operatorname{pgcd}(m,(p-1)p^{\alpha-1}).$$

(2) Plus généralement si $N = p_1^{\alpha_1} \dots p_r^{\alpha_r}$ est impair :

$$\operatorname{card}\{x \in (\mathbf{Z}/N\mathbf{Z})^* \mid x^m = 1\} = \prod_{i=1}^{r} \operatorname{pgcd}(m, (p_i - 1)p_i^{\alpha_i - 1}).$$

Preuve. Les formules de (1) résultent des considérations précédentes et du fait que $\mathbf{F}_q^*$ et $(\mathbf{Z}/p^\alpha\mathbf{Z})^*$ sont cycliques. La formule (2) se déduit de la précédente et du lemme chinois. En effet si $x \in \mathbf{Z}$ alors $x^m \equiv 1 \bmod N$ équivaut à $x^m \equiv 1 \bmod p_i^{\alpha_i}$ pour $1 \leq i \leq r$. □

Remarque. En considérant l'homomorphisme $x \mapsto x^m$ on en tire facilement, par exemple, que :

$$\operatorname{card} \mathbf{F}_q^{*m} = \operatorname{card}\{x \in \mathbf{F}_q^* \mid \exists y \in \mathbf{F}_q^*,\ x = y^m\} = \frac{q-1}{\operatorname{pgcd}(m, q-1)}$$

Par exemple, si q est impair on a $(\mathbf{F}_q^* : \mathbf{F}_q^{*2}) = 2$.

Exercice. Montrer que si N est pair et m impair, la dernière formule de la proposition reste correcte. Comment faut-il la modifier lorsque N et m sont pairs?

C. Symboles de Legendre et Jacobi.

On se propose ici d'étudier particulièrement les carrés, i.e. le cas $m = 2$ du paragraphe précédent.

Commençons par une remarque. L'application $x \mapsto x^2$ est un isomorphisme de $\mathbf{F}_2$ sur $\mathbf{F}_2$ ou plus généralement de $\mathbf{F}_{2^f}$ sur $\mathbf{F}_{2^f}$; il est donc naturel, pour étudier les carrés de se placer dans l'hypothèse $p \neq 2$ et c'est ce que nous faisons.

Définition. On définit le *symbole de Legendre* ainsi pour $a \in \mathbf{Z}$ (et $p \neq 2$) :

$$\left(\frac{a}{p}\right) := \begin{cases} 0 & \text{si } a \equiv 0 \bmod p \\ +1 & \text{si } a \text{ est un carré non nul mod } p \\ -1 & \text{si } a \text{ n'est pas un carré mod } p \end{cases}$$

Remarque. Il est clair que $\left(\frac{a}{p}\right)$ ne dépend que de $a \bmod p$; on s'autorisera donc à continuer à utiliser la même notation lorsque $a \in \mathbf{F}_p$. Si $\left(\frac{a}{p}\right) = +1$, on dira que a est un *résidu quadratique* ; si $\left(\frac{a}{p}\right) = -1$, on dira que a est un *non-résidu quadratique*.

Théorème. *Le symbole de Legendre vérifie les propriétés suivantes*

(i) Pour $a, b \in \mathbf{Z}$ on a

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

(ii) Pour tout $a \in \mathbf{Z}$ on a :

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

(iii) Pour tout $p \neq 2$ on a :

$$\left(\frac{-1}{p}\right) = (-1)^{(p-1)/2} \qquad \text{et} \qquad \left(\frac{2}{p}\right) = (-1)^{(p^2-1)/8}.$$

En particulier -1 est un carré modulo p (resp. n'est pas un carré) si $p \equiv 1 \bmod 4$ (resp. $p \equiv 3 \bmod 4$) et 2 est un carré modulo p (resp. n'est pas un carré) si $p \equiv \pm 1 \bmod 8$ (resp. $p \equiv \pm 3 \bmod 8$).

(iv) (Loi de réciprocité quadratique)

$$\left(\frac{q}{p}\right)\left(\frac{p}{q}\right) = (-1)^{\frac{(p-1)(q-1)}{4}}.$$

Preuve. La multiplicativité (i) est claire si p divise a ou b car alors les deux termes sont nuls. Si $a, b \in \mathbf{F}_p^*$ alors la formule vient de ce que $\mathbf{F}_p^*/\mathbf{F}_p^{*2}$ est de cardinal 2, donc le produit de deux non-résidus quadratiques est un résidu quadratique. Pour prouver (ii) observons que, si p divise a la formule est évidente, que si $a = b^2 \in \mathbf{F}_p^{*2}$ alors $a^{(p-1)/2} = b^{p-1} = 1$ qui est bien égal à $\left(\frac{a}{p}\right)$; si $\left(\frac{a}{p}\right) = -1$ considérons g un générateur de $\mathbf{F}_p^*$, alors $\left(\frac{g}{p}\right) = -1$ et $a = g^m$ avec m impair (sinon a serait un carré) donc $\left(\frac{a}{p}\right) = \left(\frac{g}{p}\right)^m = -1$. La première partie de (iii) découle de l'égalité (ii) ; pour la deuxième partie on introduit α racine 8-ième primitive de l'unité dans une extension de $\mathbf{F}_p$, c'est-à-dire que $\alpha^8 = 1$ mais $\alpha^4 \neq 1$, ce qui équivaut à $\alpha^4 = -1$ ou encore à $\alpha^2 = -\alpha^{-2}$. Posons $\beta := \alpha + \alpha^{-1}$ alors $\beta^2 = \alpha^2 + 2 + \alpha^{-2} = 2$; ainsi on voit que 2 est un carré dans $\mathbf{F}_p$ si et seulement si $\beta \in \mathbf{F}_p$. Or on sait que $\beta \in \mathbf{F}_p$ équivaut à $\beta^p = \beta$; calculons donc $\beta^p = \alpha^p + \alpha^{-p}$. En se souvenant que $\alpha^8 = 1$ et $\alpha^4 = -1$ on voit que, si $p \equiv \pm 1 \bmod 8$ on a $\beta^p = \beta$ et donc $\beta \in \mathbf{F}_p$ alors que si $p \equiv \pm 3 \bmod 8$ on a $\beta^p = -\beta$ et donc $\beta \notin \mathbf{F}_p$. Nous renvoyons la démonstration de la loi de réciprocité quadratique (iv) au paragraphe suivant. □

Remarque. Pour voir d'où vient le choix de "$\beta = \sqrt{2}$" on peut remarquer que si $\zeta := \exp(2\pi i/8) \in \mathbf{C}$ alors ζ est une racine huitième primitive de l'unité et $\zeta = \frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}$ et $\zeta + \zeta^{-1} = \zeta + \bar{\zeta} = \sqrt{2}$.

Introduisons maintenant une généralisation pour $N = p_1^{\alpha_1} \dots p_r^{\alpha_r}$ impair, le *symbole de Jacobi*

$$\left(\frac{a}{N}\right) := \left(\frac{a}{p_1}\right)^{\alpha_1} \dots \left(\frac{a}{p_r}\right)^{\alpha_r}.$$

On a clairement $\left(\frac{ab}{N}\right) = \left(\frac{a}{N}\right)\left(\frac{b}{N}\right)$. Les principales autres propriétés sont

Lemme. *Pour N, M impairs, on a:*
(i) $\left(\frac{-1}{N}\right) = (-1)^{\frac{N-1}{2}}$ *et* $\left(\frac{2}{N}\right) = (-1)^{\frac{N^2-1}{8}}$
(ii) $\left(\frac{M}{N}\right) = (-1)^{\frac{(N-1)(M-1)}{4}}\left(\frac{N}{M}\right)$.

Preuve. Ces formules se déduisent des formules pour M et N premiers. En effet écrivons $N = p_1 \dots p_r$ (avec répétition éventuelle) alors

$$\left(\frac{-1}{N}\right) = \prod_{i=1}^{r}\left(\frac{-1}{p_i}\right) = (-1)^{\sum_{i=1}^{r}(p_i-1)/2} = (-1)^h$$

avec h égal au nombre d'indices i avec $p_i \equiv 3 \bmod 4$. par ailleurs, on a $N \equiv 3^h \bmod 4$ donc $N \equiv 3 \bmod 4$ si h impair et $N \equiv 1 \bmod 4$ si h pair. On vérifie bien que $\sum_{i=1}^{r}(p_i - 1)/2$ est impair si h est impair et pair si h est pair. De même on a

$$\left(\frac{2}{N}\right) = \prod_{i=1}^{r}\left(\frac{2}{p_i}\right) = (-1)^{\sum_{i=1}^{r}(p_i^2-1)/8} = (-1)^h$$

où h désigne maintenant le nombre d'indices i avec $p_i \equiv \pm 3 \bmod 8$. Dans ce cas, on a $N \equiv \pm 3^h \bmod 4$. On vérifie bien que $\sum_{i=1}^{r}(p_i^2 - 1)/8$ est impair si h est impair et pair si h est pair d'où la deuxième formule de (i).

Pour prouver (ii), écrivons $M = q_1 \dots q_s$ et $N = p_1 \dots p_r$ (avec répétition éventuelle). Si h (resp. k) est le nombre d'indices i avec $p_i \equiv 3 \bmod 4$ (resp. $q_j \equiv 3 \bmod 4$) on a $\frac{N-1}{2}$ impair si h impair et $\frac{N-1}{2}$ pair si h pair (resp. $\frac{M-1}{2}$ impair si k impair et $\frac{M-1}{2}$ pair si k pair). On a alors

$$\left(\frac{M}{N}\right) = \prod_{i=1}^{r}\prod_{j=1}^{s}\left(\frac{q_j}{p_i}\right) = \prod_{i=1}^{r}\prod_{j=1}^{s}(-1)^{(p_i-1)(q_j-1)/4}\left(\frac{p_i}{q_j}\right) = (-1)^{hk}\left(\frac{N}{M}\right) = (-1)^{\frac{(N-1)(M-1)}{4}}\left(\frac{N}{M}\right).$$

□

La propriété (*ii*) est la *loi de réciprocité quadratique* de Jacobi; les deux propriétés fournissent un algorithme pour calculer le symbole de Jacobi. Attention cependant le symbole de Jacobi ne caractérise pas les carrés modulo N (si a est premier avec N et est un carré modulo N alors $\left(\frac{a}{N}\right) = 1$ mais la réciproque est fausse si N est composé).

Comme première application de la loi de réciprocité quadratique, montrons que pour d entier sans facteur carré, les nombres premiers représentables sous la forme $p = x^2 + dy^2$ doivent appartenir à certaines classes de congruences modulo $4d$.

En effet si $d = \epsilon p_1 \dots p_k$ (avec $\epsilon = \pm 1$) et $p = x^2 + dy^2$ alors p ne divise pas y car sinon p diviserait aussi x et on en déduirait p^2 divise p. On obtient donc $-d = (xy^{-1})^2 \bmod p$ si d est impair on obtient

$$1 = \left(\frac{-d}{p}\right) = (-\epsilon)^{\frac{p-1}{2}} (-1)^{\sum_{i=1}^{k} (p_i-1)(p-1)/4} \left(\frac{p}{p_1}\right) \dots \left(\frac{p}{p_k}\right)$$

On en déduit bien des congruences modulo $4p_1 \dots p_k$ pour p. Si d est pair, c'est-à-dire disons $p_1 = 2$ on calcule à part $\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}$ et on obtient des congruences modulo $8p_2 \dots p_k$ pour p.

Exemple. Si un nombre premier s'écrit $p = x^2 - 6y^2$, avec $x, y \in \mathbf{Z}$, on en tire que $\left(\frac{6}{p}\right) = 1$ ou encore $1 = (-1)^{(p^2-1)/8}(-1)^{(p-1)/2}\left(\frac{p}{3}\right)$ ce qui équivaut à $p \equiv 1$ ou $3 \bmod 8$ et $p \equiv 1 \bmod 3$, ou bien à $p \equiv -1$ ou $-3 \bmod 8$ et $p \equiv -1 \bmod 3$. Par le lemme chinois, on conclut donc que $p \equiv 1, 5, 19$ ou $23 \bmod 24$. Ainsi aucun nombre premier $p \equiv 7, 11, 13$ ou $17 \bmod 24$ ne peut être de la forme $p = x^2 - 6y^2$.

D. Sommes de Gauss.

Les somme de Gauss sont importantes en arithmétique ; nous allons les utiliser pour donner une démonstration – due à Gauss bien sûr – de la loi de réciprocité quadratique. Au paragraphe suivant nous les utilisons pour calculer le nombre de solutions modulo p d'une équation quadratique.

Remarquons que $\exp\left(\frac{2\pi i a}{p}\right)$ ne dépend que de $a \bmod p$ et garde un sens pour $a \in \mathbf{F}_p$. Nous utiliserons les formules dont la preuve est laissée en exercice (instructif) :

$$\sum_{x \in \mathbf{F}_p} \left(\frac{x}{p}\right) = \sum_{x \in \mathbf{F}_p^*} \left(\frac{x}{p}\right) = 0 \qquad \text{et} \qquad \sum_{x=0}^{n-1} \exp\left(\frac{2\pi i x y}{n}\right) = \begin{cases} n & \text{si } n \text{ divise } y \\ 0 & \text{si } n \text{ ne divise pas } y \end{cases}$$

On utilisera aussi de manière répétée la formule de changement de variables évidente : si $f : X \to Y$ est une bijection, $\sum_{y \in Y} \phi(y) = \sum_{x \in X} \phi \circ f(x)$.

Le premier exemple que nous traiterons est le suivant (pour p premier impair et a premier avec p) :

$$\tau(a) := \sum_{x=0}^{p-1} \exp\left(\frac{2\pi i a x^2}{p}\right).$$

Proposition. *Les sommes $\tau(a)$ vérifient les formules suivantes.*
(i) $\tau(a) = \left(\frac{a}{p}\right)\tau(1)$.
(ii) $|\tau(a)|^2 = p$.
(iii) $\tau(1)^2 = \left(\frac{-1}{p}\right)p$.

Preuve. Soit a un résidu quadratique et b un non-résidu quadratique modulo p.

$$\tau(a) + \tau(b) = \sum_{x=0}^{p-1} \exp\left(\frac{2\pi i a x^2}{p}\right) + \sum_{x=0}^{p-1} \exp\left(\frac{2\pi i b x^2}{p}\right) = 2 + 2 \sum_{u \in a\mathbf{F}_p^{*2}} \exp\left(\frac{2\pi i u}{p}\right) + 2 \sum_{u \in b\mathbf{F}_p^{*2}} \exp\left(\frac{2\pi i u}{p}\right) = 0$$

Ainsi on a bien $\tau(b) = -\tau(1)$ et $\tau(a) = \tau(1)$, ce qui démontre (i). Pour la deuxième formule, on calcule de deux façons $\sum_{a=1}^{p-1} |\tau(a)|^2 = (p-1)|\tau(1)|^2$ qui est aussi

$$\sum_{a=1}^{p-1} \sum_{x,y \in \mathbf{F}_p} \exp\left(\frac{2\pi i a(x^2-y^2)}{p}\right) = \sum_{a=1}^{p-1} \sum_{u,v \in \mathbf{F}_p} \exp\left(\frac{2\pi i a u v}{p}\right) = \sum_{a=1}^{p-1} p = p(p-1)$$

d'où la formule (ii) découle. Enfin on a $\overline{\tau(1)} = \tau(-1) = \left(\frac{-1}{p}\right)\tau(1)$ donc $\tau(1)^2 = \left(\frac{-1}{p}\right)|\tau(1)|^2 = \left(\frac{-1}{p}\right)p$. □

Remarque. La formule (iii) permet de dire que, si $p \equiv 1 \bmod 4$ alors $\tau(1) = \pm\sqrt{p}$, alors que si $p \equiv 3 \bmod 4$ alors $\tau(1) = \pm i\sqrt{p}$. On peut en fait montrer (c'est un peu délicat à prouver) que le signe est toujours $+$. Par exemple

$$\tau_3(1) = \sum_{x=0}^{2} \exp\left(\frac{2\pi i x^2}{3}\right) = 1 + 2\exp\left(\frac{2\pi i}{3}\right) = 1 + 2\left(-\frac{1}{2} + i\frac{\sqrt{3}}{2}\right) = i\sqrt{3}$$

$$\tau_5(1) = \sum_{x=0}^{4} \exp\left(\frac{2\pi i x^2}{5}\right) = 1 + 2\exp\left(\frac{2\pi i}{5}\right) + 2\exp\left(\frac{-2\pi i}{5}\right) = 1 + 4\cos\left(\frac{2\pi}{5}\right) = 1 + 4\left(-\frac{1}{4} + \frac{\sqrt{5}}{4}\right) = \sqrt{5}$$

On peut donner une autre expression des sommes en montrant le lemme suivant.

Lemme. *On a l'égalité :*

$$\tau(a) = \sum_{x \in \mathbf{F}_p} \left(\frac{x}{p}\right) \exp\left(\frac{2\pi i a x}{p}\right) = \sum_{x \in \mathbf{F}_p^*} \left(\frac{x}{p}\right) \exp\left(\frac{2\pi i a x}{p}\right).$$

Preuve. Remarquons que $1 + \left(\frac{x}{p}\right)$ est égal au nombre de solutions dans $\mathbf{F}_p$ de l'équation $y^2 = x$. On en déduit :

$$\sum_{x \in \mathbf{F}_p} \left(\frac{x}{p}\right) \exp\left(\frac{2\pi i a x}{p}\right) = \sum_{x \in \mathbf{F}_p} \left(1 + \left(\frac{x}{p}\right)\right) \exp\left(\frac{2\pi i a x}{p}\right) = \sum_{x \in \mathbf{F}_p} \exp\left(\frac{2\pi i a x^2}{p}\right) = \tau(a)$$

comme annoncé. □

Ceci suggère une première généralisation. Définissons un *caractère* comme un homomorphisme $\chi : \mathbf{F}_p^* \to \mathbf{C}^*$ que l'on prolonge par convention à $\mathbf{F}_p$ tout entier par $\chi(0) := 0$. On pose alors

$$G(\chi, a) = \sum_{x \in \mathbf{F}_p} \chi(x) \exp\left(\frac{2\pi i a x}{p}\right) = \sum_{x \in \mathbf{F}_p^*} \chi(x) \exp\left(\frac{2\pi i a x}{p}\right)$$

et on démontre de même :

Proposition. *Les sommes $G(\chi, a)$ vérifient les formules suivantes.*
(i) $G(\chi, a) = \bar{\chi}(a) G(\chi, 1)$.
(ii) $|G(\chi, a)|^2 = p$
(iii) $G(\chi, 1) = \chi(-1) G(\bar{\chi}, 1)$.

Preuve. Pour la première formule, notons que $\chi(a^{-1}) = \chi(a)^{-1} = \bar{\chi}(a)$. Ainsi :

$$G(\chi, a) = \sum_{x \in \mathbf{F}_p^*} \chi(x) \exp\left(\frac{2\pi i a x}{p}\right) = \chi(a^{-1}) \sum_{x \in \mathbf{F}_p^*} \chi(ax) \exp\left(\frac{2\pi i a x}{p}\right) = \chi(a^{-1}) G(\chi, 1).$$

Pour la deuxième formule $\sum_{a=1}^{p-1}|G(\chi,a)|^2=(p-1)|G(\chi,1)|^2$ et vaut aussi

$$\sum_{a=1}^{p-1}\sum_{x,y\in\mathbf{F}_p}\chi(x)\bar\chi(y)\exp\left(\frac{2\pi ia(x-y)}{p}\right)=\sum_{a=0}^{p-1}\sum_{x,y\in\mathbf{F}_p}\chi(x)\bar\chi(y)\exp\left(\frac{2\pi ia(x-y)}{p}\right)-\sum_{x,y\in\mathbf{F}_p}\chi(x)\bar\chi(y)$$

$$=p\sum_{x\in\mathbf{F}_p}\chi(x)\bar\chi(x)=p(p-1).$$

La dernière formule s'en déduit ainsi ;

$$\overline{G(\chi,1)}=G(\bar\chi,-1)=\chi(-1)G(\bar\chi,1).$$

□

Exercice. Définissons plus généralement un *caractère* modulo n comme un homomorphisme $\chi:(\mathbf{Z}/n\mathbf{Z})^*\to\mathbf{C}^*$ que l'on prolonge par convention à $\mathbf{Z}/n\mathbf{Z}$ tout entier par $\chi(x):=0$ si x non inversible. On dira que χ est *primitif* s'il ne provient pas d'un caractère modulo m avec m diviseur strict de n. On pose

$$G(\chi,a)=\sum_{x\in\mathbf{Z}/n\mathbf{Z}}\chi(x)\exp\left(\frac{2\pi iax}{n}\right)=\sum_{x\in(\mathbf{Z}/n\mathbf{Z})^*}\chi(x)\exp\left(\frac{2\pi iax}{n}\right)$$

Montrer les formules, pour a premier avec n et χ primitif modulo n.
(i) $G(\chi,a)=\bar\chi(a)G(\chi,1)$.
(ii) $|G(\chi,a)|^2=n$
(iii) $\overline{G(\chi,1)}=\chi(-1)G(\bar\chi,1)$.

Pour prouver la loi de réciprocité quadratique, on va introduire l'analogue de ces sommes en caractéristique finie. Plus précisément, si p,q sont deux nombres premiers impairs distincts on choisit α une racine primitive p-ième de l'unité dans une extension de $\mathbf{F}_q$; explicitement α est une racine de l'équation

$$\alpha^{p-1}+\alpha^{p-2}+\ldots+\alpha+1=0.$$

On définit ensuite la "somme de Gauss" dans $\mathbf{F}_q(\alpha)$ par :

$$\tau:=\sum_{x\in\mathbf{F}_p}\left(\frac{x}{p}\right)\alpha^x$$

et on démontre le lemme suivant

Lemme. *Soit τ l'élément de $\mathbf{F}_q(\alpha)$ introduit ci-dessus.*
(1) $\tau^2=\left(\frac{-1}{p}\right)p$.
(2) $\tau^{q-1}=\left(\frac{q}{p}\right)$.

Preuve. On calcule

$$\tau^2=\sum_{x,y\in\mathbf{F}_p}\left(\frac{xy}{p}\right)\alpha^{x+y}=\sum_{u\in\mathbf{F}_p}S(u)\alpha^u$$

avec $S(u):=\sum_{x+y=u}\left(\frac{xy}{p}\right)=\sum_{x\in\mathbf{F}_p}\left(\frac{x(u-x)}{p}\right)$. Pour $u=0$ on a $S(0)=\sum_{x\in\mathbf{F}_p}\left(\frac{-x^2}{p}\right)=\left(\frac{-1}{p}\right)(p-1)$. Pour $u\in\mathbf{F}_p^*$ la somme $S(u)$ vaut

$$\sum_{x\in\mathbf{F}_p}\left(\frac{x(u-x)}{p}\right)=\sum_{x\in\mathbf{F}_p^*}\left(\frac{-x^2(1-ux^{-1})}{p}\right)=\left(\frac{-1}{p}\right)\sum_{x\in\mathbf{F}_p^*}\left(\frac{1-ux^{-1}}{p}\right)=\left(\frac{-1}{p}\right)\left\{\sum_{y\in\mathbf{F}_p^*}\left(\frac{y}{p}\right)-1\right\},$$

c'est-à-dire $S(u) = -\left(\frac{-1}{p}\right)$. Ainsi :

$$\tau^2 = \left(\frac{-1}{p}\right)(p-1-\sum_{u=1}^{p-1}\alpha^u) = \left(\frac{-1}{p}\right)p.$$

Pour la deuxième formule, on écrit que, puisque la caractéristique est q impair, on a :

$$\tau^q = \sum_{x\in\mathbf{F}_p}\left(\frac{x}{p}\right)^q\alpha^{qx} = \sum_{x\in\mathbf{F}_p}\left(\frac{x}{p}\right)\alpha^{qx} = \left(\frac{q}{p}\right)\sum_{x\in\mathbf{F}_p}\left(\frac{qx}{p}\right)\alpha^{qx} = \left(\frac{q}{p}\right)\tau.$$

En utilisant le fait que $\tau \neq 0$ à cause de (i) on obtient bien (ii). □

Preuve de la loi de réciprocité quadratique. On a vu que, si q ne divise pas $a \in \mathbf{Z}$, on a $a^{(q-1)/2} \equiv \left(\frac{a}{q}\right) \bmod q$. Donc, en appliquant cela à $a = p$ on obtient les égalités suivantes dans $\mathbf{F}_q(\alpha)$, où l'on invoque successivement les formules (1) et (2) du lemme précédent.

$$\left(\frac{p}{q}\right) = p^{(q-1)/2} = \left(\left(\frac{-1}{p}\right)\tau^2\right)^{(q-1)/2} = (-1)^{(p-1)(q-1)/4}\tau^{q-1} = (-1)^{(p-1)(q-1)/4}\left(\frac{q}{p}\right)$$

On en tire l'égalité des signes (dans $\mathbf{Z}$ par exemple) :

$$\left(\frac{p}{q}\right) = (-1)^{(p-1)(q-1)/4}\left(\frac{q}{p}\right)$$

ce qui achève la preuve.

E. Applications au nombre de solutions d'équations.

Nous allons maintenant donner une autre application des sommes de Gauss, et des théorèmes élémentaires sur le nombre de solutions d'équations dans $\mathbf{F}_q$ ou $\mathbf{Z}/N\mathbf{Z}$.

Théorème. (Théorème de Chevalley-Warning) *Soit $k = \mathbf{F}_q$ un corps fini de caractéristique p. Si $P \in k[x_1, \ldots, x_n]$ avec $\deg(P) < n$ alors*

$$\mathrm{card}\{x \in k^n \mid P(x) = 0\} \equiv 0 \bmod p.$$

En particulier, si P est homogène de degré $d < n$ alors P possède un zéro non trivial (i.e. distinct de 0).

Preuve. Commençons par calculer la somme des valeurs d'un monôme.

Lemme. *Soit $x^m := x_1^{m_1}\ldots x_n^{m_n}$ un monôme, alors $\sum_{x\in k^n} x^m$ est nul sauf si chaque m_i est non nul et divisible par $(q-1)$. En particulier cette somme est nulle dès que $m_1 + \ldots + m_n < (n-1)q$.*

Preuve. Remarquons que, comme le polynôme "X^0" est le polynôme constant, il est naturel de prendre ici la convention $0^0 = 1$. Le calcul

$$\sum_{x\in k^n} x^m = \sum_{(x_1,\ldots,x_n)\in k^n} x_1^{m_1}\ldots x_n^{m_n} = \left(\sum_{x_1\in k} x_1^{m_1}\right)\ldots\left(\sum_{x_n\in k} x_n^{m_n}\right)$$

permet de se ramener au cas d'une variable. Si $m = 0$ alors $\sum_{y\in k} y^0 = q\cdot 1_k = 0$. Si m n'est pas divisible par $q-1$, prenons y_0 un générateur de k^*, alors $y_0^m \neq 1$ et donc

$$\sum_{y\in k} y^m = \sum_{y\in k}(y_0 y)^m = y_0^m\sum_{y\in k} y^m$$

entraîne $\sum_{y\in k} y^m = 0$. $\square$

On déduit du lemme que si $Q \in k[x_1, \ldots, x_n]$ avec $\deg(Q) < (q-1)n$, alors $\sum_{x\in k^n} Q(x) = 0$. Soit maintenant P le polynôme de l'énoncé du théorème de Chevalley-Warning, nous allons appliquer le résultat précédent à $Q = 1 - P^{q-1}$. Observons que $\deg(Q) = (q-1)\deg(P) < (q-1)n$ et que $Q(x) = 1$ si $P(x) = 0$ alors que $Q(x) = 0$ si $P(x) \neq 0$ et $x \in k^n$, donc on a l'égalité dans k :

$$0 = \sum_{x\in k^n} Q(x) = \sum_{\substack{x\in k^n \\ P(x)=0}} 1 = \mathrm{card}\{x \in k^n \mid P(x) = 0\}1_k$$

ce qui achève la preuve, car k est de caractéristique p et donc $m1_k = 0$ équivaut à $m \equiv 0 \bmod p$. $\square$

Exercice. Démontrer par une méthode analogue la généralisation suivante. Soient $P_1, \ldots, P_s$ des polynômes de degrés $d_1, \ldots, d_s$ avec $d_1 + \ldots + d_s < n$, montrer que

$$\mathrm{card}\{x \in k^n \mid P_1(x) = \ldots = P_s(x) = 0\} \equiv 0 \bmod p.$$

En particulier, si les polynômes sont homogènes, ils ont un zéro commun non trivial.

Definition. Si $Q(x) = \sum_{1\leq i,j\leq n} a_{ij}x_ix_j$ est une forme quadratique, on dit qu'elle est *non dégénerée* si $D_Q := \det(a_{ij}) \neq 0$.

Commençons par montrer que, si la caractéristique du corps k est différente de deux, on peut remplacer Q par une forme diagonale $Q'(y) = a_1y_1^2 + \ldots + a_ny_n^2$. Ecrivons $Q(x) = {}^txAx$, quitte à remplacer la matrice A par $\frac{1}{2}({}^tA + A)$, on peut supposer A symétrique et, si l'on introduit la forme bilinéaire symétrique $B(x, y) = {}^txAy$ on a $Q(x) = B(x, x)$ et $B(x, y) = \frac{1}{2}(Q(x + y) - Q(x) - Q(y))$. Soit F un sous-espace vectoriel de k^n, on note $F^\perp = \{x \in k^n \mid \forall y \in F,\ B(x, y) = 0\}$, alors on a $\dim F + \dim F^\perp = n$. En effet considérons $e_1, \ldots, e_r$ une base de F et posons $\Phi(x) = (B(e_1, x), \ldots, B(e_r, x))$ de k^n vers k^r ; le noyau de l'application linéaire Φ est $F^\perp$ et son image est k^r tout entier car sinon il existerait $a_1, \ldots, a_r$ non tous nuls avec $0 = a_1B(e_1, x) + \ldots + a_rB(e_r, x) = B(a_1e_1 + \ldots + a_re_r, x)$, ce qui contredirait que B (ou Q) est non dégénérée. On en tire bien $n = \dim \mathrm{Ker}\,\Phi + \dim \mathrm{Im}\,\Phi = \dim F + \dim F^\perp$. Montrons maintenant par récurrence sur n qu'il existe une base orthogonale. Choisissons e_1 tel que $Q(e_1) \neq 0$, alors $k^n = \langle e_1\rangle \oplus \langle e_1\rangle^\perp$ et on peut conclure par induction puisque $\dim\langle e_1\rangle^\perp = n-1$. Si maintenant $e_1, \ldots, e_n$ est une base orthogonale avec $Q(e_i) = a_i$ et si on note $y_1, \ldots, y_n$ les coordonnées du vecteur $(x_1, \ldots, x_n)$ dans la base $e_1, \ldots, e_n$, on a :

$$Q(x_1, \ldots, x_n) = Q(y_1e_1 + \ldots + y_ne_n) = a_1y_1^2 + \ldots + a_ny_n^2.$$

Remarquons que, si l'on appelle Q' la forme quadratique de droite et U la matrice de changement de base on a $D_Q = \det(U)^2 D_{Q'}$; en particulier si l'on travaille sur $\mathbf{F}_p$ on a $\left(\frac{D_Q}{p}\right) = \left(\frac{D_{Q'}}{p}\right)$. Cette remarque est utilisée dans la preuve du théorème suivant.

Théorème. *Soit Q une forme quadratique en n variables, non dégénérée, à coefficients dans $\mathbf{F}_p$ (où $p \neq 2$) alors*

$$\mathrm{card}\,\{x \in (\mathbf{F}_p)^n \mid Q(x) = 0\} = p^{n-1} + \epsilon(p-1)p^{\frac{n}{2}-1}$$

avec

$$\epsilon = \begin{cases} 0 & \text{si } n \text{ est impair} \\ \left(\frac{(-1)^{n/2}D_Q}{p}\right) & \text{si } n \text{ est pair} \end{cases}$$

Preuve. D'après ce qui précède, nous pouvons nous ramener au cas où la forme Q est diagonale, c'est-à-dire $Q(x) = a_1x_1^2 + \ldots + a_nx_n^2$; notons N_p le cardinal que nous voulons calculer. On a :

$$\begin{aligned}
pN_p &= \sum_{a=0}^{p-1} \sum_{x\in\mathbf{F}_p^n} \exp\left(\frac{2\pi i a Q(x)}{p}\right) \\
&= p^n + \sum_{a=1}^{p-1} \sum_{x\in\mathbf{F}_p^n} \exp\left(\frac{2\pi i a Q(x)}{p}\right) \\
&= p^n + \sum_{a=1}^{p-1} \sum_{x_1,\ldots,x_n\in\mathbf{F}_p} \exp\left(\frac{2\pi i a(a_1x_1^2 + \ldots + a_nx_n^2)}{p}\right) \\
&= p^n + \sum_{a=1}^{p-1} \prod_{j=1}^{n} \sum_{x_j\in\mathbf{F}_p} \exp\left(\frac{2\pi i a a_j x_j^2}{p}\right) = p^n + \sum_{a=1}^{p-1} \prod_{j=1}^{n} \tau(aa_j) \\
&= p^n + \tau(1)^n \left(\frac{a_1\ldots a_n}{p}\right) \sum_{a=1}^{p-1} \left(\frac{a}{p}\right)^n .
\end{aligned}$$

Or $a_1 \ldots a_n = D_Q$ et la somme $\sum_{a=1}^{p-1} \left(\frac{a}{p}\right)^n$ vaut 0 (resp. $p-1$) si n est impair (resp. si n pair). On en tire déjà que $N_p = p^{n-1}$ si n est impair. Si n est pair on remarque que

$$\tau(1)^n = \left(\tau(1)^2\right)^{n/2} = \left(\frac{-1}{p}\right)^{n/2} p^{n/2}$$

et on obtient bien la formule annoncée pour N_p. □

Remarque. Cet énoncé permet de retrouver de manière beaucoup plus précise le théorème de Chevalley-Warning pour le cas des formes quadratiques. C'est clair si la forme quadratique est non dégénérée ; il faut ajouter qu'une forme dégénérée peut s'écrire après changement de variables $Q(x_1, \ldots, x_n) = a_1x_1^2 + \ldots + a_rx_r^2$ avec $r < n$ et $D'_Q := a_1 \ldots a_r \neq 0$. On en tire que dans ce cas

$$N_p = p^{n-r}\left(p^{r-1} + \epsilon(p-1)p^{\frac{r}{2}-1}\right) = p^{n-1} + \epsilon(p-1)p^{n-\frac{r}{2}-1}$$

où maintenant ϵ est nul si r est impair et vaut $\left(\frac{(-1)^{r/2}D'_Q}{p}\right)$ si r est pair.

Considérons maintenant une forme quadratique $Q(x) = \sum_{1\leq i,j\leq n} a_{ij}x_ix_j$ à coefficients *entiers*. Si l'on veut compter maintenant le nombre de solutions modulo N avec N non nécessairement premier, on pourra avoir recours aux deux lemmes suivant (dont le premier est une variante du lemme chinois).

Lemme. *Posons $\psi_Q(N) := \operatorname{card}\{x \bmod N \mid Q(x) \equiv 0 \bmod N\}$ alors, si M et N sont premiers entre eux, on a $\psi_Q(MN) = \psi_Q(M)\psi_Q(N)$.*

Preuve. C'est un corollaire du lemme chinois : on a $Q(x) \equiv 0 \bmod MN$ si et seulement si $Q(x) \equiv 0 \bmod N$ et $Q(x) \equiv 0 \bmod M$ et de plus chaque paire de classes de congruence $x \equiv a \bmod M$, $x \equiv b \bmod N$ correspond à une classe de congruence $\bmod MN$. □

Ce lemme permet de se ramener à compter des solutions modulo p^m. Ceci peut se faire grâce au lemme suivant qui est un cas particulier du "Lemme de Hensel".

Lemme. *Soit p un nombre premier impair ne divisant pas D_Q, définissons l'ensemble des solutions modulo p^m et "non singulières" par*

$$\mathcal{C}_Q(p^m) := \{x \bmod p^m \mid Q(x) \equiv 0 \bmod p^m \text{ et } x \not\equiv 0 \bmod p\}.$$

On a la formule

$$\operatorname{card} \mathcal{C}_Q(p^m) = p^{(m-1)(n-1)} \operatorname{card} \mathcal{C}_Q(p) = p^{(m-1)(n-1)} \left(p^{n-1} - 1 + \epsilon(p-1)p^{\frac{n}{2}-1}\right).$$

Preuve. La deuxième égalité est bien sûr un corollaire de la première et du théorème précédent. On a une application évidente de $\mathcal{C}_Q(p^{m+1})$ vers $\mathcal{C}_Q(p^m)$ qui à un n-uplet d'entiers modulo p^{m+1} associe le même n-uplet d'entiers modulo p^m. Il suffit de montrer que cette application est surjective et que chaque fibre a cardinal p^{n-1} car alors on aura $\operatorname{card} \mathcal{C}_Q(p^{m+1}) = p^{n-1} \operatorname{card} \mathcal{C}_Q(p^m)$ et le lemme en découle aisément. Soit donc un n-uplet d'entiers x_0 tel que $Q(x_0) \equiv 0 \bmod p^m$, ou encore tel que $Q(x) = p^m a_0$, on remarque que :

$$Q(x_0 + p^m z) = Q(x_0) + 2p^m B(x_0, z) + p^{2m} Q(z) \equiv p^m \left(a_0 + 2B(x_0, z)\right) \bmod p^{m+1}$$

donc est nul modulo p^{m+1} si et seulement si

$$a_0 + 2B(x_0, z) \equiv 0 \bmod p.$$

Comme $x_0 \not\equiv 0 \bmod p$ et B est une forme bilinéaire non dégénérée par hypothèse, cette dernière équation est celle d'un hyperplan (affine) das $\mathbf{F}_p^n$; il y a donc exactement p^{n-1} solutions modulo p en z. □

Première partie (applications I) : Algorithmes, primalité et factorisation.

A. Algorithmes de base.
B. Cryptographie, RSA.
C. Test de Primalité (I).
D. Test de Primalité (II).
E. Factorisation.

A. Algorithmes de base.

On décrit succintement d'une part les principaux algorithmes ainsi que leur complexité ou temps de calcul théorique. On utilise la notation $O(f(n))$ pour désigner une fonction $\leq Cf(n)$; par ailleurs, les constantes apparaissant n'ayant – au moins d'un point de vue théorique – aucune importance seront négligées.

Soit n un entier, une fois qu'on a choisi une base $b \geq 2$, on écrit n en base b c'est-à-dire avec des *chiffres* $a_i \in [0, b-1]$:

$$n = a_0 + a_1 b + \ldots + a_r b^r = \overline{a_0 a_1 a_2 \ldots a_r}^{\,b}, \qquad \text{avec disons } a_r \neq 0$$

(les deux choix les plus courants étant $b = 10$ – écriture décimale usuelle – et $b = 2$ – écriture binaire, particulièrement adaptée à la programmation sur machine). On considérera une opération sur les chiffres comme une unique opération (ou encore comme une opération nécessitant $O(1)$ temps machine). Il est naturel d'appeler *complexité* du nombre n le nombre de chiffres nécessaires pour le décrire, c'est-à-dire r ; comme $b^r \leq a_r b^r \leq n \leq b^{r+1}$ on voit que

$$r \leq \frac{\log n}{\log b} \leq r + 1$$

et on décrira donc cette complexité comme proportionnelle à $\log n$. Il est clair que la manipulation de nombres quelconques de taille n recquiert au moins $\log n$ opérations élémentaires ; on considère, tant d'un point de vue pratique que théorique, qu'un "bon" algorithme est un algorithme *polynomial* c'est-à-dire utilisant $O\left((\log n)^\kappa\right)$ opérations élémentaires. Inversement on considère qu'un algorithme *exponentiel*, c'est-à-dire ayant un temps d'exécution ou requérant un nombre d'opérations supérieur à $\exp(\kappa \log n) = n^\kappa$ est impraticable (pour n grand bien sûr).

Addition. Pour additionner $n + m$ deux nombres avec au plus r chiffres, on doit faire au plus r additions de deux chiffres et (éventuellement) propager une retenue. Le coût est donc $O\left(\log \max(n, m)\right)$. Le coût d'une soustraction est similaire.

Multiplication. Pour calculer $n \times m$, où n et m sont deux nombres avec au plus r chiffres (avec l'algorithme appris à l'école primaire) on effectue au plus r^2 multiplications élémentaires et r additions, éventuellement avec retenues, d'où un coût en $O\left(r^2\right) = O\left((\log \max(n, m))^2\right)$.

Remarque. L'algorithme d'addition est optimal (aux constantes près) mais des méthodes plus sophistiquées ("transformation de Fourier rapide") peuvent permettre d'effectuer des multiplications avec un coût en $O\left(\log \max(n, m) \log\log \max(n, m)^2\right)$ par exemple.

Division euclidienne. Etant donné a et $b \geq 1$, si on calcule (q, r) tels que $a = qb + r$ et $0 \leq r \leq b - 1$ avec (une variante de) l'algorithme appris à l'école primaire, on utilise un nombre d'opérations élémentaires similaire à celui de la multiplication i.e. $O(\log \max(a, b)^2)$. Pour donner un exemple d'*algorithme tortue* (à ne pas utiliser!), proposons la procédure suivante. On commence par poser $q_0 = 0$ et $r_0 = a$. On a donc $a = q_0 b + r_0$, si $r_0 < b$, on s'arrête, sinon on calcule $q_1 = q_0 + 1$ et $r_1 = r_0 - b$ de sorte que $a = q_1 b + r_1$ et on obtient le résultat en itérant et s'arrêtant quand $r_n < b$ et $a = q_n b + r_n$. Si, disons $a > b$ on doit effectuer environ a/b soustractions, donc le coût est $O((\log a) \times (a/b))$ (ce qui est exponentiel).

Algorithme d'Euclide (étendu). Il s'agit, étant donné deux entiers a, b de calculer $d := \text{pgcd}(a, b)$ et $(u, v) \in \mathbf{Z}^2$ tel que $au + bv = d$ (théorème de Bézout). Le principe est le suivant : on effectue la division de a par b, disons $a = bq_1 + r_1$; puis la division de b par r_1, disons $b = r_1 q_2 + r_2$ et plus généralement la division de r_n

par r_{n+1}, disons $r_n = r_{n+1}q_{n+2} + r_{n+2}$; on observe que la suite r_n est strictement décroissante et on s'arrête quand disons $r_{n+1} = 0$ et alors $\mathrm{pgcd}(a, b) = r_n$. En effet

$$\mathrm{pgcd}(a, b) = \mathrm{pgcd}(b, r_1) = \mathrm{pgcd}(r_1, r_2) = \ldots = \mathrm{pgcd}(r_n, r_{n+1}) = r_n.$$

Pour le calcul de (u, v) on peut procéder ainsi : on pose $u_0 = 1$, $u_1 = 0$, $v_0 = 0$ et $v_1 = 1$ puis définir par récurrence $u_n = u_{n-2} - q_n u_{n-1}$ et $v_n = v_{n-2} - q_n v_{n-1}$. On vérifie immédiatement par récurrence que $au_n + bv_n = r_n$. Estimons maintenant le nombre maximal de divisions euclidiennes à effectuer. On peut supposer $r_0 = a \geq r_1 = b$, on a ensuite $r_n = r_{n+1}q_{n+2} + r_{n+2} \geq r_{n+1} + r_{n+2}$. Si $r_0 > r_1 > \ldots > r_n = d$ est la suite donnant le pgcd, posons $d_i = r_{n-i}$, on a alors $d_{i+2} \geq d_{i+1} + d_i$. Posons $\alpha := (1 + \sqrt{5})/2$ la racine positive de $X^2 = X + 1$, on a alors $d_i \geq \alpha^i$. En effet $d_0 = d \geq 1 = \alpha^0$, $d_1 \geq d_0 + 1 \geq 2 \geq \alpha^1$ et si l'inégalité est vraie jusqu'à $i+1$ on a $d_{i+2} \geq d_{i+1} + d_i \geq \alpha^{i+1} + \alpha^i = \alpha^{i+2}$. On conclut que $a = d_n \geq \alpha^n$ ou encore que le nombre d'étapes est majoré par $\log(a)/\log(\alpha) = O(\log a)$. Le coût total est donc $O\left(\log \max\{|a|, |b|\}^3\right)$.

Calculs dans $\mathbf{Z}/N\mathbf{Z}$. Il s'agit de faire les opérations d'addition et de multiplications sur deux entiers inférieurs à N puis de prendre le reste pour la division euclidienne par N. Pour calculer l'inverse de a modulo N. Le calcul est un corollaire du précédent : si a est un entier, l'algorithme d'Euclide étendu nous répond soit que $\mathrm{pgcd}(a, N) > 1$ – auquel cas a n'est pas inversible modulo N – soit qu'il existe u, v (calculés par l'algorithme) tels que $au + Nv = 1$ et alors l'inverse de a est la classe de u modulo N. Le coût est donc le même que celui de l'algorithme d'Euclide étendu.

Exponentiation. Pour calculer a^m bien sûr on pourrait calculer $a \times a \times \ldots a$ mais cela obligerait à effectuer $m-1$ multiplications ; on peut faire beaucoup mieux et effectuer le calcul avec $O(\log m)$ multiplication. Par exemple, si $m = 2^r$ on élevera r fois au carré, i.e. on effectuera r multiplications. Dans le cas général on écrira m en binaire $m = \epsilon_0 + \epsilon_1 2 + \ldots + \epsilon_r 2^r$ et on calculera :

$$a^m = \left(\left(\left(a^{\epsilon_r}\right)^2 a^{\epsilon_{r-1}}\right)^2 a^{\epsilon_{r-2}} \cdots\right)^2 a^{\epsilon_0}$$

On peut décrire l'algorithme itérativement ainsi : on part des données initiales $(u, v, n) := (1, a, m)$ et on itère ainsi : si n est pair, on remplace (u, v, n) par $(u, v^2, n/2)$ et si n est impair, on remplace (u, v, n) par $(uv, v^2, (n-1)/2)$, on s'arrête quand $n = 0$ et on a alors $u = a^m$. Comme n est au moins divisé par deux à chaque pas, le nombre r d'étapes est tel que $2^r \leq m$ donc on doit calculer $O(\log m)$ multiplications. Si on calcule $\bmod N$ on réduit chaque résultat $\bmod N$ et on fait donc à chaque étape la multiplication d'entiers $\leq N$. Le coût total pour calculer $a^m \bmod N$ est donc $O\left(\log m (\log N)^2\right)$.

Calculs dans $\mathbf{F}_q$ et $\mathbf{F}_q^*$. Nous supposerons que le corps fini $\mathbf{F}_q = \mathbf{F}_{p^f}$ est décrit par un polynôme $P(X) = X^f + p_{f-1}X^{f-1} + \ldots + p_0 \in \mathbf{F}_p[X]$ unitaire irréductible de degré f ; on identifie donc $\mathbf{F}_q$ à $\mathbf{F}_p[X]/P\mathbf{F}_p[X]$ ou encore à l'espace vectoriel sur $\mathbf{F}_p$ de base $1, x, x^2, \ldots, x^{f-1}$ avec l'addition coordonnée par coordonnée et la multiplication définie par $x^i \cdot x^j = x^{i+j}$ et $x^f = -p_{f-1}x^{f-1} - \ldots - p_0$. Un élément de $\mathbf{F}_q$ est donc vu comme un f-uplet d'entiers modulo p ou comme un polynôme de degré $\leq f - 1$. Pour effectuer une addition, on doit effectuer f additions dans $\mathbf{F}_p$, soit un coût $O(f \log p) = O(\log q)$. Pour effectuer une multiplication on fait le produit des deux polynômes, soit essentiellement f^2 multiplications dans $\mathbf{F}_p$, puis on fait la division euclidienne du résultat par $P(X)$, soit essentiellement $O(f)$ divisions et $O(f^2)$ multiplications dans $\mathbf{F}_p$. Le coût d'une multiplication dans $\mathbf{F}_q$ est donc $O(f^2(\log p)^2) + O(f(\log p)^3)$. On remarque que ce coût est toujours $O((\log q)^3)$ mais que si on choisi $q = 2^f$ par exemple, il est $O(f^2) = O((\log q)^2)$.

B. Cryptographie, RSA.

On ne s'intéresse ici qu'à un seul aspect de la cryptographie, et un seul système à "clefs publiques", dit RSA du nom de ses trois inventeurs Rivest, Shamir et Adleman – c'est en fait un des plus utilisés.

La cryptographie est l'art (ou la science) des messages secrets : on veut pouvoir envoyer des informations sans qu'une autre personne que le destinaire puisse en bénéficier. Un problème annexe est de pouvoir identifier avec certitude l'auteur du message. On pense communément que le seul moyen est d'utiliser un "code secret";

en fait l'originalité de la cryptographie "à clefs publiques" réside précisément dans le fait que le code n'est pas secret mais connu (au moins en grande partie) de tous! Ce n'est pas seulement une curiosité mathématique, c'est aussi le principe régissant les cartes bancaires, les transactions sur Internet, etc ...

Le principe général est le suivant : on appelle $\mathcal{M}$ l'ensemble des messages (en pratique on prendra $\mathcal{M} = [0, N-1]$ ou encore $\mathcal{M} = \mathbf{Z}/N\mathbf{Z}$) ; deux personnes A et B souhaitant échanger des messages sans qu'une troisième personne C puisse les déchiffrer choisissent chacun des bijections $f_{\mathrm{A}}, f_{\mathrm{B}} : \mathcal{M} \to \mathcal{M}$; l'ensemble $\mathcal{M}$ (disons l'entier N) est connu de tous, de même que f_{A} et f_{B}, par contre – et c'est l'idée centrale – la bijection réciproque f_{A}^{-1} (resp. f_{B}^{-1}) n'est connue que de A (resp. de B). Cela ne signifie pas bien sûr que, connaissant f_{A}, il est impossible de calculer f_{A}^{-1} mais que ce calcul serait si long qu'il est irréalisable en pratique ; nous verrons plus loin comment construire de telles fonctions.

Quand A veut envoyer à B un message $m \in \mathcal{M}$ (disons un entier modulo N), elle envoie en clair $m' = f_{\mathrm{B}} \circ f_{\mathrm{A}}^{-1}(m)$; noter qu'elle connaît f_{B} (qui est dans l'annuaire) et f_{A}^{-1} (qui est son secret). Pour déchiffrer le message B calcule $f_{\mathrm{A}} \circ f_{\mathrm{B}}^{-1}(m')$ qui lui redonne m ; noter qu'elle connaît f_{A} (qui est dans l'annuaire) et f_{B}^{-1} (qui est son secret). Le système possède un double avantage : non seulement C ne pourra déchiffrer le message qu'en calculant f_{B}^{-1} (ce qui est supposé impraticable) mais B peut être sûre que c'est bien A qui lui a envoyé le message puisque celui-ci a dû être codé en utilisant f_{A}^{-1} que seul A connait!

Ce procédé est une forme simplifié de procédures connues sous le nom de protocole de Diffie-Hellman (1976) ; sa sécurité repose sur le choix de fonctions f "à sens unique" c'est-à-dire telles que f soit facile (rapide) à calculer mais f^{-1} soit impossible en pratique à calculer. Plusieurs constructions de fonctions ont été proposées mais une des plus robustes et des plus utilisées repose sur le constat simple que, si p, q sont de très grands nombres premiers (disons une centaine de chiffres) alors le calcul de leur produit $N := pq$ peut s'effectuer très rapidement (disons dix mille opérations élémentaires), cependant, si l'on ne connaît que N, le calcul de la factorisation est extrêmement long, voire impossible en pratique.

Construisons maintenant les fonctions f_A du système RSA. On choisit deux très grands nombres premiers p et q, on calcule $N := pq$ et on choisit également un entier d (de taille moyenne) premier avec $\phi(N) = (p-1)(q-1)$. La clef publique est alors (N, d), par contre p et q sont secrets et on pose, pour a entier inférieur à N :

$$f(a) := a^d \bmod N.$$

Pour décoder, on calcule e l'inverse de d modulo $\phi(N)$ et on observe que

$$f^{-1}(b) = b^e \bmod N$$

puisque $\left(a^d\right)^e = a^{ed} \equiv a \bmod N$ car $a^{\phi(N)} \equiv 1 \bmod N$.

Remarques. 1) Il y a une petite contrainte sur le "message" a : il doit être premier avec N[(1)]. Cependant on observera que la proportion des entiers premiers avec N est $\phi(N)/N = (1-1/p)(1-1/q)$; ainsi, si p, q sont par exemple $\geq 10^{50}$, la proportion d'entiers non premiers à N est $\leq 2.10^{-50}$. 2) Une fois choisis p, q et d, le calcul de N, $\phi(N)$ et e s'effectue en temps polynomial (rapide) ; de même l'opération $a \mapsto f(a)$ est rapide tout comme $a \mapsto f^{-1}(a)$ *si l'on connait* e. 3) On peut voir que, au moins heuristiquement, la connaissance de e permet de factoriser N : si on écrit $de - 1 = 2^r M$ (avec M impair), en calculant $\mathrm{pgcd}(a^{2^j M} \pm 1, N)$ pour $j = 1, 2, \ldots$ et quelques valeurs de a, on a de bonnes chances de factoriser rapidement N. 4) Ainsi, si l'on connait seulement la clef publique (N, d), on doit *a priori* factoriser N pour calculer $\phi(N)$ puis e. En effet la connaissance de $\phi(N)$ équivaut à celle de p et q puisque $\phi(N) = N - (p+q) + 1$ (la connaissance du produit et de la somme de deux entiers permet de déterminer facilement la paire d'entiers).

Ce système soulève plusieurs problèmes pour lesquelles les réponses sont plus ou moins satisfaisantes.
(i) Comment fabriquer de (très) grands nombres premiers?
(ii) Quelles sont les méthodes pour factoriser un entier?

(1) Si par mégarde on envoyait un message $a = pa'$, on pourrait certes encore le décoder par $f(a)^e = p^{de} a'^{ed} = p^{ed} a' = a$ mais C, ou n'importe qui, n'aura qu'à calculer $\mathrm{pgcd}(a, N)$ pour découvrir p et casser le code!

(iii) Comment doivent être choisis p et q dans RSA pour résister aux méthodes de factorisation?

Comme il est clair, par la question (iii), que les nombres premiers ne doivent pas être trop "spéciaux" la question (i) est essentiellement équivalente à la question :

(I) (Test de Primalité) Donner un algorithme rapide décidant si un nombre N donné est premier ou composé.

Si l'on dispose d'un tel algorithme $\mathcal{P}$, on pourra en effet décider d'une taille d'entier (par exemple $N \sim 10^{50}$) choisir aléatoirement un entier N_1 impair de cette taille, tester si $\mathcal{P}(N_1)$ puis $\mathcal{P}(N_1+2)$, $\mathcal{P}(N_1+4)$ jusqu'à trouver un nombre premier. D'après les théorèmes de répartition des nombres premiers, le nombre de nombres premiers dans un intervalle $[N_1, N_1+H]$ est environ $H/\log(N_1)$; on peut donc s'attendre à trouver un nombre premier au bout de $O(\log(N_1))$ essais.

Nous allons voir qu'on dispose de réponses satisfaisantes pour (I) mais seulement de réponses très partielles pour les autres questions.

C. Test de Primalité (I).

On considère N *impair* et le problème de déterminer si N est premier ou non. On note (M,N) le PGCD de M et N. La lettre p est réservé à un nombre dont on sait qu'il est premier. Le premier, et en quelque sorte le parent, de tous les tests de primalité est

Lemme. (Fermat) *Si N est premier et $(a,N)=1$ alors $a^{N-1} \equiv 1 \bmod N$.*

Preuve. Le groupe $\mathbf{Z}/N\mathbf{Z}^*$ est d'ordre $N-1$, le lemme en découle d'après le théorème de Lagrange.[1] □

Ce test est "bon" dans le sens où calculer $a^{N-1} \bmod N$ requiert $O(\log N)$ mutiplications (à condition d'utiliser bien sûr l'écriture binaire de $N-1$). Cependant il est "mauvais" car il existe des nombres, dit *nombres de Carmichael* qui vérifient le test sans être premiers. On sait même qu'il en existe une infinité, le plus petit étant $561 = 3.11.17$. On voit facilement qu'un nombre N est de Carmichael si et seulement si N est sans facteur carré et $p-1$ divise $N-1$ pour tout p divisant N. En fait, plus généralement on peut introduire $\lambda(N)$, l'exposant du groupe $(\mathbf{Z}/N\mathbf{Z})^*$, appelée parfois *indicateur de Carmichael* ; c'est le plus petit entier naturel (au sens de la divisibilité ou de l'ordre usuel) tel que pour tout a premier avec N on ait $a^{\lambda(N)} \equiv 1 \bmod N$. D'après ce que nous avons vu, on sait que, si $N = p_1^{m_1} \dots p_k^{m_k}$ est impair on a

$$\lambda(N) = \operatorname{ppcm}\left(p_1^{m_1-1}(p_1-1), \dots, p_k^{m_k-1}(p_k-1)\right)$$

On a toujours $\lambda(N)$ divise $\phi(N)$ et on a égalité seulement si $(\mathbf{Z}/N\mathbf{Z})^*$ est cyclique, i.e. si $N = p^\alpha$ ou $2p^\alpha$.

Lemme. (Euler[2]) *Si N est premier et $(a,N)=1$, alors $a^{\frac{N-1}{2}} \equiv \left(\frac{a}{N}\right) \bmod N$.*

Preuve. Si a est un carré b^2 alors $a^{\frac{N-1}{2}} = b^{N-1} = 1$; les carrés forment un sous-groupe d'indice 2 dans le groupe cyclique $\mathbf{Z}/N\mathbf{Z}^*$ (si N premier) donc si $a^{\frac{N-1}{2}} = 1$ alors a est un carré. □

Ce test, dit test de Solovay-Strassen, est toujours polynomial (pour une valeur de a) et est meilleur que celui de Fermat; en particulier :

Lemme. *Soit $H := \left\{ a \in \mathbf{Z}/n\mathbf{Z}^* \mid a^{\frac{N-1}{2}} \equiv \left(\frac{a}{N}\right) \bmod N \right\}$, alors $H = \mathbf{Z}/n\mathbf{Z}^*$ si et seulement si N est premier.*

Preuve. On a vu que si N est premier, alors $H = \mathbf{Z}/n\mathbf{Z}^*$. Si p^2 divise N, il existe a d'ordre $p(p-1)$ or p ne divise pas $N-1$ donc $a^{N-1} \neq 1$. Si $N = pp_2 \dots p_r$ avec $r \geq 2$, choisissons (par le lemme chinois) $a \equiv 1$ modulo $p_2, \dots, p_r$ et non carré modulo p; alors $\left(\frac{a}{N}\right) = -1$ mais $a^{(N-1)/2} \equiv 1 \bmod p_2 \dots p_r$ donc $a^{(N-1)/2} \not\equiv -1 \bmod N$. □

Applications.

---

[1] Montrer le petit théorème de Fermat à partir du théorème de Lagange est évidemment un anachronisme.

[2] Appeler lemme d'Euler un énoncé utilisant le symbole de Jacobi ou Legendre est aussi un anachronisme.

(i) Test probabiliste polynomial. Si N est composé on a $(\mathbf{Z}/N\mathbf{Z}^* : G) \geq 2$ donc en prenant a aléatoirement, on a au moins une chance sur deux d'avoir $a \notin G$. Ainsi, si N passe successivement k tests, on peut dire qu'il est premier avec une probabilité supérieure à $1 - 2^{-k}$.

(ii) Test déterministe polynomial (sous GRH). La théorie analytique permet de démontrer que, si les fonctions $L(\chi, s)$ ne s'annulent pas pour $\mathrm{Re}(s) > 1/2$ (Hypothèse de Riemann généralisée, GRH), alors pour tout caractère $\chi : \mathbf{Z}/n\mathbf{Z}^* \to \mathbf{C}^*$, il existe un $a \leq 2(\log N)^2$ tel que $\chi(a) \neq 0, 1$. Ainsi on voit que si N est composé, il existera $a \leq 2(\log N)^2$ qui ne passera pas le test de Solovay-Strassen. En essayant tous les $a \in [2, 2(\log N)^2]$, on obtient donc un certificat de primalité conditionnel à l'hypothèse de Riemann.

On peut améliorer le test et l'algorithme de Solovay-Strassen ainsi.

Lemme. (Rabin-Miller) *Soit N impair, posons $N-1 = 2^s M$ avec M impair. Si N est premier et $(a, N) = 1$, alors ou bien $a^M \equiv 1 \bmod N$ ou bien il existe $0 \leq r \leq s-1$ tel que $a^{2^r M} \equiv -1 \bmod N$.*

Preuve. L'ordre de a modulo N est $2^t M'$ avec $0 \leq t \leq s$ et M' impair divisant M. Si $t = 0$ alors $a^{M'} = 1$ donc $a^M = 1$. Si $t \geq 1$ alors, comme N est premier, $a^{2^{t-1}M'} = -1$ donc $a^{2^{t-1}M} = -1$. □

Ce test est meilleur que celui d'Euler car, d'une part la propriété pour la paire a, N entraîne celle d'Euler pour la paire a, N, d'autre part, si N est composé, la proportion de a passant le test raffiné est $\leq 1/4$ et même souvent plus petite. On a bien sûr une version probabiliste polynomiale du test raffiné et une version déterministe polynomiale sous l'hypothèse de Riemann.

Remarque. Si $N \equiv 3 \bmod 4$ alors "Rabin-Miller = Solovay-Strassen" et équivaut même à $a^{(N-1)/2} \equiv \pm 1 \bmod N$. En effet $(N-1)/2$ est impair et on peut observer que si $\epsilon = \pm 1$ alors $\left(\frac{\epsilon}{N}\right) = \epsilon$ et que si $a^{(N-1)/2} \equiv \pm 1 \bmod N$ alors

$$\left(\frac{a}{N}\right) = \left(\frac{a \cdot (a^2)^{(N-3)/4}}{N}\right) = \left(\frac{a^{(N-1)/2}}{N}\right) = a^{(N-1)/2} \bmod N.$$

Preuve. (que "Rabin-Miller > Solovay-Strassen") On sait donc que $a^{(N-1)/2} = a^{2^{s-1}M}$ vaut $-1 \bmod N$ si $r = s-1$ et $1 \bmod N$ dans tous les autres cas. Cherchons donc à calculer $\left(\frac{a}{N}\right)$. Si $a^M \equiv 1 \bmod N$ alors $\left(\frac{a}{N}\right) = \left(\frac{a}{N}\right)^M = \left(\frac{a^M}{N}\right) = 1$ donc on a bien $a^{\frac{N-1}{2}} \equiv \left(\frac{a}{N}\right) \bmod N$. Supposons maintenant que $a^{2^r M} \equiv -1 \bmod N$. Soit p_i divisant N, écrivons $p_i - 1 = 2^{s_i} M_i$, alors, comme $a^{2^r M} \equiv -1 \bmod p_i$, l'ordre de a modulo p_i est de la forme $2^{r+1} L_i$ (avec L_i impair). Ainsi, modulo $\bmod p_i$, on a :

$$\left(\frac{a}{p_i}\right) \equiv a^{(p_i-1)/2} \equiv a^{2^{s_i-1}M_i} \equiv \begin{cases} 1 & \text{si } s_i > r+1 \\ -1 & \text{si } s_i = r+1 \end{cases}.$$

Remarquons qu'on a toujours $r + 1 \leq s_i$. Appelons h le nombre des p_i (éventuellement avec répétition) tels que $s_i = r + 1$. On a alors $\left(\frac{a}{N}\right) = (-1)^h$. D'autre part modulo 2^{r+2} on a $N = 1 + 2^s M = \prod_i p_i = \prod_i (1 + 2^{s_i}) \equiv 1 + h2^{r+1} \bmod 2^{r+2}$. Dans le cas où $r < s-1$, on doit avoir h pair donc $\left(\frac{a}{N}\right) = 1$ et on a bien $a^{(N-1)/2} \equiv 1 \bmod N$. Dans le cas où $r = s-1$ alors h est impair et $\left(\frac{a}{N}\right) = -1 \equiv a^{(N-1)/2} \bmod N$. □

On peut résumer les discussions précédentes en introduisant les ensembles suivants :

$G_0 := (\mathbf{Z}/N\mathbf{Z})^*$
$G_1 := \left\{a \in (\mathbf{Z}/N\mathbf{Z})^* \mid a^{N-1} \equiv 1 \bmod N\right\}$
$G_2 := \left\{a \in (\mathbf{Z}/N\mathbf{Z})^* \mid a^{(N-1)/2} \equiv \pm 1 \bmod N\right\}$
$G_3 := \left\{a \in (\mathbf{Z}/N\mathbf{Z})^* \mid a^{(N-1)/2} \equiv \left(\frac{a}{N}\right) \bmod N\right\}$
$S := \left\{a \in (\mathbf{Z}/N\mathbf{Z})^* \mid a^M \equiv 1 \bmod N \text{ ou il existe } r \in [0, s-1], \text{ tel que } a^{2^r M} \equiv -1 \bmod N\right\}$

On a toujours $S \subset G_3 \subset G_2 \subset G_1 \subset G_0$ et on a égalité partout si et seulement si N est premier ou encore si et seulement si $G_3 = G_0$. Par ailleurs, G_1, G_2 et G_3 sont des sous-groupes mais pas en général S, même si, dans le cas $N \equiv 3 \bmod 4$ on a vu que $G_2 = G_3 = S$. En effet S est stable par inversion et si $a, b \in S$ ne vérifient pas la même congruence ou tous deux $a^M = b^M = 1$ alors $ab \in S$ mais si $a^{2^r M} = b^{2^r M} = -1$ on peut avoir $ab \notin S$. Par exemple si $\epsilon^2 = 1$ mais $\epsilon \neq \pm 1$ et si $a^{2M} = -1$ (ce qui impose $N \equiv 1 \bmod 4$) alors $a \in S$

et $a\epsilon \in S$ puisque $(a\epsilon)^{2M} = -1$. Cependant $(\epsilon a^2)^M = \epsilon^M a^{2M} = -\epsilon \neq \pm 1$ et $(\epsilon a^2)^{2M} = 1$ donc $\epsilon a^2 \notin S$. En considérant $a \mapsto \left(\frac{a}{N}\right) a^{(N-1)/2}$ de G_2 vers $\{\pm 1\}$, on voit que $(G_2 : G_3) = 1$ ou 2. On va maintenant tenter de calculer le cardinal de l'ensemble S.

Définition. Soit A, B des entiers, on pose

$$\phi(A; B) = \operatorname{card}\left\{a \in (\mathbf{Z}/A\mathbf{Z})^* \mid a^B \equiv 1 \bmod A\right\}.$$

Lemme. *Soit $t \geq 0$, $N = 1 + 2^s M = p_1^{\alpha_1} \dots p_k^{\alpha_k}$ (avec M impair), posons $p_i - 1 = 2^{s_i} M_i$, $s_i' = \min(t, s_i)$ et $t_i := pgcd(M, M_i)$, alors*

$$\phi(N, 2^t M) = 2^{s_1' + \dots + s_k'} t_1 \dots t_k.$$

Par ailleurs le cardinal de l'ensemble

$$\left\{a \in (\mathbf{Z}/N\mathbf{Z})^* \mid a^{2^t M} \equiv -1 \bmod N\right\}$$

est nul si $t \geq \min_i s_i$ et égal à $\phi(N, 2^t M) = 2^{tk} t_1 \dots t_k$ si $t < \min_i s_i$.

Preuve. Notons $N = p_1^{\alpha_1} \dots p_k^{\alpha_k}$. On a $a^{2^t M} \equiv 1 \bmod N$ si et seulement si $a^{2^t M} \equiv 1 \bmod p_j^{\alpha_j}$ pour $j = 1, \dots, k$. Maintenant $(\mathbf{Z}/p_j^{\alpha_j}\mathbf{Z})^*$ est cyclique de cardinal $(p_j - 1)p_j^{\alpha_j - 1}$ donc le nombre de solution est

$$pgcd(2^t M, (p_j - 1)p_j^{\alpha_j - 1}) = pgcd(2^t M, 2^{s_j} M_j) = 2^{\min(t, s_j)} t_j.$$

Par le lemme chinois, le nombre de solution modulo N est donc le produit de ces nombres comme annoncé. Pour la deuxième affirmation, on voit tout de suite que soit il n'existe aucune solution soit il en existe une et alors l'ensemble des solutions est en bijection avec les solutions de la congruence précédente. La congruence $a^{2^t M} \equiv -1 \bmod p_j^{\alpha_j}$ est résoluble si et seulement si 2^{t+1} divise $(p_j - 1)p_j^{\alpha_j - 1}$, c'est-à-dire si et seulement si $t + 1 \leq s_j$ d'où le résultat. □

Supposons $s_1 \leq s_2 \leq \dots \leq s_k$. En décomposant l'ensemble S en $S_0 := \left\{a \in (\mathbf{Z}/N\mathbf{Z})^* \mid a^M \equiv 1 \bmod N\right\}$ et $T_j := \left\{a \in (\mathbf{Z}/N\mathbf{Z})^* \mid a^{2^j M} \equiv -1 \bmod N\right\}$ pour $0 \leq j \leq s_1 - 1$, on obtient, en appliquant le lemme à chacun de ces ensembles :

$$\operatorname{card}(S) = t_1 \dots t_k \left(1 + 1 + 2^k + \dots + 2^{k(s_1 - 1)}\right) = t_1 \dots t_k \left(\frac{2^{k s_1} + 2^k - 2}{2^k - 1}\right).$$

La proportion de $a \in G_0$ qui passe le test de Rabin-Miller est donc

$$\frac{\operatorname{card}(S)}{\operatorname{card}(G_0)} = \frac{t_1 \dots t_k}{M_1 \dots M_k} \frac{2^{-s_1 - \dots - s_k}}{p_1^{\alpha_1 - 1} \dots p_k^{\alpha_k - 1}} \left(\frac{2^{k s_1} + 2^k - 2}{2^k - 1}\right).$$

Remarque Les deux cas les plus "méchants" sont les cas suivants :

(i) On a $N = pq$ avec $q = 2p - 1$ et $p \equiv 3 \bmod 4$. Par exemple $N = 3 \cdot 5$, $N = 7 \cdot 13$ etc. On a alors $p = 1 + 2M_1$ et $q = 1 + 4M_1$ et $N = (1 + 2M_1)(1 + 4M_1) = 1 + 2M_1(3 + 4M_1)$ donc $t_1 = t_2 = M_1 = M_2$ et ainsi

$$\frac{\operatorname{card}(S)}{\operatorname{card}(G_0)} = \frac{1}{4}.$$

(ii) On a $N = pqr = 1 + 2M$ avec $p = 1 + 2M_1$, $q = 1 + 2M_2$, $r = 1 + 2M_3$ et M_i divise M. On obtient alors la proportion $1/4$ également. Exemple : $N = 8911 = 7 \cdot 19 \cdot 67$ (on a $M_1 = 3$, $M_2 = 9$, $M_3 = 33$ et $M = 4455 = 3^4 \cdot 5 \cdot 11$).

En effet, on peut supposer $\alpha_1 = \dots = \alpha_k = 1$ sinon la proportion est $\leq 1/p_i$ qu'on peut supposer dans la pratique arbitrairement petit[(2)]. Si l'un des M_i est distinct de t_i alors $t_1 \dots t_k / M_1 \dots M_k \leq 1/3$. Ensuite

$$2^{-s_1 - \dots - s_k} \left(\frac{2^{k s_1} + 2^k - 2}{2^k - 1}\right) \leq 2^{-k s_1} \frac{2^k - 2}{2^k - 1} + \frac{1}{2^k - 1} \leq 2^{1-k},$$

(2) Néanmoins remarquons que pour $N = 3^2$, on trouve $|S|/|G_0| = 1/3$

donc la proportion est $\leq 1/8$ si $k \geq 4$ et $\leq 1/4$ si $k = 3$. Si $k = 2$ et si l'un des M_i est distinct de t_i alors la proportion est $\leq 1/6$. Si $k = 2$ et $M_1 = t_1$ (i.e. M_1 divise M) et $M_2 = t_2$ (i.e. M_2 divise M) on voit que $M_1 = M_2$ donc $s_1 < s_2$ (sinon $p_1 = p_2$). On écrit que la proportion est $\leq (2^{s_1-s_2} + 2^{1-s_1-s_2})/3$ donc $\leq 1/4$ si $s_1 = 1$ et $s_2 = 2$, mais $\leq 1/8$ si $s_1 = 1$ et $s_2 \geq 3$ et $\leq 3/16$ si $s_2 > s_1 \geq 2$. Revenons au cas $k = 3$ et les M_i sont égaux aux t_i; la proportion est égale à $2^{-s_1-s_2-s_3}(2^{3s_1} + 6)/7$. Si $s_1 \geq 2$ on trouve une proportion $\leq (1 + 3 \cdot 2^{-5})/7 = 5/32 < 1/6$, si $s_1 = 1$ et $s_3 \geq 2$, on trouve une proportion $\leq (1/2 + 3 \cdot 2^{-3})/7 = 1/8$. Dans le cas où $s_1 = s_2 = s_3 = 1$, on retrouve $1/4$.

D. Test de Primalité (II).

On expose dans ce paragraphe l'algorithme d'Agrawal-Kayal-Saxena [A-K-S] datant de juillet 2002, exposé dans leur article " PRIMES is in P *", donnant un test de primalité en temps polynomial.*

L'idée de départ est de pratiquer des tests dans $\mathbf{Z}[X]$. Par exemple on a facilement que si N est premier alors $(X-a)^N \equiv X^N - a \bmod N$ mais ce test a évidemment l'inconvénient majeur de nécessiter le calcul de N coefficients. Qu'à cela ne tienne!

Lemme. *Soit N premier et $h(X) \in \mathbf{Z}[X]$ un polynôme de degré r, alors*

$$(X-a)^N \equiv X^N - a \ \bmod(N, h(X)).$$

Rappelons que, dans un anneau, la notation $a \equiv b \bmod I$ signifie que $a - b$ appartient à l'idéal I et que, $(a_1, \ldots, a_m)$ désigne l'idéal engendré par $a_1, \ldots, a_m$; ainsi la congruence du lemme se traduit par : il existe $P, Q \in \mathbf{Z}[X]$ tels que $(X-a)^N - (X^N - a) = NP(X) + h(X)Q(X)$.

Il faut remarquer que, si r est $O((\log N)^k)$, alors le test reste polynomial. Le problème est de choisir les paires a, $h(X)$ de sorte qu'elles détectent la non primalité. La solution de [A-K-S] est de choisir $h(X) = X^r - 1$ avec r premier "très bien choisi", en particulier $r = O((\log N)^k)$, et de montrer qu'il suffit alors de tester les $a \in [1, L]$ avec $L = O(\sqrt{r} \log N)$ pour s'assurer que N est premier, ou éventuellement une puissance d'un nombre premier – ce qui n'est pas très gênant.

L'argument est essentiellement algébrique et combinatoire mais utilise un argument de répartition des nombres premiers, en fait une forme faible du théorème des nombres premiers qui dit que la somme des $\log p$ pour p premier inférieur à x est $\geq c_1 x$ avec une constante $c_1 > 0$. Résumons ce qu'on utilise dans un lemme.

Lemme. *Soit $Y > 1$, soit $N \geq 2$ entier, il existe un nombre premier r vérifiant :*
(i) L'ordre de N modulo r est au moins Y.
(ii) On a $r = O\left(Y^2 \log N\right)$.

Preuve. Posons $A := \prod_{1 \leq y \leq Y} (N^y - 1)$; soit r le plus petit nombre premier ne divisant pas A, alors pour $y \leq Y$ on a $N^y \not\equiv 1 \bmod r$ donc la condition (i) est satisfaite. Par ailleurs chaque $p < r$ divise A alors que $A \leq N^{Y(Y+1)/2}$ donc

$$c_1 r \leq \sum_{p<r} \log p \leq \log A \leq \frac{Y(Y+1)}{2} \log N.$$

On en tire bien $r = O\left(Y^2 \log N\right)$. □

Remarque. On peut ajouter que, comme l'ordre de N modulo r divise $r - 1$, on a forcément $r > Y$.

On utilisera également le lemme combinatoire élémentaire suivant

Lemme. *Le cardinal de l'ensemble des monômes de degré $\leq k$ est*

$$\operatorname{card}\{(m_1, \ldots, m_L) \mid m_i \geq 0 \text{ et } m_1 + \ldots + m_L \leq k\} = C^k_{L+k} = \binom{L+k}{k}.$$

De plus on a l'estimation

$$\binom{L+k}{k} \geq 2^{\min(L,k)}.$$

Preuve. La première formule est classique et peut, par exemple, se démontrer par récurrence (appeler $f(L,k)$ le cardinal en question, vérifier que $f(L,0)=1$, $f(1,k)=k+1$ et montrer que $f(L,k)=f(L,k-1)+f(L-1,k)$). Pour la minoration on observe que, si $k \leq L$ on a

$$\binom{L+k}{k} = \frac{(L+k)!}{k!L!} = \frac{(L+k)(L+k-1)\dots(L+2)(L+1)}{k(k-1)\dots 2.1} = \prod_{i=0}^{k-1}\left(\frac{L+k-i}{k-i}\right) \geq 2^k$$

et si $L \geq k$ on renverse le rôle de L et k. □

Remarque. On peut souvent améliorer l'inégalité ; par exemple, si $1 \leq k \leq L$ alors $\binom{L+k}{k} \geq 2^k(L+1)/2$ donc, si $L \geq 5$ on a $\binom{L+k}{k} \geq 2^{k+1}$.

Enonçons maintenant une version du théorème principal d'Agrawal-Kayal-Saxena.

Théorème. *Soit $N \geq 2$ et soit r un nombre premier tel que*
(i) Aucun nombre premier $\leq r$ ne divise N.
(ii) On a $\mathrm{ord}(N \bmod r) \geq (2\log N/\log 2)^2 + 1$.
(iii) Pour $1 \leq a \leq r-1$ on a

$$(X-a)^N \equiv X^N - a \bmod (N, X^r-1).$$

Alors N est une puissance d'un nombre premier.

Remarques. Pour démontrer le théorème on supposera seulement l'hypothèse (iii) vérifiée pour $1 \leq a \leq L$ et on verra qu'on peut prendre L plus petit que $r-1$. D'après le lemme analytique, on peut choisir $r = O\left((\log N)^5\right)$ tel que (ii) soit vérifié et on aura forcément $r \geq (2\log N/\log 2)^2+1$. Ainsi il est clair que le théorème implique que l'algorithme suivant est correct et polynomial.

ALGORITHME. [A-K-S] *On rentre N et l'algorithme sort "Premier" ou "Composé".*
(1) On vérifie si $N = a^b$ avec $b \geq 2$, si oui STOP
(2) On essaie $r = 2, 3\dots$ premiers. Si r divise N, STOP, *sinon on vérifie si r est premier avec $N^y - 1$ pour $y = 1, 2, \dots, Y$, avec $Y = [(2\log N/\log 2)^2] + 1$, si oui, on garde ce r et on passe à l'étape suivante, sinon, on cherche r plus grand*
(3) On essaie pour $a = 1, 2, 3, 4, \dots$ (on s'arrêtera à $r-1$) si $(X-a)^N \not\equiv X^N - a \bmod (N, X^r-1)$. Si oui, alors STOP, *sinon on passe à $a+1$*
(4) Si on rencontre un STOP, *on affiche "Composé", sinon on affiche "Premier".*

Discutons brièvement de la complexité (sans chercher à l'optimiser). On voit facilement que la partie la plus longue est la vérification (3) qui requiert $O(r\log N)$ multiplications dans l'anneau $\mathbf{Z}[X]/(N, X^r-1)$ dont chacune utilise au plus $O((r\log N)^2)$ opérations élémentaires donc en tout on obtient $O((r\log N)^3)$. Si on ajoute que $r = O\left((\log N)^5\right)$ on obtient une complexité au plus $O\left((\log N)^{18}\right)$.

Passons à la preuve du théorème. Notons p un diviseur premier de N. On notera $d_1 := \mathrm{ord}(N \bmod r)$, $d_2 = \mathrm{ord}(p \bmod r)$ et $d := \mathrm{ppcm}(d_1, d_2)$. Remarquons que d_1 (resp. d_2) est l'ordre du sous-groupe engendré par N (resp. par p) dans $(\mathbf{Z}/r\mathbf{Z})^*$ et que d est donc l'ordre du sous-groupe engendré par N et p dans $(\mathbf{Z}/r\mathbf{Z})^*$. Choisissons ensuite $h(X)$ un facteur irréductible de $\Phi_r(X) := (X^r-1)/(X-1)$ dans $\mathbf{F}_p[X]$. Remarquons, même si nous n'en aurons pas besoin qu'on peut montrer que $\deg(h) = d_2$. On va travailler dans le *corps* $K := \mathbf{F}_p[X]/(h(X))$; c'est un corps fini (isomorphe à $\mathbf{F}_{p^{d_2}}$) qu'on obtient donc en rajoutant à $\mathbf{F}_p$ une racine primitive r-ième de l'unité. Il est naturel de considérer G le sous-groupe de K^* engendré par les classes de $(X-a)$ pour $1 \leq a \leq L$. Le coeur de la preuve consiste à majorer et minorer le cardinal de G.

Lemme. *On a la minoration :*

$$\mathrm{card}(G) \geq \binom{L+d-1}{d-1} \geq 2^{\min(L,d-1)}.$$

Remarque. En reprenant la remarque suivant le lemme combinatoire, on obtient par exemple que, si $1 \leq d-1 \leq L$ alors $\mathrm{card}(G) \geq 2^d$ et si $L \leq d$ alors $\mathrm{card}(G) \geq 2^{L+1}$.

Preuve. Au vu du lemme combinatoire rappelé ci-dessus, il suffit de montrer que les classes des éléments

$$\prod_{1\leq a\leq L}(X-a)^{m_a}, \qquad \text{pour } m_a \geq 0 \text{ et} \qquad \sum_{a=1}^{L} m_a \leq d-1$$

sont toutes distinctes dans K. Tout d'abord les a sont distincts modulo p car sinon on aurait $p \leq L < r$, or on a supposé N non divisible par les premiers inférieurs à r donc $p > r$; ainsi nos polynômes restent distincts dans $\mathbf{F}_p[X]$. On fait ensuite la remarque-clef que si $P = \prod_{1\leq a\leq L}(X-a)^{m_a}$ alors on a d'une part $P(X)^N \equiv P(X^N) \bmod (N, X^r-1)$ mais également $P(X)^p \equiv P(X^p) \bmod p$ donc les deux congruences sont valables $\bmod(p, X^r-1)$. On obtient ainsi

$$\text{Pour } m = N^i p^j, \text{ on a } P(X)^m \equiv P(X^m) \bmod (p, X^r-1), \text{ ou encore } \bmod(p, h(X)).$$

Soient maintenant P, Q deux polynômes de la forme ci-dessus (vus dans $\mathbf{F}_p[X]$) et supposons qu'ils aient la même classe dans K, i.e. supposons $P \equiv Q \bmod (p, h(X))$. Soit x la classe de X ; c'est une racine primitive r-ième dans K et on a

$$(P-Q)(x^m) = 0, \qquad \text{pour} \quad m \in \langle N, p\rangle \subset (\mathbf{Z}/r\mathbf{Z})^*.$$

Mais l'on sait que N et p engendre un sous-groupe de cardinal d dans $(\mathbf{Z}/r\mathbf{Z})^*$ donc le polynôme $P-Q$ posède au moins d racines et comme $\deg(P-Q) \leq d-1$ on conclut bien que $P = Q$ (d'abord dans $\mathbf{F}_p[X]$ puis, si l'on veut, dans $\mathbf{Z}[X]$). □

Pour majorer $|G|$ nous noterons g un générateur de G (c'est un sous-groupe de K^* donc il est cyclique) et introduisons l'ensemble ci-dessous.

Définition. On pose $\mathcal{I} = \mathcal{I}_g := \{m \in \mathbf{N} \mid g(X)^m \equiv g(X^m) \bmod (X^r-1, p)\}$.

Les propriétés principales de $\mathcal{I}$ sont :

Lemme. *L'ensemble $\mathcal{I}$ vérifie les propriétés suivantes.*
(i) N et p sont dans $\mathcal{I}$.
(ii) $\mathcal{I}$ est multiplicatif i.e, si m_1 et $m_2 \in \mathcal{I}$ alors $m_1 m_2 \in \mathcal{I}$.
(iii) Si m_1 et $m_2 \in \mathcal{I}$ vérifient $m_1 \equiv m_2 \bmod r$ alors en fait $m_1 \equiv m_2 \bmod \mathrm{card}(G)$.

Preuve. La première propriété a déjà été prouvé. Pour (ii) écrivons

$$g(X)^{m_1 m_2} = (g(X)^{m_1})^{m_2} \equiv (g(X^{m_1}))^{m_2} \bmod (p, X^r-1)$$

et observons que, puisque $m_2 \in \mathcal{I}$, on a $g(Y)^{m_2} \equiv g(Y^{m_2}) \bmod (p, Y^r-1)$ et donc en substituant $Y = X^{m_1}$ on obtient

$$(g(X^{m_1}))^{m_2} = g(X^{m_1 m_2}) + pQ_1(X^{m_1}) + (X^{m_1 r}-1)Q_2(X^{m_1}) \equiv g(X^{m_1 m_2}) \bmod (p, X^r-1).$$

Pour démontrer (iii) supposons donc m_1 et $m_2 \in \mathcal{I}_g$ et $m_2 = m_1 + kr$ avec $k \geq 0$. On a donc

$$g(X)^{m_2} \equiv g(X^{m_2}) \bmod (X^r-1, p) \qquad \text{et donc} \qquad \bmod (h(X), p)$$

ainsi $g(X)^{m_1+kr} = g(X^{m_1+kr})$ dans K. Mais $X^{m_1+kr} \equiv X^{m_1} \bmod (X^r-1)$ et donc $\bmod(h(X))$. Ainsi on obtient dans K^* l'égalité :

$$g(X)^{m_1} g(X)^{kr} = g(X^{m_1}) = g(X)^{m_1}$$

la dernière égalité provenant de l'hypothèse $m_1 \in \mathcal{I}$. On en tire bien sûr $g(X)^{kr} = 1 \in K^*$ et donc $\mathrm{card}(G)$ divise $kr = m_2 - m_1$. □

Pour appliquer le lemme, on utilise que N, p et donc tous les $N^i p^j$ sont dans $\mathcal{I}$ et on se rappelle que ces éléments engendrent un sous-groupe de cardinal d dans $(\mathbf{Z}/r\mathbf{Z})^*$. Si l'on pose

$$E := \{(i,j) \in \mathbf{N} \times \mathbf{N} \mid 0 \leq i,j \leq [\sqrt{d}]\}$$

alors le cardinal de E est $([\sqrt{d}]+1)^2 > d$. Par le principe des tiroirs(*), il y a deux éléments $N^{i_1}p^{j_1}$ et $N^{i_2}p^{j_2}$ congrus modulo r avec (i_1,j_1) et (i_2,j_2) distincts dans E. Ces deux éléments $N^{i_1}p^{j_1}$ et $N^{i_2}p^{j_2}$ sont donc congrus modulo $\mathrm{card}(G)$. Supposons d'abord que $N^{i_1}p^{j_1} \neq N^{i_2}p^{j_2}$, ce qui entraîne donc que

$$\mathrm{card}(G) \leq |N^{i_1}p^{j_1} - N^{i_2}p^{j_2}| \leq N^{2\sqrt{d}}.$$

Si l'on combine cette majoration avec la minoration obtenue précédemment, on voit que :

$$\min(L+1,d)\log 2 \leq (2\sqrt{d})\log N.$$

Or, si l'on avait $L \geq d$, on en tirerait $\sqrt{d} \leq 2\log N/\log 2$ ou encore $d \leq (2\log N/\log 2)^2$. Mais cette inégalité est contradictoire car, par construction, $d \geq d_1$ et, par hypothèse, $d_1 > (2\log N/\log 2)^2$. Si $L < d$, on en tire $(L+1)\log 2 \leq (2\sqrt{d})\log N$ et, comme $d \leq r-1$ on aurait $(L+1)\log 2 \leq 2\sqrt{r-1}\log N$. Il suffit donc que $L \geq 2\sqrt{r-1}\log N/\log 2$ soit suffisamment grand pour conclure que $N^{i_1}p^{j_1} = N^{i_2}p^{j_2}$. Le choix $L = r-1$ convient(**) puisqu'alors l'inégalité voulue équivaut à $\sqrt{r-1} \geq 2\log N/\log 2$ ou encore $r \geq (2\log N/\log 2)^2 + 1$. On termine en remarquant que l'égalité $N^{i_1}p^{j_1} = N^{i_2}p^{j_2}$ entraîne aisément que $N = p^\alpha$. □

E. Factorisation.

On considère brièvement et, par la force des choses, de manière très insatisfaisante, le problème de la factorisation : ayant établi, par un test de primalité, qu'un entier N n'est pas premier, comment calculer sa factorisation ? Commençons par observer que le problème de la factorisation (complète) est essentiellement équivalent au problème de trouver un facteur, en effet l'itération de ce procédé donnera bien la factorisation complète.

La méthode *naïve* pour factoriser consiste à voir si 2 divise N, puis si 3 divise N, etc. Si $N = pq$ avec p et q sensiblement de la même taille, i.e. $p \sim q \sim \sqrt{N}$ on voit qu'on devra effectuer $O(\sqrt{N})$ divisions avant d'arriver à factoriser N. L'algorithme naïf est donc exponentiel.

Il existe des algorithmes plus performants, en fait un des meilleurs algorithmes connus (utilisant les "courbes elliptiques") a un nombre d'opérations estimé par $\exp(C\sqrt{\log p \log\log p})$ ou p est le plus petit facteur premier de N. Dans le cas où $N = pq$ avec $p \sim q \sim \sqrt{N}$ on obtient donc un algorithme avec une complexité en $\exp(C'(\log N)^\kappa)$ (avec $\kappa < 1$) qui croit beaucoup moins vite que N^κ mais beaucoup plus vite que $(\log N)^\kappa$. On dit qu'on a un algorithme sous-exponentiel. Un autre algorithme ("crible du corps de nombres") produit une complexité en $\exp\left(C(\log N)^{1/3}(\log\log N)^{2/3}\right)$. En pratique, on sait aujourd'hui (2004) factoriser en quelques heures un nombre entier de cent chiffres, on arrive à factoriser en plusieurs mois avec plusieurs ordinateurs un nombre de cent cinquante chiffres et on ne sait pas factoriser sur la durée d'une vie humaine un nombre RSA de disons trois cents chiffres. Une observation curieuse est que la complexité des divers algorithmes (prouvée, probabilistique ou heuristique) tend à prendre la forme d'une fonction

$$L(b,N) := \exp\left(C(\log N)^b(\log\log N)^{1-b}\right).$$

Le cas $b = 0$, c'est-à-dire $(\log N)^C$ correspond aux algorithmes polynomiaux, le cas $b = 1$ c'est-à-dire N^C correspond aux algorithmes exponentiels et le cas $0 < b < 1$ aux algorithmes sous-exponentiels ; les deux algorithmes cités ci-dessus ont une complexité estimée à $L(1/2,N)$ et $L(1/3,N)$.

(*) le principe des tiroirs dit que si l'on range $n+1$ chaussettes dans n tiroirs, un des tiroirs au moins contiendra deux chaussettes.

(**) Cependant on notera qu'on peut prendre $L = O(\sqrt{r}\log N)$, ce qui permet d'améliorer un peu l'estimation de la complexité.

Nous n'allons pas présenter les algorithmes les plus puissants qui recquièrent des outils dépassant le niveau de ce cours mais seulement un algorithme améliorant l'algorithme naïf et esquisser un autre algorithme plus puissant.

On note p un facteur premier de N.

Algorithme ρ de Pollard. On procède ainsi : on choisi a_0 entre 1 et N et on calcule la suite définie par $a_{i+1} = f(a_i)$ où $f(a) := a^2+1 \bmod N$. On choisit k "assez grand mais pas trop" et on calcule $\mathrm{pgcd}(a_{2k}-a_k, N)$ en espérant qu'il soit non trivial ; si c'est le cas on a trouvé une factorisation, sinon on réessaie avec k plus grand. Nous expliquons ci-dessous pourquoi, au moins statistiquement, il existe un k d'ordre $O(\sqrt{p})$ avec p divisant $\mathrm{pgcd}(a_{2k} - a_k, N)$, admettant cela, on voit que la complexité de l'algorithme est, avec une bonne probabilité, $O(\sqrt{p})$ donc en particulier $O(\sqrt[4]{N})$.

L'analyse de la complexité est basée sur l'hypothèse que la suite a_i modulo p est suffisamment "aléatoire", ce qui est assez bien vérifié par l'expérience et la pratique. Or la probabilité que r nombres modulo p tirés "au hasard" soient tous distincts est(*)

$$P_r = \left(1 - \frac{1}{p}\right)\left(1 - \frac{2}{p}\right)\ldots\left(1 - \frac{r-1}{p}\right) \leq \exp\left(-\frac{r(r-1)}{2p}\right)$$

Si l'on prend r de l'ordre de $\sqrt{p}$ disons $r \geq 2\sqrt{p}$, la probabilité que deux des nombres soient égaux (modulo p) sera $> 1/2$ donc on a une bonne chance d'avoir deux indices $i < j < r$ avec $a_i \equiv a_j \bmod p$. Vu la construction de la suite on aura $a_{i+m} \equiv a_{j+m} \bmod p$ pour tout $m \geq 0$ et en particulier en prenant $m = j - 2i$ et $k = j - i$ on aura $a_k \equiv a_{2k} \bmod p$.

Le deuxième algorithme, que nous ne ferons qu'esquisser, est basé sur la remarque que le nombre d'éléments $a \in (\mathbf{Z}/N\mathbf{Z})^*$ tels que $a^2 = 1$ est au moins égal à 4 si N possède au moins deux facteurs premiers distincts. Si on savait calculer une racine carrée dans $(\mathbf{Z}/N\mathbf{Z})^*$, disons $\mathcal{A}(x)$ par un algorithme rapide $\mathcal{A}$, alors on pourrait factoriser N ainsi : on prend a au hasard et on calcule $b = \mathcal{A}(a^2)$; on a alors bien sûr $a^2 \equiv b^2 \bmod N$ ou encore N divise $(a+b)(a-b)$, or il y a (au moins) une chance sur deux pour que $\pm b \bmod N$ ne soit pas la racine carrée calculée par $\mathcal{A}$ et dans ce cas le calcul de $\mathrm{pgcd}(N, a+b)$ ou de $\mathrm{pgcd}(N, a-b)$ fournira une factorisation. Malheureusement ou heureusement on ne connaît pas d'algorithme $\mathcal{A}$ rapide (on peut même penser qu'il n'en existe pas). Une extension de cette idée est la suivante : au lieu de chercher directement une égalité $a^2 \equiv b^2 \bmod N$ on cherche à la fabriquer. Pour cela on prend au hasard a proche de $\sqrt{N}$ et réduit a^2 modulo N (on a intérêt à prendre le représentant dans $[-N/2, N/2]$) et on regarde si on peut le factoriser avec des petits nombres premiers. Une fois qu'on a obtenu quelques a_i et b_j de ce type, on cherche une combinaison de ceux-ci qui fournissent une égalité du type $\prod_i a_i^2 \equiv \prod_j b_j^2 \bmod N$. Convenablement quantifié, cet algorithme a une complexité moyenne $L(1/2, N)$ – ce qui est déjà remarquable même si insuffisant pour factoriser de très grands nombres.

Exemples de précautions à prendre dans le choix de p et q pour la méthode RSA (nous ne donnons que quelques indications très élémentaires, la question est assez complexe et en fait largement ouverte).

1) Il faut que $|p-q|$ soit grand. En effet écrivons $q = p+\delta$ et supposons δ beaucoup plus petit que p. Comme $N = pq$ alors $\sqrt{N} = p\sqrt{1+\delta/p} \sim p + \delta$ et on pourra trouver p avec un algorithme naïf en $O(\delta)$ étapes!

2) Il faut que $p-1$ (resp. $q-1$) ne soit pas trop *friable* c'est à dire ne puisse pas se factoriser trop vite, par exemple ne soit pas le produit de nombres premiers petits ; en effet choisissons $C > 0$ et notons $p_1, \ldots, p_k$ les premiers inférieurs à C, l'ensemble $S := \{s = p_1^{m_1} \ldots p_k^{m_k} \mid s \leq N\}$ a un cardinal $O((\log N)^k)$ et on peut donc calculer $\mathrm{pgcd}(a^s - 1, N)$ pour quelques valeurs de a et $a \in S$ en temps polynomial. Si $p - 1 \in S$ (c'est-à-dire si $p-1$ n'a que des facteurs premiers $\leq C$) on a de très bonnes chances d'arriver à factoriser N ainsi.

3) Une contrainte moins évidente est qu'il faut que l'exposant "secret" e ne soit pas trop petit. Il est clair que si $e = O(\log N)$ par exemple alors en faisant $O(\log N)$ essais on trouvera e mais en fait on peut montrer qu'il faut éviter que $e \leq N^{1/4}$.

(*) Exemple. Si $n \geq 23$, la probabilité que, parmi n personnes, deux aient le même anniversaire est supérieure à $1/2$.

Ces remarques relativement triviales peuvent faire douter de la sécurité du système RSA. Un support théorique est néanmoins fourni par les considérations suivantes. Appelons P la classe des problèmes pour lesquels il existe un algorithme polynomial (par exemple le problème de décider si un nombre est premier est dans P d'après Agrawal-Kayal-Saxena). On peut définir une classe NP a priori plus vaste qui est celle des problèmes pour lesquels il existe un algorithme de vérification polynomial (par exemple le problème de la factorisation d'un nombre est clairement dans NP puisque, si l'on donne une factorisation, on peut la vérifier en temps polynomial). La sécurité du système RSA repose, du point de vue théorique, sur l'hypothèse que le problème de la factorisation n'est pas dans P. En fait c'est un cas spécial du grand problème de la théorie de la complexité[(**)] :

A-t-on $P \neq NP$?

[(**)] Le problème $P \neq NP$ est un des sept problèmes pour la solution desquels la fondation Clay offre un million de dollars.

Première partie (applications II) : Codes correcteurs.

A. Généralités sur les codes correcteurs.
B. Codes linéaires cycliques.

A. Généralités sur les codes correcteurs.

On donne un bref aperçu d'une autre application industrielle de l'algèbre et l'arithmétique : la construction de "codes correcteurs d'erreurs" permettant, dans une certaine mesure de restituer un message d'information si la transmission est légèrement défectueuse. Cette technique est à la base par exemple de la fabrication de lecteurs de compact disque, de la transmission d'images par les sondes spatiales, etc. Pour le lecteur que cette initiation laisserait sur sa faim, je recommande les derniers chapitres du livre de Demazure, Cours d'algèbre.

Pour transmettre des informations, on admettra qu'on utilise un alphabet fini $\mathcal{Q}$ comportant q symboles ou lettres et qu'on envoie des mots de longueur n fixée; un mot est donc un élément de $\mathcal{Q}^n$. On peut penser au language binaire, i.e. $\mathcal{Q} := \{0, 1\}$, ou aux codes génétiques, par exemple $\mathcal{Q} := \{A, C, G, U\}$ - les bases de l'ARN sont A pour adénine, C pour cytidine, G pour guanine et U pour uracile. Nous prendrons le plus souvent l'exemple de $\mathcal{Q} := \mathbf{F}_q$, ce qui a l'inconvénient de limiter les valeurs possibles de q mais l'avantage de donner une structure plus riche.

L'ensemble des mots $\mathcal{Q}^n$ peut être muni de la *distance de Hamming* définie comme suit. Si $x = (x_1, \ldots, x_n) \in \mathcal{Q}^n$ et $x' = (x'_1, \ldots, x'_n) \in \mathcal{Q}^n$ alors

$$d(x, x') := \text{card}\{i \in [1, n] \mid x_i \neq x'_i\}.$$

On vérifie aisément que c'est bien une distance.

Un *code* est un sous-ensemble $\mathcal{C} \subset \mathcal{Q}^n$ comportant au moins deux éléments $\mathcal{Q}^n$; on définit la *distance* d'un code comme

$$d(\mathcal{C}) := \min_{x \neq x' \in \mathcal{C}} d(x, x').$$

Le principe consiste, une fois choisi un code $\mathcal{C}$, à n'envoyer que des messages avec des mots appartenant à $\mathcal{C}$. Observons que l'on peut par ce procédé repérer jusqu'à $d(\mathcal{C}) - 1$ erreurs de transmission sur un mot; de plus si t erreurs ont été commises durant la transmission d'un mot et si $2t + 1 \leq d(\mathcal{C})$ on voit qu'il existe un seul mot de $\mathcal{C}$ situé à une distance $\leq t$ du mot reçu; en conclusion le code permet de corriger t erreurs, on dit qu'il est t-correcteur. Si l'on note $d = d(\mathcal{C})$ la distance du code et $t = t(\mathcal{C})$ le nombre maximal d'erreurs qui est systématiquement corrigé par le code, on voit facilement que la relation entre les deux est donnée par $t = \left[\frac{d-1}{2}\right]$ ou inversement $d = 2t + 1$ ou $2t + 2$. Sauf sur des exemples, nous laisserons de côté la question du *décodage*, c'est-à-dire essentiellement l'étude d'algorithmes permettant de trouver le mot du code situé à distance minimale d'un mot donné (noter que l'on ne peut en général garantir a priori l'unicité de ce mot que sous certaines conditions). Une des qualités demandées à un code est évidemment de corriger ou repérer le plus possible d'erreurs (on peut aussi demander que le décodage soit le plus simple possible), une qualité intuitivement évidente est d'occuper le moins de place possible; on peut formaliser cette idée en introduisant le *taux de correction* t/n, et le *taux d'information* qu'on définit comme la proportion $\log \text{card}(\mathcal{C}) / n \log q$. La théorie de l'information, développée par Shannon, indique que, si l'on accepte d'envoyer des messages de plus en plus long (i.e. de prendre n très grand), il existe des codes aussi sûrs que l'on veut, avec un taux d'information proche de 1 ; cependant le théorème de Shannon est un théorème d'existence, il ne dit pas comment construire les codes en question.

Nous allons en fait exclusivement nous occuper des *codes linéaires* pour lesquels l'alphabet est (en bijection avec) un corps fini $\mathbf{F}_q$, l'espace des mots est donc (en bijection avec) l'espace vectoriel $(\mathbf{F}_q)^n$ et $\mathcal{C}$ est un sous-espace vectoriel. Dans le cas $q = 2$ on parle de codes binaires, dans le cas $q = 3$ on parle de codes ternaires, etc.

Les paramètres les plus importants d'un code linéaire sont le cardinal de l'alphabet $q = \text{card}\, \mathcal{Q}$, sa *longueur* disons n, sa *dimension* disons $k := \dim \mathcal{C}$, sa distance $d(\mathcal{C})$, son taux de correction et son taux d'information k/n.

Remarque. Soit $\mathcal{C} \subset \mathbf{F}_q^n$ un code linéaire, définissons le *poids* d'un élément $w(x)$ comme le nombre de composantes non nulle de x (en anglais "*weight*"). On a visiblement

$$d(\mathcal{C}) = \min_{0 \neq x \in \mathcal{C}} d(0,x) = \min_{0 \neq x \in \mathcal{C}} w(x).$$

Exemples.

1) L'exemple le plus basique de code est l'utilisation d'un *bit de parité* : pour transmettre un mot $x = (x_1, \ldots, x_{n-1}) \in (\mathbf{F}_2)^{n-1}$ on envoie $\bar{x} = (x_1, \ldots, x_{n-1}, x_1 + \ldots + x_{n-1}) \in (\mathbf{F}_2)^n$. Pour voir si le message reçu $x' = (x_1, \ldots, x_n)$ est correct on vérifie si $x_n = x_1 + \ldots + x_{n-1}$. Ce code est de longueur n, de dimension $n-1$ et permet de repérer une erreur, mais pas de la corriger, sa distance est 2.

2) Code de Hamming ; prenons l'ensemble des mots de sept chiffres binaires $q = 2$, $n = 7$, et $\mathcal{C}$ le code ayant pour base

$$e_0 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad e_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad e_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

Le principe du codage est simple : pour transmettre un message $m = (m_0, m_1, m_2, m_3)$, on transmet $x = m_0 e_0 + m_1 e_1 + m_2 e_2 + m_3 e_3$. Expliquons sur cet exemple simple le décodage *sous l'hypothèse qu'une erreur au plus a été commise*. Après avoir reçu le message $x = (x_0, \ldots, x_6)$, on vérifie si $L(x) = 0$ avec

$$L(x) = (x_0 + x_3 + x_5 + x_6, x_1 + x_3 + x_4 + x_6, x_2 + x_4 + x_5 + x_6).$$

Si $L(x) = 0$, le message est correct, si $L(x) = (1,0,0)$ il faut corriger x_0, si $L(x) = (0,1,0)$ il faut corriger x_1, si $L(x) = (1,0,1)$ il faut corriger x_5, si $L(x) = (1,1,1)$ il faut corriger x_6; on a alors $m = (x_0, x_1, x_5, x_6)$.

Notons $T(x_1, \ldots, x_7) := (x_7, x_1, \ldots, x_6)$ le "décalage", de sorte que $T(e_0) = e_1$, $T(e_1) = e_2$, $T(e_2) = e_3$ et $T(e_3) = e_0 + e_1 + e_2$, ainsi $T(\mathcal{C}) = \mathcal{C}$ (on dira que $\mathcal{C}$ est cyclique). Il est facile de voir que chaque vecteur non nul de $\mathcal{C}$ possède au moins trois coordonnées non nulles et d'en déduire $d(\mathcal{C}) = 3$. Ainsi ce code est 1-correcteur et repère deux erreurs.

Illustration amusante. Le code précédent suggère qu'il est possible de retrouver un élément de $\mathbf{F}_2^4$ (soit disons un entier entre 0 et 15) à partir d'un élément de $\mathbf{F}_2^7$ (soit disons sept informations oui/non) si au plus une erreur est commise (soit au plus une des informations est fausse). En voici une version, avec les sept questions suivantes, pour trouver un entier N entre 0 et 15 :

1) L'entier N est-il ≥ 8 ?
2) L'entier N est-il dans $\{4,5,6,7,12,13,14,15\}$?
3) L'entier N est-il dans $\{2,3,6,7,10,11,14,15\}$?
4) L'entier N est-il impair ?
5) L'entier N est-il dans $\{1,2,4,7,9,10,12,15\}$?
6) L'entier N est-il dans $\{1,2,5,6,8,11,12,15\}$?
7) L'entier N est-il dans $\{1,3,4,6,8,10,13,15\}$?

On laisse en exercice la justification de l'algorithme suivant. On note $m = (m_1, \ldots, m_7)$ les réponses ($m_i = 1$ si la i-ème réponse est oui, $m_i = 0$ sinon) et on calcule $a_1 = m_4 + m_5 + m_6 + m_7$, $a_2 = m_2 + m_3 + m_6 + m_7$ et $a_3 = m_1 + m_3 + m_5 + m_7$. Si $a_1 = a_2 = a_3 = 0$, on conclut qu'il n'y a pas eu d'erreur, sinon on change la r-ème réponse m_r avec $r = \overline{a_1 a_2 a_3}$ (écriture en numération binaire) et le nombre cherché s'écrit alors $N = \overline{m_1 m_2 m_3 m_4}$.

Montrons maintenant comment caractériser et fabriquer des codes et comment déduire de nouveaux codes à partir de codes donnés, en utilisant l'algèbre linéaire élémentaire. On note n la longueur des codes et k leur dimension, sauf mention contraire.

Définition. Une matrice *génératrice* d'un code $\mathcal{C}$ est une matrice dont les lignes forment une base de $\mathcal{C}$. (C'est donc une matrice de rang k ayant k lignes et n colonnes). Une matrice *vérificatrice* d'un code $\mathcal{C}$ est une matrice dont les lignes forment une base de des formes linéaires s'annulant sur $\mathcal{C}$. (C'est donc une matrice de rang $n-k$ ayant $n-k$ lignes et n colonnes).

Remarques. Se donner une matrice génératrice équivaut bien sûr à se donner une base de l'espace vectoriel $\mathcal{C}$, se donner une matrice vérificatrice équivaut bien sûr à se donner une base des équations linéaires définissant $\mathcal{C}$ dans $\mathbf{F}_q^n$. Si A est une matrice génératrice, B une matrice vérificatrice, on voit aisément que $A\,{}^tB = 0$ ou encore $B\,{}^tA = 0$. Par ailleurs, on peut reconnaître la distance du code comme le plus petit nombre d tel qu'il existe d vecteurs colonnes de B distincts et liés.

Supposons un code $\mathcal{C}$ donné avec une matrice vérificatrice B et supposons que le code est 1-correcteur, montrons comment décoder un message reçu x' différant du message envoyé x en au plus une coordonnée. Tout d'abord si l'on note $\epsilon = x' - x$ l'erreur commise, on voit que $B(x') = B(\epsilon)$. Calculons donc $B(x')$, si ce dernier est nul, alors aucune erreur n'a été commise; sinon on calcule les images des vecteurs e_i de la base canonique $f_i = B(e_i)$. Si une seule erreur a été commise on trouve un unique i tel que $B(x')$ soit proportionnel à f_i, disons $B(x') = a_i f_i$, et alors $\epsilon = a_i e_i$ et $x = x' - a_i e_i$.

Soit $\mathcal{C}$ un code de longueur n sur le corps $\mathbf{F} = \mathbf{F}_q$.

(i) Code raccourci. Soit $d(\mathcal{C}) \leq \ell \leq n$, on pose $\mathcal{C}^{(\ell)} := \left\{x \in \mathbf{F}_q^\ell \mid (x; 0, \ldots, 0) \in \mathcal{C}\right\}$. C'est un code de longueur ℓ et on voit aisément que $d\left(\mathcal{C}^{(\ell)}\right) \geq d(\mathcal{C})$.

(ii) Code étendu. On fabrique l'analogue du "bit de parité" en construisant $\bar{\mathcal{C}} := \{(x_1, \ldots, x_{n+1}) \in \mathbf{F}_q^{n+1} \mid (x_1, \ldots, x_n) \in \mathcal{C} \quad \text{et} \quad x_1 + \ldots + x_n + x_{n+1} = 0\}$. On voit aisément que $d(\mathcal{C}) \leq d\left(\bar{\mathcal{C}}\right) \leq d(\mathcal{C}) + 1$. Une variante est le *sous-code pair* défini comme $\mathcal{C}' = \{x \in \mathcal{C} \mid x_1 + \ldots + x_n = 0\}$. On a $d(\mathcal{C}) \leq d(\mathcal{C}')$.

(iii) Code dual. On définit un "produit scalaire" $\langle x, y\rangle := x_1y_1 + \ldots + x_ny_n$ et on pose $\mathcal{C}^* := \{x' \in \mathbf{F}_q^n \mid \forall x \in \mathcal{C}, \quad \langle x, x'\rangle = 0\}$. On a $\dim \mathcal{C}^* = n - \dim \mathcal{C}$. Une catégorie intéressante de codes binaires est celle des codes *auto-duaux* i.e; tels que $\mathcal{C}^* = \mathcal{C}$; de tels codes sont de dimension $n/2$ et le poids d'un élément est pair car $w(x) \equiv \langle x, x\rangle \bmod 2$.

A titre d'exercice on pourra chercher comment construire une matrice vérificatrice (ou génératrice) de chacun des codes construits à partir d'une matrice vérificatrice (ou génératrice) du code de départ.

Lemme. *Soit $\mathcal{C}$ un code de dimension k et de longueur n sur $\mathbf{F}_q$, on a les inégalités :*

(i) $d(\mathcal{C}) \leq n + 1 - k$.

(ii) Si $\mathcal{C}$ est t-correcteur $1 + C_n^1(q-1) + C_n^2(q-1)^2 + \ldots + C_n^t(q-1)^t \leq q^{n-k}$.

Preuve. (i) Les vecteurs de la forme $(x_1, \ldots, x_{n+1-k}, 0, \ldots, 0)$ forment un sous-espace vectoriel $\mathcal{D}$ de $(\mathbf{F}_q)^n$, comme $\dim \mathcal{D} + \dim \mathcal{C} = n+1$, on voit que $\mathcal{D} \cap \mathcal{C} \neq \{0\}$, d'où l'existence d'un vecteur non nul de $\mathcal{D}$ de poids $\leq n + 1 - k$. Pour (ii) observons qu'on a toujours $\operatorname{card}(B(x,t)) = 1 + C_n^1(q-1) + C_n^2(q-1)^2 + \ldots + C_n^t(q-1)^t$; si le code est t-correcteur, les boules $B(x,t)$ de centre $x \in \mathcal{C}$ sont disjointes et donc

$$\operatorname{card}\left(\cup_{x\in\mathcal{C}} B(x,t)\right) = q^k \operatorname{card}(B(0,t)) \leq q^n.$$

□

Définition. Un code tel que $d(\mathcal{C}) = n+1-k$ sera dit MDS *maximal distance separable* (traduction française proposée : "*Codes de distance de séparation maximale*"). Un code t-correcteur tel que $\mathcal{C} = \cup_{x\in\mathcal{C}} B(x,t)$ (union forcément disjointe) est dit t-correcteur *parfait*.

Le code de Hamming de longueur 7 étudié en exemple est 1-correcteur parfait puisque dans ce cas on vérifie que $\operatorname{card} B(x,1) = 1 + 7 = 8$ et $8 \operatorname{card} \mathcal{C} = 2^7$. On peut observer que ce code n'est pas MDS (en effet $d(\mathcal{C}) = 3 < 4 = n - k + 1$).

B. Codes linéaires cycliques.

On décrit explicitement une classe intéressante de codes, qui comprend notamment des codes classiques comme ceux de Hamming, Reed-Solomon et Golay et nous amènera à l'étude des polynômes cyclotomiques.

Définition. Un code linéaire cyclique est un code $\mathcal{C}$ linéaire de longueur n, stable par la permutation $T(a_1, \ldots, a_n) = (a_n, a_1, \ldots, a_{n-1})$.

On peut donner une caractérisation algébrique agréable des codes cycliques en introduisant l'isomorphisme naturel d'espaces vectoriels $\mathbf{F}_q^n \cong \mathbf{F}_q[X]_n \cong \mathbf{F}_q[X]/Q\mathbf{F}_q[X]$, où $\mathbf{F}_q[X]_n$ désigne les polynômes de degré $< n$ et où Q est un polynôme de degré n. Comme l'endomorphisme T a pour polynôme caractéristique (ou minimal) $Q = X^n - 1$ on choisit donc cette valeur. Notons donc $\psi : \mathbf{F}_q^n \to \mathbf{F}_q[X]_n \cong \mathbf{F}_q[X]/(X^n - 1)$ défini par $\psi(a_0, a_1, \ldots, a_{n-1}) \mapsto a_0 + a_1 X + \ldots + a_{n-1}X^{n-1} \bmod(X^n - 1)$, on voit immédiatement que

$$\psi \circ T(a_0, a_1, \ldots, a_{n-1}) = X(a_0 + a_1 X + \ldots + a_{n-1}X^{n-1}) \bmod(X^n - 1).$$

Ainsi un sous-espace vectoriel $\mathcal{C} \subset \mathbf{F}_q^n$ est stable par T si et seulement son image par ψ est stable par multiplication par X. Observons maintenant qu'un $\mathbf{F}_q$-sous-espace vectoriel $\mathbf{F}_q[X]/(X^n - 1)$ stable par multiplication par X n'est autre chose qu'un *idéal* de $\mathbf{F}_q[X]/(X^n - 1)$. Enfin les idéaux de $\mathbf{F}_q[X]/(X^n - 1)$ correspondent aux idéaux de $\mathbf{F}_q[X]$ contenant le polynôme $X^n - 1$ et donc de la forme $P\mathbf{F}_q[X]$ avec P divisant $X^n - 1$. On peut résumer cette discussion par l'énoncé suivant :

Théorème. *Soit $K := \mathbf{F}_q$ et un code $\mathcal{C}$ de longueur n, identifions K^n et $K[X]/(X^n - 1)$ en associant $(a_1, a_2, \ldots, a_n) \mapsto a_1 + a_2 X + \ldots + a_n X^{n-1}$, on obtient des bijections entre les objets suivants :*
(i) Un code cyclique de longueur n ;
(ii) Un idéal de $K[X]/(X^n - 1)$;
(iii) Un polynôme unitaire divisant $X^n - 1$ dans $K[X]$;
les bijections sont données par : d'une part, à P divisant $X^n - 1$ on associe l'idéal $\mathcal{C}$ de $K[X]/(X^n - 1)$ engendré par sa classe modulo X^n et d'autre part un idéal de $K[X]/(X^n - 1)$ est aussi un sous-espace vectoriel et correspond donc à un sous-espace vectoriel $\mathcal{C}$ de K^n; de plus $\dim \mathcal{C} = n - \deg(P)$.

On aboutit ainsi au problème suivant : comment se décompose le polynôme $X^n - 1$ dans $\mathbf{F}_q[X]$?

Bien sûr il vaut mieux commencer par la décomposition dans $\mathbf{Z}[X]$ (ou $\mathbf{Q}[X]$) ; celle-ci est fournie par les *polynômes cyclotomiques*. Pour les définir nous noterons $\mu_n = \{\zeta \in \mathbf{C} \mid \zeta^n = 1\}$ le groupe des racines n-ièmes de l'unité et μ_n^* le sous-ensemble des racines n-ièmes primitives de l'unité, ainsi $\operatorname{card} \mu_n = n$ et $\operatorname{card} \mu_n^* = \phi(n)$.

Nous aurons besoin du corollaire du lemme de Gauss suivant :

Lemme. *Soit $\alpha \in \mathbf{C}$ racine d'un polynôme unitaire à coefficients entiers, alors le polynôme unitaire minimal de α est à coefficients entiers.*

Preuve. Soit P, a priori dans $\mathbf{Q}[X]$, le polynôme minimal de α et Q unitaire à coefficients entiers tel que $Q(\alpha) = 0$ alors $Q = PR$ avec R dans $\mathbf{Q}[X]$. Le lemme de Gauss nous dit qu'il existe $\lambda \in \mathbf{Q}^*$ tel que $R' = \lambda R$ et $P' = \lambda^{-1} P$ soient à coefficients entiers. En observant que $Q = P'R'$ on voit que le coefficient dominant de P' est inversible et donc $P = \pm P'$ est bien à coefficients entiers. □

Définition. Le n-ième polynôme cyclotomique, noté Φ_n est le polynôme

$$\Phi_n(X) := \prod_{\zeta \in \mu_n^*} (X - \zeta).$$

Ces polynômes sont a priori à coefficients complexes mais en fait à coefficients entiers et fournissent la décomposition en facteurs irréductibles de $X^n - 1$ comme le montre le théorème suivant.

Théorème. *Les polynômes Φ_n ont les propriétés suivantes.*
(i) $\Phi_n \in \mathbf{Z}[X]$ et $\deg \Phi_n = \phi(n)$.
(ii) $X^n - 1 = \prod_{d \mid n} \Phi_n(X)$.

(ii) Les polynômes Φ_n sont irréductibles dans $\mathbf{Z}[X]$ (ou $\mathbf{Q}[X]$).

Avec la définition donnée $\Phi_n \in \mathbf{C}[X]$, la formule (ii) est claire ainsi que le fait que $\deg(\Phi_n) = \phi(n)$; cependant il est moins évident qu'en fait $\Phi_n \in \mathbf{Z}[X]$ et que Φ_n est irréductible dans $\mathbf{Q}[X]$ (ou $\mathbf{Z}[X]$). Commençons par voir que les coefficients de Φ_n sont entiers. Il est clair que $\Phi_1(X) = X - 1 \in \mathbf{Z}[X]$. On peut alors démontrer ce que l'on veut par induction sur n en utilisant la formule (ii). En effet le polynôme $B := \prod_{d \mid n, d \neq n} \Phi_d(X)$ est unitaire et, par hypothèse de récurrence, à coefficients entiers; on peut donc effectuer dans $\mathbf{Z}[X]$ la division euclidienne $X^n = BQ + R$. La formule (ii) garantit alors que $R = 0$ et $Q = \Phi_n$. Montrons maintenant que Φ_n est irréductible dans $\mathbf{Z}[X]$. Soit ζ une racine primitive n-ème de l'unité et P son polynôme minimal sur $\mathbf{Q}$, on veut montrer que $P = \Phi_n$. Observons d'abord que $P \in \mathbf{Z}[X]$. Choisissons ensuite p un nombre premier ne divisant pas n alors ζ^p est encore une racine primitive n-ème de l'unité. Soit Q son polynôme minimal qui est également dans $\mathbf{Z}[X]$. Si P et Q étaient distincts, le produit PQ diviserait Φ_n. Mais comme $Q(\zeta^p) = 0$ on voit que ζ est racine de $Q(X^p)$ et donc $Q(X^p) = P(X)R(X)$ pour un certain $R \in \mathbf{Z}[X]$. En réduisant les coefficients modulo p on obtient:

$$\bar{Q}(X^p) = \bar{Q}(X)^p = \bar{P}(X)\bar{R}(X).$$

ou encore $\bar{P}(X)$ divise $\bar{Q}(X)^p$ dans $(\mathbf{Z}/p\mathbf{Z})[X]$ mais les facteurs de $X^n - 1$ et donc de $\bar{P}(X)$ sont simples dans $(\mathbf{Z}/p\mathbf{Z})[X]$ (la dérivée de $X^n - 1$ est nX^{n-1} et on a pris soin de choisir p ne divisant pas n) donc en fait $\bar{P}(X)$ divise $\bar{Q}(X)$. Mais alors $\bar{P}(X)^2$ divise $\bar{\Phi}_n(X)$ dans $(\mathbf{Z}/p\mathbf{Z})[X]$, ce qui contredit le fait que les facteurs de $\bar{\Phi}_n(X)$ sont simples. En résumé on a prouvé que, pour p premier ne divisant pas n, le polynôme minimal de ζ annulait ζ^p. On en tire aisément que, si m est premier avec n alors $P(\zeta^m) = 0$. Ainsi $\deg(P) \geq \phi(n)$ et comme P divise Φ_n, on a donc $P = \Phi_n$ et ce dernier est donc irréductible. □

Exercice. 1) Montrer les relations suivantes (on pourra comparer les degrés et les racines de chaque côté) :

$$\Phi_n(X^p) = \begin{cases} \Phi_{np}(X) & \text{si } p \text{ divise } n \\ \Phi_{np}(X)\Phi_n(X) & \text{si } p \text{ ne divise pas } n \end{cases}$$

2) Montrer que $\Phi_{p^r} = X^{p^{r-1}(p-1)} + X^{p^{r-1}(p-2)} + \ldots + X^{p^{r-1}} + 1$.

Comme Φ_n est à coefficients entiers, on peut réduire ses coefficients modulo p et le voir comme un polynôme dans $\mathbf{F}_p[X]$ (ou dans $\mathbf{F}_q[X]$ avec $q = p^f$).

Théorème. *La décomposition en facteurs irréductibles du polynôme $\Phi_n \in \mathbf{F}_q[X]$ (avec $q = p^f$) s'écrit ainsi :*
(i) Si $n = p^s m$ avec $p \not| m$, on a $\Phi_n(X) = \Phi_m(X)^{p^s - p^{s-1}}$.
(ii) Si $\operatorname{pgcd}(n, q) = 1$, notons r l'ordre de $q \bmod n$ dans $(\mathbf{Z}/n\mathbf{Z})^*$, alors Φ_n se décompose en le produit de $\phi(n)/r$ facteurs irréductibles de degré r.

Supposons d'abord que $n = p^r m$. Alors en utilisant le petit théorème de Fermat et les formules de l'exercice précédent on obtient $\Phi_m(X)^p \equiv \Phi_m(X^p) = \Phi_{mp}(X)\Phi_m(X)$ donc $\Phi_{mp}(X) \equiv \Phi_m(X)^{p-1}$ et ensuite

$$\Phi_{mp^r}(X) = \Phi_{mp}\left(X^{p^{r-1}}\right) \equiv \Phi_{mp}(X)^{p^{r-1}} \equiv \Phi_m(X)^{p^{r-1}(p-1)}$$

d'où la première formule. Supposons désormais p premier avec n. Soit β une racine primitive n-ème dans une extension de $\mathbf{F}_q$. Tout facteur de Φ_n s'écrit $Q = \prod_{i \in I}(X - \beta^i)$ avec $I \subset (\mathbf{Z}/n\mathbf{Z})^*$. Le polynôme Q est à coefficients dans $\mathbf{F}_q$ si et seulement si

$$Q(X)^q = Q(X^q) \qquad (*)$$

En effet $\left(\sum_j a_j X^j\right)^q = \sum_j (a_j)^q X^{qj}$ et $a \in \mathbf{F}_q$ si et seulement si $a^q = a$. Ainsi le polynôme Q est à coefficient dans $\mathbf{F}_q$ si et seulement si

$$\prod_{i \in I}(X^q - \beta^{iq}) = \prod_{i \in I}(X - \beta^i)^q = \prod_{i \in I}(X^q - \beta^i).$$

On voit ainsi que Q est à coefficient dans $\mathbf{F}_q$ si et seulement si I est une partie stable par multiplication par q (dans $(\mathbf{Z}/n\mathbf{Z})^*$). Les plus petites parties stables sont clairement du type $I := \{i, iq, iq^2, \ldots, iq^{r-1}\}$. Ainsi les facteurs irréductibles de $\Phi_n(X)$ dans $\mathbf{F}_q[X]$ sont de la forme

$$Q = \prod_{s=0}^{r-1} (X - \beta^{iq^s})$$

et en particulier ont tous degré r. □

Exemples. 1) Prenons $n = 11$ et $q = 3$; on voit que l'ordre de $3 \bmod 11$ est égal à 5. Ainsi $X^{11} - 1 = (X-1)\Phi_{11}(X)$ dans $\mathbf{Z}[X]$ et $\Phi_{11} = P_1 P_2 \in \mathbf{F}_3[X]$ avec $\deg(P_i) = 5$. On pourra vérifier que

$$X^{11} - 1 = (X-1)(X^5 - X^3 + X^2 - X - 1)(X^5 + X^4 - X^3 + X^2 - 1) \in \mathbf{F}_3[X].$$

2) Prenons $n = 23$ et $q = 2$; on voit que l'ordre de $2 \bmod 23$ est égal à 11. Ainsi $X^{23} - 1 = (X-1)\Phi_{23}(X)$ dans $\mathbf{Z}[X]$ et $\Phi_{23} = P_1 P_2 \in \mathbf{F}_2[X]$ avec $\deg(P_i) = 11$. On pourra vérifier que

$$X^{23} - 1 = (X-1)(X^{11} + X^{10} + X^6 + X^5 + X^4 + X^2 + 1)(X^{11} + X^9 + X^7 + X^6 + X^5 + X + 1) \in \mathbf{F}_2[X].$$

3) Prenons $n = 15$ et $q = 2$; ainsi $X^{15} - 1 = (X-1)\Phi_3(X)\Phi_7(X)\Phi_{15}(X)$ dans $\mathbf{Z}[X]$ avec $\Phi_{15} = X^8 - X^7 + X^5 - X^4 + X^3 - X + 1$. L'ordre de $2 \bmod 3$ est égal à 2, l'ordre de $2 \bmod 5$ est égal à 4 et l'ordre de $2 \bmod 15$ est égal à 4. Donc $\Phi_3 = X^2 + X + 1$ et $\Phi_5 = X^4 + X^3 + X^2 + X + 1$ sont irréductibles dans $\mathbf{F}_2[X]$ et $\Phi_{15} = P_1 P_2 \in \mathbf{F}_2[X]$ avec $\deg(P_i) = 4$. On pourra vérifier que

$$X^{15} - 1 = (X-1)(X^2 + X + 1)(X^4 + X^3 + X^2 + X + 1)(X^4 + X^3 + 1)(X^4 + X + 1) \in \mathbf{F}_2[X].$$

Théorème. *Soit $\mathcal{C}$ un code linéaire cyclique de longueur n sur $\mathbf{F}_q$ associé à $I \subset (\mathbf{Z}/n\mathbf{Z})^*$, supposons qu'il existe i et s tels que $\{i+1, i+2, \ldots, i+s\} \subset I$ alors $d(\mathcal{C}) \geq s+1$.*

Preuve. Soit β une racine primitive n-ème dans une extension de $\mathbf{F}_q$. Soit Q un polynôme modulo $X^n - 1$, il appartient à $\mathcal{C}$ si et seulement si $Q(\beta^{i+j}) = 0$ pour $j = 1, \ldots, s$; supposons que le poids w de Q (vu comme élément de $\mathbf{F}_q^n$) soit $\leq s$, ce qui signifie que $Q = a_1 X^{i_1} + \ldots + a_w X^{i_w}$ avec disons $0 \leq a_1 < a_2 < \ldots < a_w < n$. Il faut montrer qu'en fait Q doit être nul. Or on dispose des équations $a_1 \beta^{i_1(i+j)} + \ldots + a_w \beta^{i_w(i+j)} = 0$ pour $j = 1, \ldots, s$. Posons $a'_1 = a_1 \beta^{i_1 i}$, $\ldots$, $a'_w = a_w \beta^{i_w i}$, les équations se réécrivent :

$$\beta^{i_1 j} a'_1 + \ldots + \beta^{i_w j} a'_w = 0, \quad \text{pour } j = 1, \ldots, s.$$

Mais la matrice des $\beta^{i_r j}$ est extraite d'une matrice de Vandermonde avec $\beta^{i_r} \neq \beta^{i_{r'}}$, puisque β est d'ordre n, et possède donc rang $w = \min\{w, s\}$, ce qui impose donc $a'_1 = \ldots = a'_w = 0$ et donc $a_1 = \ldots = a_w = 0$. □

Remarque. La borne du théorème n'est pas, en général, optimale comme on pourra le constater plus loin sur l'exemple des codes de Golay.

Exemples de codes linéaires cycliques.

Un tel exemple s'obtient en choisissant q, n et une partie $I \subset (\mathbf{Z}/n\mathbf{Z})^*$ stable par multiplication par q. Pour être rigoureux, il faut préciser que le code que l'on construit dépend aussi de la racine n-ème primitive "β" que l'on choisit ; toutefois il n'est pas difficile de voir que les divers codes obtenus selon les choix de β sont tous isomorphes et nous omettrons donc β.

Codes de Hamming. Un premier choix intéressant de paramètres est $n = (q^r - 1)/(q-1)$ qui entraîne, bien sûr, que l'ordre de $q \bmod n$ est r. On pose alors $I := \{1, q, q^2, \ldots, q^{r-1}\}$, ce qui définit un code $\mathcal{C}$ de dimension $n - r$ (une fois choisi β racine n-ème primitive de l'unité). Vérifions directement que $d(\mathcal{C}) \geq 3$: en effet un polynôme de poids 2 s'écrit $f = aX^i + bX^j$ avec disons $0 \leq i < j \leq n-1$ et la condition qu'il s'annule en β^{q^ℓ} pour $0 \leq \ell \leq r-1$ s'écrit donc $a + b\beta^{(j-i)q^\ell} = 0$ et comme β est d'ordre n on voit que ceci est

impossible sauf si $a = b = 0$. Ainsi le code $\mathcal{C}$ est 1-correcteur et comme $\operatorname{card} B(x,1) = 1 + n(q-1) = q^r$ on voit que $\mathcal{C}$ est parfait 1-correcteur et ainsi $d(\mathcal{C}) = 3$ ou 4 (on montre ci-dessous que la distance est 3, on peut en déduire que le code est MDS si et seulement si $r = 2$). Les codes de Hamming binaires s'obtiennent en spécialisant $q = 2$ et en choisissant $I := \{1, 2, 4, \ldots, 2^{r-1}\}$ et donc $k = n - r = 2^r - r - 1$. Comme $\{1, 2\} \subset I$ on retrouve que $d(\mathcal{C}) \geq 3$. Pour $r = 3$, $q = 2$, $n = 7$ on retrouve le code étudié comme premier exemple.

On peut donner une autre présentation des codes de Hamming en choisissant des représentants $e_1, e_2, \ldots, e_n$ des vecteurs non nuls de $\mathbf{F}_q^r$ à colinéarité près. On a $n = (q^r - 1)/(q - 1)$. Notons V la matrice dont les colonnes sont les vecteurs e_i et appelons $\ell_1, \ldots, \ell_r$ ses lignes. Alors V est la matrice vérificatrice d'un code $\mathcal{C}$, explicitement :

$$\mathcal{C} := \left\{ x \in \mathbf{F}_q^n \mid \langle \ell_i, x \rangle = 0, \text{ pour } 1 \leq i \leq r \right\}.$$

Ce code est isomorphe au code de Hamming construit précédemment ; comme deux des vecteurs e_i distincts ne sont jamais liés par construction mais qu'il existe bien sûr des triplets linéairement dépendants, on vérifie bien que $d(\mathcal{C}) = 3$.

Codes de Reed-Solomon. Ces codes correspondent au choix $n = q - 1$ avec, le plus souvent, $q = 2^f$; soit α un générateur de $\mathbf{F}_q^*$, une fois choisi k on pose

$$g(X) := \prod_{i=1}^{q-1-k} \left(X - \alpha^i\right).$$

On a bien sûr $k = \dim \mathcal{C}$ et, comme $I = \{1, 2, 3, \ldots, q - 1 - k\}$ on a $d(\mathcal{C}) \geq q - k$. Mais on sait que, pour tout code linéaire $d(\mathcal{C}) \leq n + 1 - k$ donc $d(\mathcal{C}) = q - k$ et le code ainsi construit est MDS. Supposons maintenant que $q = 2^f$, on peut voir $\mathcal{C}$ comme un code *binaire* $\mathcal{C}'$ de paramètre $n' = (2^f - 1)f$, $k' = kf$ et distance $d(\mathcal{C}') \geq 2^f - k$. Une particularité de ce code est de corriger de large bouffées d'erreurs : si t vérifie $2t+1 \leq d(\mathcal{C}) = q-k$, le code corrige t éléments de $\mathbf{F}_{2^f}$ donc tf erreurs binaires si celles-ci se répartissent par paquets! Cette particularité explique pourquoi ce type de code est utilisé dans la technologie des compacts disques.

Code ternaire de Golay. On a $3^5 - 1 = 11.23$. On choisit $q = 3$, $n = 11$ et la partie de $(\mathbf{Z}/11\mathbf{Z})^*$ engendrée par 3, c'est-à-dire $I := \{1, 3, 4, 5, 9\}$; le code noté $\mathcal{G}_{11}$ est donc de dimension 6. On peut remarquer (ce dont nous ne nous servirons pas) que $I = \mathbf{F}_{11}^{*2}$. D'après le théorème sur la distance d'un code cyclique, on voit que $d(\mathcal{G}_{11}) \geq 4$ et, en considérant la factorisation de Φ_{11} dans $\mathbf{F}_3[X]$ on voit que $\mathcal{G}_{11}$ contient un polynôme de poids 5 donc $d(\mathcal{G}_{11}) \leq 5$. Un calcul exhaustif (ou voir l'exercice ci-dessous) permet d'établir qu'en fait $d(\mathcal{G}_{11}) = 5$. Ainsi $\mathcal{G}_{11}$ est 2-correcteur et comme $\operatorname{card} B(x,2) = 1 + 2C_{11}^1 + 2^2 C_{11}^2 = 3^5$ on voit que le code $\mathcal{G}_{11}$ est 2-correcteur parfait (noter qu'il n'est pas MDS).

Code binaire de Golay. On a $2^{11} - 1 = 23.89$ (c'est le plus petit nombre de la forme $2^p - 1$ qui n'est pas premier). Choisissons donc $q = 2$, $n = 23$ et I la partie de $(\mathbf{Z}/23\mathbf{Z})^*$ engendrée par 2, c'est-à-dire $I := \{1, 2, 3, 4, 6, 8, 9, 12, 13, 16, 18\}$ et notons $\mathcal{G}_{23}$ le code associé. On peut remarquer que $I = \mathbf{F}_{23}^{*2}$. D'après le théorème sur la distance d'un code cyclique, on voit que $d(\mathcal{G}_{23}) \geq 5$ et, en considérant la factorisation de Φ_{23} dans $\mathbf{F}_2[X]$ on voit que $\mathcal{G}_{23}$ contient un polynôme de poids 7 donc $d(\mathcal{G}_{23}) \leq 7$. Un calcul exhaustif (voir aussi l'exercice proposé ci-dessous) permet d'établir qu'en fait $d(\mathcal{G}_{23}) = 7$. Ainsi $\mathcal{G}_{23}$ est 3-correcteur et comme $\operatorname{card} B(x,3) = 1 + C_{23}^1 + C_{23}^2 + C_{23}^3 = 2^{11}$ on voit que le code $\mathcal{G}_{23}$ est 3-correcteur parfait (noter qu'il n'est pas MDS).

Remarque. On peut montrer que, si l'on exclut les codes triviaux (i.e. de dimension 1, $n - 1$ ou n), les seuls codes t-correcteurs parfaits sont ceux que nous avons construits : les codes 1-correcteurs de Hamming et les deux codes de Golay binaire et ternaire.

Exercices. (Où l'on montre que $d(\mathcal{G}_{11}) = 5$ et $d(\mathcal{G}_{23}) = 7$ et utilise la notion de code autodual).

A) Soit $\mathcal{C}$ un code cyclique de longueur n engendré par le polynôme $g = g(X)$ de degré d, soit $\mathcal{C}'$ son sous-code pair et $\mathcal{C}^*$ son code dual.

1) Montrer que $\mathcal{C}' = \mathcal{C}$ si et seulement si $g(1) = 0$; si $g(1) \neq 0$, vérifier que $\mathcal{C}'$ est cyclique engendré par le polynôme $(X - 1)g(X)$.

2) Montrer que $\mathcal{C}^*$ est cyclique engendré par le polynôme $h^*(X) = X^{n-d}h(1/X)$ où $g(X)h(X) = X^n - 1$. [Indication : on pourra montrer que si $\deg(f) \leq n-d-1$ et $\deg(e) \leq d-1$ alors $\langle fg, eh^* \rangle$ est égal au coefficient de X^{n-1} dans le produit $f(X)g(X)e^*(X)h(X) = f(X)e^*(X)(X^n-1)$ et est donc nul.]

B) On suppose $\mathcal{C} \subset \mathcal{C}^*$ (i.e. pour tout $x, y \in \mathcal{C}$ on a $\langle x, y \rangle = 0$).

1) Si $q = 2$, montrer que pour tout $x, y \in \mathcal{C}$ on a $w(x+y) \equiv w(x) + w(y) \bmod 4$.

2) Si $q = 3$, montrer que pour tout $x, y \in \mathcal{C}$ on a $w(x+y) \equiv w(x) + w(y) \bmod 3$.

C) On introduit $\mathcal{D}$ le sous-code de $\mathcal{G}_{11}$ formé des vecteurs dont la somme des coordonnées est nulle (le sous-code "pair").

1) Montrer que si $g(X)$ est le polynôme générateur de $\mathcal{G}_{11}$, le code $\mathcal{D}$ est cyclique de générateur $(X-1)g(X)$.

2) Montrer que $\mathcal{D} \subset \mathcal{D}^*$ (i.e. pour tout $x, y \in \mathcal{D}$ on a $\langle x, y \rangle = 0$). En déduire que pour tout $x \in \mathcal{D}$ on $w(x) \equiv 0 \bmod 3$.

3) On note $\bar{\mathcal{D}}$ et $\bar{\mathcal{G}}_{11}$ les codes étendus, on pose $e_{11} = (1, \ldots, 1) \in \mathbf{F}_3^{11}$ et $e_{12} = (1, \ldots, 1) \in \mathbf{F}_3^{12}$. Montrer que $e_{11} \in \mathcal{G}_{11}$, $e_{12} \in \bar{\mathcal{G}}_{11}$ et donc $\bar{\mathcal{G}}_{11} = \bar{\mathcal{D}} \oplus \mathbf{F}_3 e_{12}$.

4) Montrer que $\bar{\mathcal{G}}_{11}$ est autodual, en déduire que, pour tout $x, y \in \bar{\mathcal{G}}_{11}$, on a $w(x+y) \equiv w(x) + w(y) \bmod 3$ et donc $d(\bar{\mathcal{G}}_{11}) \equiv 0 \bmod 3$.

5) Sachant que $4 \leq d(\mathcal{G}_{11}) \leq 5$ et $d(C) \leq d(\bar{C}) \leq d(C) + 1$, conclure que $d(\mathcal{G}_{11}) = 5$ et $d\left(\bar{\mathcal{G}}_{11}\right) = 6$.

D) Soit p premier impair tel que $\left(\frac{2}{p}\right) = 1$, $S := \mathbf{F}_p^{*2}$ et soit $\mathcal{C}$ un code binaire de longueur p correspondant à la partie S (qui est stable par multiplication par 2 par hypothèse); soit $\bar{\mathcal{C}}$ le code étendu de longueur $p+1$.

1) Si $g = g(X)$ est un générateur de $\mathcal{C}$ et si $g^*(X) = X^{(p-1)/2}g(1/X)$ est son polynôme réciproque, montrer que si $p \equiv 1 \bmod 8$ alors $g(X) = g^*(X)$ alors que si $p \equiv -1 \bmod 8$ on a $\Phi_p(X) = g(X)g^*(X)$.

2) On suppose désormais $p \equiv -1 \bmod 8$. Montrer que $\bar{\mathcal{C}}$ est auto-dual (i.e. $\bar{\mathcal{C}} = \bar{\mathcal{C}}^*$ ou encore pour tout $\bar{x}, \bar{y} \in \bar{\mathcal{C}}$ on a $\langle \bar{x}, \bar{y} \rangle = 0$).

3) Soit $x = \sum_{i \in I} X^i$, $y = \sum_{i \in J} X^i$, vérifier que $\langle x, y \rangle = |I \cap J| \bmod 2$, que $w(x+y) = |I| + |J| - 2|I \cap J|$ et conclure que, si $\langle x, y \rangle = 0$ alors $w(x+y) \equiv w(x) + w(y) \bmod 4$.

4) Déduire de la question précédente que si $\mathcal{D}$ est un code auto-dual engendré par des éléments de poids multiples de 4, alors tout élément de $\mathcal{D}$ est de poids multiple de 4 et en particulier $d(\mathcal{D}) \equiv 0 \bmod 4$.

5) Appliquer ce qui précède au cas $p = 23$, observer que, si g est le générateur de $\mathcal{C} = \mathcal{G}_{23}$ on a $w(g) = 7$ donc $w(\bar{g}) = 8$ et conclure que $d(\bar{C}) \equiv 0 \bmod 4$. Sachant que $5 \leq d(\mathcal{G}_{23}) \leq 7$ et $d(C) \leq d(\bar{C}) \leq d(C) + 1$, conclure que $d(\mathcal{G}_{23}) = 7$ et $d\left(\bar{\mathcal{G}}_{23}\right) = 8$.

Deuxième partie : Algèbre et équations diophantiennes.

A. Sommes de carrés.
B. Equation de Fermat (n=3 et 4).
C. Equation de Pell-Fermat $x^2 - dy^2 = 1$.
D. Anneaux d'entiers algébriques.

Nous abordons dans cette partie des problèmes classiques de théorie des nombres comme la résolution en nombres entiers d'équations polynomiales. Les exemples traités couvrent 1) la décomposition d'un entier n en somme de deux trois ou quatre carrés, autrement dit la recherche de solutions de l'équation $n = x_1^2 + x_2^2 + \ldots + x_k^2$, 2) le "théorème de Fermat" (démontré par Andrew Wiles en 1995) : les seules solutions de l'équation $x^n + y^n = z^n$ pour $n \geq 3$ sont les solutions triviales (celles avec $xyz = 0$) 3) les solutions de l'équation de Pell-Fermat $x^2 - dy^2 = 1$ (ou plus généralement $x^2 - dy^2 = n$). L'étude des congruences - thème de la première partie - permet de donner des conditions nécessaires à l'existence de solutions, les méthodes introduites dans ce chapitre sont d'une part l'utilisation d'anneaux plus généraux que $\mathbf{Z}$, d'autre part le recours à des énoncés d'approximation rationnelle. On étudiera ainsi d'une part des anneaux comme $\mathbf{Z}[i]$, $\mathbf{Z}[\exp(2\pi i/n)]$, $\mathbf{Z}[\sqrt{d}]$ et même un anneau non commutatif, l'anneau des quaternions d'Hurwitz (sous-anneau du corps des quaternions défini par Hamilton), et d'autre part la rapidité avec laquelle un réel peut être approché par des rationnels. On termine avec un aperçu sur les propriétés générales de ces anneaux en introduisant qulques notions supplémentaires d'algèbre : entiers algébriques, anneaux de Dedekind.

A. Sommes de carrés.

Nous cherchons à quelle condition, un entier $n \in \mathbf{N}$ peut s'écrire comme somme de carrés. Etudions tout d'abord les contraintes provenant de congruences.

On sait que $x^2 \equiv 0$ ou 1 modulo 4, donc un nombre $n = 4n' + 3$ ne peut pas être somme de deux carrés. Plus précisément on peut remarquer que si $p \equiv 3 \bmod 4$ et si p divise $n = x^2 + y^2$ alors p doit diviser y ; en effet sinon on pourrait écrire $(xy^{-1})^2 \equiv -1 \bmod p$ et en déduire que -1 est un carré modulo p , ce qui est faux. Comme p divise y, il divise aussi x et on conclut que $x = px'$, $y = py'$ et $n = p^2 n'$. En répétant le raisonnement on voit que :

Si $p \equiv 3 \bmod 4$ et si $n = p^{2a+1}m$ (avec m premier avec p) alors n n'est pas somme de deux carrés.

Observons que si x est pair alors $x^2 \equiv 0$ ou 4 modulo 8 alors que si x est impair $x^2 \equiv 1$ modulo 8. On voit donc que $x^2 + y^2 + z^2$ n'est jamais congru à 7 modulo 8. On peut légèrement raffiner ce raisonnement : si $n = 4n'$ et si $n = x^2 + y^2 + z^2$ alors on voit que x, y, z doivent être pairs donc $x = 2x'$, $y = 2y'$ et $z = 2z'$ avec $n' = x'^2 + y'^2 + z'^2$. En répétant le raisonnement on voit que :

Si n est de la forme $n = 4^a(8m + 7)$ alors n n'est pas somme de deux carrés.

Il est remarquable que les obstructions données par les congruences soient, dans ce cas, les seules.

Théorème. (théorème des 2 carrés) *Un entier $n \in \mathbf{N}$ peut s'écrire comme somme de deux carrés d'entiers si et seulement si chaque nombre premier p congru à 3 modulo 4 apparaît avec un exposant pair dans la décomposition en facteurs premiers de n.*

Théorème. (théorème des 3 carrés) *Un entier $n \in \mathbf{N}$ peut s'écrire comme somme de trois carrés d'entiers si et seulement si il n'est pas de la forme $n = 4^a(8m + 7)$.*

Théorème. (théorème des 4 carrés) *Soit $n \in \mathbf{N}$ alors il existe des entiers x, y, z, t tels que $n = x^2+y^2+z^2+t^2$.*

Nous ne prouverons pas le second théorème (voir par exemple de livre de Serre *Cours d'arithmétique* chez P.U.F.) ; pour prouver le premier théorème nous allons introduire l'anneau $\mathbf{Z}[i]$, pour prouver le troisième, nous allons introduire l'anneau des quaternions d'Hurwitz.

Au vu des théorèmes, on s'aperçoit que l'ensemble des sommes de deux carrés (resp. de quatre carrés) est stable par produit mais pas l'ensemble des sommes de trois carrés. En effet $18 = 2.3^2 = 4^2 + 1^2 + 1^2$

et $14 = 2.7 = 3^2 + 2^2 + 1^2$ mais $18.14 = 4.9.7$ n'est pas somme de trois carrés. On peut expliquer la multiplicativité de l'ensemble des sommes deux carrés (resp. de quatre carrés) par les formules :

$$(x^2 + y^2)(a^2 + b^2) = (ax - by)^2 + (ay + bx)^2$$

et $\quad (x^2 + y^2 + z^2 + t^2)(a^2 + b^2 + c^2 + d^2) =$

$$(ax - by - cz - dt)^2 + (ay + bx - ct + dz)^2 + (az + bt + cx - dy)^2 + (at - bz + cy + dx)^2.$$

Bien entendu l'origine de ces formules sera claire une fois donnée l'interprétation par $\mathbf{Z}[i]$ ou les quaternions.

Si on pose

$$\mathcal{C}_2 := \{n \in \mathbf{N} \mid \exists x, y \in \mathbf{N},\ n = x^2 + y^2\} \text{ et } \mathcal{C}_4 := \{n \in \mathbf{N} \mid \exists x, y, z, t \in \mathbf{N},\ n = x^2 + y^2 + z^2 + t^2\}$$

on voit qu'il suffit de montrer que tout nombre premier congru à 1 modulo 4 est dans $\mathcal{C}_2$ et que tout nombre premier est dans $\mathcal{C}_4$.

Nous allons construire l'exemple classique de corps non commutatif : le corps des quaternions découvert par Hamilton, et développer une application arithmétique : la preuve du théorème des quatre carrés.

La façon la plus concrète de construire le corps des quaternions est comme un espace vectoriel réel de dimension 4 muni d'une base $\mathbf{1}, I, J, K$ et d'une multiplication $\mathbf{R}$-bilinéaire définie par le fait que $\mathbf{1}$ est élément neutre et les formules

$$I^2 = J^2 = K^2 = -\mathbf{1}, \quad IJ = -JI = K, \quad JK = -KJ = I \quad \text{et} \quad KI = -IK = J \qquad (*)$$

Il faut alors vérifier "à la main" l'associativité : par exemple $(IJ)K = K^2 = -\mathbf{1}$ et $I(JK) = I^2 = -\mathbf{1}$. Pour s'épargner cette vérification on peut aussi définir $\mathbf{H}$ comme sous-algèbre des matrices 2×2 complexes ou 4×4 réelles (l'associativité est alors immédiate mais il faut vérifier que les matrices introduites vérifient les formules $(*)$). On peut ainsi définir

$$\mathbf{H} = \left\{ \begin{pmatrix} \alpha & -\beta \\ \bar{\beta} & \bar{\alpha} \end{pmatrix} \middle| \ \alpha, \beta \in \mathbf{C} \right\}$$

avec $\mathbf{1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $I = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$, $J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ et $K = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$ ou encore

$$\mathbf{H} = \left\{ \begin{pmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{pmatrix} \middle| \ a, b, c, d \in \mathbf{R} \right\}$$

avec

$$\mathbf{1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},\ I = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix},\ J = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix},\ K = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Remarque. Une fois construit $\mathbf{H}$, on peut remarquer que c'est une $\mathbf{R}$-algèbre engendrée par deux éléments i, j avec les relations $i^2 = j^2 = -\mathbf{1}$ et $ij = -ji$. En effet en posant $k := ij$ on en déduit la table de multiplication puisque $k^2 = ijij = -iijj = -\mathbf{1}$ et $ik = iij = -j = (ii)j = -iji = -ki$ etc. Le fait que $\mathbf{H}$ ne soit pas commutatif se lit déjà sur la table de multiplication.

On introduit le *conjugué* d'un quaternion $z = a\mathbf{1} + bI + cJ + dK$ comme $\bar{z} = a\mathbf{1} - bI - cJ - dK$ ainsi que sa *trace* $\mathrm{Tr}(z) = z + \bar{z}$ et sa *norme* $\mathrm{N}(z) = z\bar{z}$. On vérifie alors

Lemme. *Soient $z, w \in \mathbf{H}$, $\overline{z+w} = \bar{z} + \bar{w}$, $\overline{zw} = \bar{w} \cdot \bar{z}$ et si $z = a\mathbf{1} + bI + cJ + dK$, alors $\mathrm{N}(z) = z\bar{z} = \bar{z}z = (a^2+b^2+c^2+d^2)\mathbf{1}$ et $\mathrm{Tr}(z) = 2a\mathbf{1}$; de plus $\mathrm{Tr}(z+z') = \mathrm{Tr}(z) + \mathrm{Tr}(z')$, $\mathrm{N}(zz') = \mathrm{N}(z)\,\mathrm{N}(z')$ et z est racine du polynôme $X^2 - \mathrm{Tr}(z)X + \mathrm{N}(z) \in \mathbf{R}[X]$.*

Preuve. Des calculs directs (laissés au lecteur) permettent de vérifier ces formules. Noter que la conjugaison est un *anti-isomorphisme* de corps, i.e. qu'elle renverse l'ordre de la multiplication. □

On voit immédiatement comme corollaire que $\mathbf{H}$ est un corps puisque, si $z = a\mathbf{1} + bI + cJ + dK$ est un quaternion non nul, alors $\mathrm{N}(z) := a^2+b^2+c^2+d^2 \in \mathbf{R}^*$ et $z\bar{z}/\,\mathrm{N}(z) = \mathbf{1}$ donc $z^{-1} = \bar{z}/\,\mathrm{N}(z)$.

Introduisons maintenant l'anneau $\mathbf{Z}[i]$ et les deux anneaux

$$A_0 = \mathbf{Z}\mathbf{1} + \mathbf{Z}I + \mathbf{Z}J + \mathbf{Z}K \quad \text{ et } A = A_0 + \mathbf{Z}\left(\frac{1+I+J+K}{2}\right).$$

L'ensemble A est un sous-anneau de $\mathbf{H}$ car, si on note $\delta := (1+I+J+K)/2$, on a $\delta^2 = \delta - 1$. Il est clair que $\mathcal{C}_2 = \{\mathrm{N}(z) \mid z \in B\}$ et $\mathcal{C}_4 = \{\mathrm{N}(z) \mid z \in A_0\}$, en fait on a aussi $\mathcal{C}_4 = \{\mathrm{N}(z) \mid z \in A\}$ car si $x, y, z, t \in \mathbf{Z} + 1/2$ alors $\mathrm{N}(x\mathbf{1} + yI + zJ + tK) \in \mathbf{N}$ et on a le résultat élémentaire suivant :

Lemme. *Soit $\alpha = \frac{x\mathbf{1}+yI+zJ+tK}{2} \in A$ avec x, y, z, t entiers impair, alors il existe $\epsilon = \frac{\pm\mathbf{1}\pm I\pm J\pm K}{2}$ tel que $\epsilon\alpha$ soit dans A_0 et $\mathrm{N}(\alpha) = \mathrm{N}(\epsilon\alpha)$.*

Preuve. On écrit $x = 4x' + \epsilon_1$, $y = 4y' + \epsilon_2$, $z = 4z' + \epsilon_3$, $t = 4t' + \epsilon_4$, avec $\epsilon_i = \pm 1$. Si on pose $\epsilon := \frac{\epsilon_1\mathbf{1}-\epsilon_2 I-\epsilon_3 J-\epsilon_4 K}{2}$ on a bien $\mathrm{N}(\epsilon) = 1$ donc $\mathrm{N}(\alpha\epsilon) = \mathrm{N}(\alpha)$ et

$$\alpha\epsilon = 4\left(\frac{x'\mathbf{1} + y'I + z'J + t'K}{2}\right)\epsilon + \mathrm{N}(\epsilon) = (x'\mathbf{1} + y'I + z'J + t'K)\,(2\epsilon) + 1 \in A_0.$$

□

Le lemme suivant sera également utile.

Lemme. *Dans les anneaux $\mathbf{Z}[i]$, A_0 et A, un élément est inversible si et seulement si sa norme vaut 1.*

Preuve. Si α est inversible, alors $1 = \mathrm{N}(\alpha\alpha^{-1}) = \mathrm{N}(\alpha)\,\mathrm{N}(\alpha^{-1})$ donc $\mathrm{N}(\alpha) = 1$. Inversement si $\mathrm{N}(\alpha) = 1$ alors $\alpha\bar{\alpha} = 1$ or les anneaux considérés sont stables par conjugaison donc $\bar{\alpha}$ est un élément de l'anneau et α est bien inversible. □

Enfin, la norme étant multiplicative, il suffit de montrer que tout nombre premier p (resp. tout nombre premier $\equiv 1 \bmod 4$) est une norme de quaternion d'Hurwitz (resp. une norme de $\mathbf{Z}[i]$). Comme $2 = 1^2 + 1^2$ il suffit d'ailleurs de le faire pour p premier impair. Pour cela nous allons montrer d'abord que $\mathbf{Z}[i]$ est principal et A est *principal* à gauche (ou à droite).

Proposition. *L'anneau $\mathbf{Z}[i]$ est euclidien donc principal. L'anneau A est euclidien à gauche, donc principal à gauche (idem à droite).*

Preuve. Notons B l'anneau A ou $\mathbf{Z}[i]$, l'énoncé signifie que pour $\alpha \in B$ et $\beta \in B \setminus \{0\}$, il existe $q, r \in B$ tel que $\alpha = q\beta + r$ avec $\mathrm{N}(r) < \mathrm{N}(\beta)$ (lorsque l'anneau est A, il faut faire attention au sens des multiplications). Supposons ceci démontré, on en tire aussitôt que $\mathbf{Z}[i]$ est principal, en fait la "même" démonstration montre que A est principal (à gauche). Soit donc I un idéal à gauche non nul de A (i.e. $A.I \subset I$), il contient un élément $\beta \neq 0$ de norme minimale et on a clairement $A\beta \subset I$. Inversement, soit $\alpha \in I$, écrivons $\alpha = q\beta + r$ avec $\mathrm{N}(r) < \mathrm{N}(\beta)$, on a alors $r = \alpha - q\beta \in I$ donc r est nul et on a bien $I = A\beta$. Montrons maintenant que A et $\mathbf{Z}[i]$ sont euclidiens. La preuve est basée sur le lemme élémentaire suivant dont la preuve est laissée au lecteur.

Lemme. *Soit $x \in \mathbf{R}$, il existe $m \in \mathbf{Z}$ tel que $|x - m| \leq 1/2$ et il existe $n \in \mathbf{Z}$ tel que $|x - n/2| \leq 1/4$.*

Soit donc $\alpha \in \mathbf{Z}[i]$ et $\beta \in \mathbf{Z}[i] \setminus \{0\}$, alors $\alpha/\beta = x + iy \in \mathbf{Q}[i]$ et il existe $m, n \in \mathbf{Z}$ tels que $|x - m| \leq 1/2$ et $|y - n| \leq 1/2$ donc

$$\mathrm{N}\left((x+iy) - (m+in)\right) = (x-m)^2 + (y-n)^2 \leq \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

d'où, si l'on note $q := m + ni$ l'inégalité cherchée

$$\mathrm{N}(\alpha - q\beta) \leq \frac{\mathrm{N}(\beta)}{2} < \mathrm{N}(\beta).$$

Soit maintenant $\alpha \in A$ et $\beta \in A \setminus \{0\}$, alors $\alpha\beta^{-1} = x + yI + zJ + tK \in \mathbf{H}$ et il existe $m \in \mathbf{Z}$ tel que $|x - m/2| \leq 1/4$. On choisit alors $q = (m + nI + hJ + \ell K)/2$ avec m, n, h, ℓ entiers de même parité (de sorte que $q \in A$) et tel que $|y - n/2|$, $|z - h/2|$ et $|t - \ell/2|$ soient $\leq 1/2$. On obtient alors

$$\mathrm{N}(\alpha\beta^{-1} - q) = \left(x - \frac{m}{2}\right)^2 + \left(y - \frac{n}{2}\right)^2 + \left(z - \frac{h}{2}\right)^2 + \left(t - \frac{\ell}{2}\right)^2 \leq \frac{1}{16} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4} < 1$$

d'où l'inégalité cherchée

$$\mathrm{N}(\alpha - q\beta) < \mathrm{N}(\beta).$$

□

On peut maintenant achever la preuve des deux théorèmes. On rappelle que les notions d'élément *premier* (i.e. l'idéal engendré par l'élément est premier ou encore si l'élément divise un produit, il doit diviser un des facteurs) et *irréductible* (i.e. si l'élément s'écrit comme un produit de deux facteurs l'un de ceux-ci est inversible) sont en général distinctes dans un anneau mais coïncide dans le cas d'un anneau principal ou factoriel.

(Somme de deux carrés). L'anneau $\mathbf{Z}[i]$ est principal donc factoriel et on voit facilement que $\mathbf{Z}[i]^* = \{\pm 1, \pm i\}$. Soit donc $p \equiv 1 \bmod 4$. On sait qu'il existe $a \in \mathbf{Z}$ tel que $a^2 \equiv -1 \bmod p$. Ainsi on a une égalité de la forme $(a+i)(a-i) = pm$. Il est clair que p ne divise ni $(a+i)$ ni $(a-i)$ donc n'est pas premier (dans $\mathbf{Z}[i]$) donc n'est pas irréductible. On a ainsi une décomposition $p = \alpha\beta$ avec α, β non inversible. Donc $\mathrm{N}(\alpha\beta) = \mathrm{N}(p) = p^2$ et $\mathrm{N}(\alpha) = \mathrm{N}(\beta) = p$, ce qui démontre le théorème des deux carrés.

(Somme des quatres carrés). Il suffit de montrer que, si p est un nombre premier impair, il est la norme d'un élément de A. Le nombre de carrés dans $\mathbf{Z}/p\mathbf{Z}$ est $(p+1)/2$ donc le polynôme $-1 - X^2$ prend au moins une fois pour valeur un carré ; en d'autre termes, il existe $a, b \in \mathbf{Z}$ tels que $a^2 + b^2 + 1 \in p\mathbf{Z}$. On en tire que $(\mathbf{1} + aI + bJ)(\mathbf{1} - aI - bJ) \in pA$. Considérons donc l'idéal (à gauche) $\mathcal{I}$ engendré par p et $\mathbf{1} + aI + bJ$, on a $\mathcal{I} = A\beta$ puisque A est principal (à gauche) et d'autre part des inclusions $pA = Ap \subset \mathcal{I} \subset A$. Ainsi $p = \alpha\beta$; vérifions que β et α ne sont pas inversibles ou encore que les inclusions d'idéaux ci-dessus sont strictes. Si α était inversible on aurait p divise $1 + aI + bJ$ ou encore $(1 + aI + bJ) = p(x + yI + zJ + tK)/2$ donc $px = 2$, ce qui est impossible (p est premier impair). Si β était inversible, on aurait $\mathcal{I} = A$ donc $1 = q(1 + aI + bJ) + q'p$ et, en multipliant (à droite) par $(1 - aI - bJ)$ on obtiendrait $(1 - aI - bJ) = q''p$, ce qui est également absurde. On conclut donc que $\mathrm{N}(p) = \mathrm{N}(\alpha)\,\mathrm{N}(\beta) = p^2$ avec $\mathrm{N}(\alpha)$ et $\mathrm{N}(\beta)$ différents de 1 donc égaux à p. □

Exercice. Montrer que

$$\mathbf{Z}[i]^* = \{\pm 1, \pm i\}, \qquad A_0^* = \{\pm 1, \pm I, \pm J, \pm K\} \qquad \text{et} \qquad A^* = A_0^* \cup \left\{\frac{\pm 1 \pm I \pm J \pm K}{2}\right\}$$

(A_0^* et A^* sont les groupes quaternioniques d'ordre 8 et 24 respectivement). Le groupe A^* est-il isomorphe à $\mathcal{S}_4$? Montrer que A_0 n'est pas principal (à gauche). En déduire également qu'un élément de norme égale à un nombre premier est irréductible.

Supplément : une preuve via la géométrie des nombres.

On peut aussi démontrer le théorème des deux carrés à l'aide du théorème de Minkowski ci-dessous.

Rappelons d'abord l'énoncé de topologie suivant :

Proposition. *Un sous-groupe G discret dans $\mathbf{R}^n$ possède une $\mathbf{Z}$-base formé de r vecteurs $\mathbf{R}$-linéairement indépendants (avec $r \leq n$) ; en particulier $G \cong \mathbf{Z}^r$.*

Remarque. Lorsque $r = n$ on dit que G est un *réseau* dans $\mathbf{R}^n$; il revient au même de demander que G soit discret et que $\mathbf{R}^n/G$ soit compact. On alors définit le volume ou déterminant d'un réseau G comme la valeur absolue du déterminant d'une base de G.

Théorème. (Minkowski) *Soit $K \subset \mathbf{R}^n$ un compact convexe et symétrique (i.e. $x \in K$ entraîne $-x \in K$). Supposons $\mathrm{vol}(K) \geq 2^n$ alors il existe x non nul dans $K \cap \mathbf{Z}^n$.*

Plus généralement si Λ est un réseau et $\mathrm{vol}(K) \geq 2^n \det(\Lambda)$ alors il existe x non nul dans $K \cap L$.

Remarque. L'énoncé est optimal puisque par exemple le cube défini par $\max_i |x_i| \leq 1 - \epsilon$ est compact, convexe, symétrique et a pour volume $2^n(1-\epsilon)^n$.

Preuve. La deuxième affirmation se déduit de la première en faisant un changement de variables linéaire ramenant le réseau Λ sur $\mathbf{Z}^n$; ici $\det(\Lambda)$ désigne la valeur absolue du déterminant d'une matrice dont les colonnes sont formées par les vecteurs d'une base de Λ.

Posons $C := [0,1[^n$. Soit $T \subset \mathbf{R}^n$ et supposons que $(T+\lambda) \cap (T+\mu) = \emptyset$ pour $\lambda \neq \mu \in \mathbf{Z}^n$. On a $T = \cup_{\lambda \in \mathbf{Z}^n} (T \cap (C+\lambda))$ donc

$$\mathrm{vol}(T) = \sum_{\lambda \in \mathbf{Z}^n} \mathrm{vol}\,(T \cap (C+\lambda)) = \sum_{\lambda \in \mathbf{Z}^n} \mathrm{vol}\,((T-\lambda) \cap C) = \mathrm{vol}\,((\cup_{\lambda \in \mathbf{Z}^n}(T-\lambda)) \cap C) \leq \mathrm{vol}(C) = 1.$$

Inversement, si $\mathrm{vol}(T) > 1$ on en déduit qu'il existe $x \in T \cap (T+\lambda)$ avec $0 \neq \lambda \in \mathbf{Z}^n$ ou encore qu'il existe un tel λ dans $T - T$. Revenons à la preuve du théorème ; introduisons $T := \frac{1}{2}K = \left\{\frac{x}{2} \mid x \in K\right\}$. On a alors $K = T - T$ et $\mathrm{vol}(T) = 2^{-n}\mathrm{vol}(K)$ donc on conclut que, si $\mathrm{vol}(K) > 2^n$ alors $K \cap (\mathbf{Z}^n \setminus \{0\}) \neq \emptyset$. Si on suppose de plus K compact, on voit que la condition $\mathrm{vol}(K) \geq 2^n$ suffit en introduisant $K_\epsilon = (1+\epsilon)K$ qui contient un élément non nul du réseau $\mathbf{Z}^n$, or la compacité donne $K = \cap_{\epsilon > 0} K_\epsilon$ et chaque K_ϵ contient un point non nul de $\mathbf{Z}^n$ qui est discret donc il existe $x \in \mathbf{Z} \setminus \{0\}$ contenu dans tous les K_ϵ et donc dans K. □

Exercice. Modifier la démonstration pour obtenir $\mathrm{card}\,(K \cap \Lambda) \geq 2^n \mathrm{vol}(K)/\det(\Lambda)$.

Application. *Un nombre premier $p \equiv 1 \bmod 4$ est somme de deux carrés.*

Choisissons, $a \in \mathbf{Z}$ tel que $a^2 + 1 \equiv 0 \bmod p$ et soit le réseau

$$\Lambda := \{(x,y) \in \mathbf{Z}^2 \mid y \equiv ax \bmod p\}.$$

On a $\det(\Lambda) = p$, et $\mathrm{vol}(B(0,r)) = \pi r^2$, donc, dès que $\pi r^2 \geq 4p$, il existe un vecteur non nul dans $B(0,r) \cap \Lambda$. En particulier il existe un tel vecteur $(x,y) \in \Lambda$ avec

$$0 < x^2 + y^2 \leq r^2 = 4p/\pi < 2p.$$

Or on a $x^2 + y^2 \equiv (1+a^2)x^2 \equiv 0 \bmod p$ donc en fait $x^2 + y^2 = p$. □

Remarque. On peut également démontrer le théorème des quatre carrés en prouvant l'identité de Jacobi :

$$\mathrm{card}\{(x,y,z,t) \in \mathbf{Z}^4 \mid x^2+y^2+z^2+t^2 = n\} = 8 \sum_{\substack{d|n \\ 4 \not| d}} d = \begin{cases} 8\sum_{d|n} d & \text{si } n \text{ impair} \\ 24 \sum_{d|n, 2 \not| d} d & \text{si } n \text{ pair} \end{cases}$$

(où $n > 0$). En effet le membre de droite de l'égalité est clairement strictement positif. On pourra vérifier cette formule sur les premières valeurs : $r_4(1) = 8$, $r_4(2) = 24$, $r_4(3) = 32$, $r_4(4) = 24$, $r_4(5) = 48$, $r_4(6) = 96$, $r_4(7) = 64$, $r_4(8) = 48$, $r_4(9) = 104$, $r_4(10) = 144$. Cette formule peut s'écrire en terme de fonctions génératrices où l'on note $r_k(n)$ le nombre de représentations en somme de k carrés, i.e. $r_k(n) := \mathrm{card}\left\{(x_1, \ldots, x_k) \in \mathbf{Z}^k \mid x_1^2 + \ldots + x_k^2 = n\right\}$.

Définissons les séries suivantes (formelle ou convergente si $|q| < 1$) :

$$\Theta(q) = \sum_{n \in \mathbf{Z}} q^{n^2}$$

de sorte que

$$\Theta(q)^k = \left(\sum_{n \in \mathbf{Z}} q^{n^2} \right)^k = \sum_{n \in \mathbf{N}} r_k(n) q^n,$$

et également

$$Z(q) = \sum_{n \in \mathbf{N}^*} \sigma(n) q^n, \qquad \text{où } \sigma(n) = \textstyle\sum_{d|n} d$$

alors la formule de Jacobi s'écrit

$$\Theta(q)^4 = 1 + 8 \left(Z(q) - 4Z(q^4) \right).$$

B. Equation de Fermat (n=3 et 4).

Un des problèmes mathématiques les plus célèbres (dit "théorème de Fermat") a été résolu par Andrew Wiles en 1995 :

Théorème. *Soit $n \geq 3$, soient x, y, z entiers tels que $x^n + y^n = z^n$ alors $xyz = 0$.*

Bien sûr il "suffit" de le démontrer pour $n = 4$ et $n = p$ premier impair; d'autre part on peut supposer que x, y, z sont premiers entre eux deux à deux. Nous nous contenterons de le démontrer pour $n = 3$ et 4 en utilisant le principe de la descente infinie de Fermat. La démonstration proposée pour $n = 4$ reste dans $\mathbf{Z}$ mais celle que nous donnons pour $n = 3$ passe par $\mathbf{Z}[j]$ (avec $j = \exp(2\pi i/3)$). L'approche classique, due à Kummer, est basée sur la factorisation suivante : on introduit $\zeta = \exp(2\pi i/p)$ et on obtient alors dans l'anneau $\mathbf{Z}[\zeta]$:

$$x^p + y^p = (x + y)(x + \zeta y) \dots (x + \zeta^{p-1} y) = z^p.$$

Donnons quelques calculs dans l'anneau $\mathbf{Z}[\zeta]$ en posant $\lambda = 1 - \zeta$.

Lemme. *L'élément λ est premier et $\mathbf{Z}[\zeta]/\lambda\mathbf{Z}[\zeta] \cong \mathbf{F}_p$; on a la décomposition*

$$p = \prod_{k=1}^{p-1} (1 - \zeta^k) = \epsilon \lambda^{p-1}, \qquad \textit{avec } \epsilon \in \mathbf{Z}[\zeta]^*.$$

Les éléments $\eta_k := \sin(k\pi/p)/\sin(\pi/p)$ de même que $\epsilon_k := (1 - \zeta^k)/(1 - \zeta)$ sont des unités de $\mathbf{Z}[\zeta]$ pour $1 \leq k \leq p - 1$ et η_k/ϵ_k est une racine de l'unité.

Preuve. Partons de la factorisation (sur $\mathbf{C}$) du p-ième polynôme cyclotomique $\Phi_p(X) = X^{p-1} + X^{p-2} + \dots + X + 1 = \prod_{k=1}^{p-1}(X - \zeta^k)$. On en tire la formule $p = \Phi_p(1) = \prod_{k=1}^{p-1}(1 - \zeta^k)$. Ainsi λ divise p, ou encore $p \in \lambda\mathbf{Z}[\zeta]$ ainsi tout élément de $\mathbf{Z}[\zeta]$ est congru modulo λ à un entier compris entre 0 et $p-1$, ce qui montre bien que $\mathbf{Z}[\zeta]/\lambda\mathbf{Z}[\zeta] \cong \mathbf{F}_p$. Comme $1 - \zeta^k = (1 - \zeta)(1 + \dots + \zeta^{k-1})$ on voit que ϵ_k est bien dans $\mathbf{Z}[\zeta]$; le même raisonnement (en utilisant h l'inverse de k modulo p et $1 - \zeta = (1 - \zeta^k)(1 + \dots + \zeta^{k(h-1)})$) montre que ϵ_k^{-1} est aussi un entier et donc que $\epsilon_k \in \mathbf{Z}[\zeta]^*$. Enfin si k est impair on a :

$$\epsilon_k = \frac{1 - \zeta^k}{1 - \zeta} = \exp\left(\frac{\pi i (k-1)}{p} \right) \frac{\exp(\pi i k/p) - \exp(-\pi i k/p)}{\exp(\pi i/p) - \exp(-\pi i/p)} = \zeta^{\frac{k-1}{2}} \frac{\sin(\pi k/p)}{\sin(\pi/p)} = \zeta^{\frac{k-1}{2}} \eta_k$$

tandis que si k est pair $\epsilon_k = -\zeta^k \epsilon_{p-k}$. Enfin comme $1 - \zeta^k = \epsilon_k \lambda$, on a bien $p = \epsilon \lambda^{p-1}$ avec $\epsilon = \epsilon_1 \dots \epsilon_{p-1} \in \mathbf{Z}[\zeta]^*$. □

Remarque. On peut bien sûr écrire d'autres formules donnant des unités comme :

$$2\cos\left(\frac{2\pi}{p}\right) = \zeta + \zeta^{-1} = \zeta^{-1}(1+\zeta^2) = \zeta^{-1}\frac{1-\zeta^4}{1-\zeta^2} = \zeta^{-1}\epsilon_4\epsilon_2^{-1}.$$

Revenons à la méthode de Kummer pour l'équation de Fermat sous la forme factorisée

$$(x+y)(x+\zeta y)\dots(x+\zeta^{p-1}y) = z^p.$$

Soit $\delta \in \mathbf{Z}[\zeta]$ divisant deux des facteurs de l'équation ci-dessus, disons $x+\zeta^i y$ et $x+\zeta^j y$ alors il divise $(\zeta^i-\zeta^j)y$ et $(\zeta^i-\zeta^j)x$ donc $(\zeta^i-\zeta^j)$ et donc δ divise λ, donc $\delta = 1$ ou λ (à une unité près). Si z n'est pas divisible par p, les facteurs sont premiers entre eux et, *si l'on vérifie que* $\mathbf{Z}[\zeta]$ *est factoriel*, on en tire que :

$$\text{pour } i = 0,\dots,p-1, \qquad x+\zeta^i = u_i\alpha_i^p, \qquad \text{avec } u_i \text{ unité et } \alpha_i \in \mathbf{Z}[\zeta].$$

Si z est divisible par p on montre de même, toujours si $\mathbf{Z}[\zeta]$ est factoriel, des identités similaires avec des puissances de λ en plus. Cependant cette approche se heurte au fait que justement l'anneau $\mathbf{Z}[\zeta]$ n'est pas factoriel en général. En fait si $n = p$ premier, il n'est pas factoriel dès que $p \geq 23$. On cherche donc un substitut au lemme (dont la preuve est laissée en exercice) :

Lemme. *Soit A un anneau factoriel, si les éléments $a_1,\dots,a_m \in A$ sont premiers entre eux deux à deux et si $a_1\dots a_m = a^p$ alors, à une unité près, les a_i sont des puissances p-ièmes.*

D'abord décrivons les solutions de l'équation de Fermat lorsque $n = 2$.

Proposition. *Soient x,y,z des entiers (premiers entre eux) tels que $x^2+y^2=z^2$ alors (quitte à échanger x et y) il existe u,v entiers (premiers entre eux) tels que*

$$x = u^2-v^2, \qquad y = 2uv \qquad \text{et} \qquad z = u^2+v^2.$$

Preuve. Après avoir simplifié par leur pgcd, on peut supposer x,y,z premiers entre eux deux à deux. Remarquons qu'on a bien $(u^2-v^2)^2+(2uv)^2 = (u^2+v^2)^2$. Les considérations de parité montrent que z est impair et que x et y sont de parité opposée; nous supposerons donc x impair, y pair. Ecrivons $y^2 = z^2-x^2 = (z-x)(z+x)$. Observons que si d divise $z-x$ et $z+x$ alors il divise $2x$ et $2z$ donc 2 (puisque x et z sont premiers entre eux), ainsi $\text{pgcd}(z-x,z+x) = 2$. Les entiers $(z-x)/2$ et $(z+x)/2$ étant donc premiers entre eux et leur produit étant un carré, ils sont eux-mêmes des carrés et on a : $z-x = 2v^2$ et $z+x = 2u^2$ et $y = 2uv$ d'où $x = u^2-v^2$ et $z = u^2+v^2$ comme annoncé. □

Théorème. *L'équation $x^4+y^4 = z^2$ n'a pas de solutions entières avec $xyz \neq 0$. Par conséquent l'équation de Fermat avec $n = 4$ non plus.*

Le principe de la preuve est de montrer que, si l'équation a une solution (x,y,z) avec $xyz \neq 0$, alors elle possède une autre solution (x_1,y_1,z_1) avec $x_1y_1z_1 \neq 0$ et $|z_1| < |z|$. Ceci aboutirait à une contradiction car une suite décroissante d'entiers positifs est nécessairement constante à partir d'un certain rang.

Soit donc (x,y,z) solution. On peut supposer x,y,z premiers entre eux alors la proposition précédente montre que $x^2 = u^2-v^2$, $y^2 = 2uv$ et $z = u^2+v^2$, avec u,v premiers entre eux. On voit que u et v sont de parité différente et donc u impair, $v = 2w$ (sinon on aurait $x^2 \equiv -1 \bmod 4$). En considérant $y^2 = 4uw$ on voit que u et w sont des carrés, disons $u = z_1^2$ et $w = a^2$. Par ailleurs, en appliquant de nouveau la proposition précédente à $x^2+v^2 = u^2$ on obtient $x = b^2-c^2$, $v = 2bc$ et $u = b^2+c^2$ avec b et c premiers entre eux ; mais en se rappelant que $v = 2w = 2a^2$, on voit que b et c sont des carrés, disons $b = x_1^2$ et $c = y_1^2$. On obtient alors

$$u = z_1^2 = b^2+c^2 = x_1^4+y_1^4$$

et on vérifie que $|z_1| < |z|$ par exemple en observant que $z = u^2+v^2 = z_1^4+4a^4 > z_1$. □

Théorème. *L'équation $x^3+y^3=z^3$ n'a pas de solutions entières avec $xyz\neq 0$. Plus généralement il n'existe pas d'entiers algébriques $x,y,z\in\mathbf{Z}[j]$ et d'unité $u\in\mathbf{Z}[j]^*$ tels que $x^3+y^3=uz^3$ et $xyz\neq 0$.*

Il est commode de diviser la preuve en deux cas. L'un (facile) où on suppose xyz premier avec 3, l'autre (plus difficile) où disons z n'est pas premier avec 3. Le principe de la preuve du deuxième cas est de montrer que, si l'équation a une solution, alors elle possède une autre plus petite en un certain sens.

Commençons par rappeler que $A:=\mathbf{Z}[j]$ est principal, donc factoriel et que le groupe des unités est formé de $\pm 1,\pm j,\pm j^2$. En particulier on vérifie directement(*) que si $u\in A^*$ et $u\equiv\pm 1 \bmod \lambda^2$ alors $u=\pm 1$ (rappelons que λ désigne l'élément premier $1-j$).

Premier cas : λ ne divise pas xyz. Observons que si $x\equiv 1\bmod\lambda$ alors $x^3-1=(x-1)(x-j)(x-j^2)\equiv 0\bmod\lambda^4$; ainsi on a $x^3\equiv\pm 1\bmod\lambda^4$ et donc une solution de l'équation de Fermat entraînerait $\pm 1\pm 1\pm u\equiv 0\bmod\lambda^4$ (avec $u\in A^*$). On vérifie directement qu'une telle congruence est impossible.

Deuxième cas : λ divise xyz. On se ramène facilement au cas où λ divise z et ne divise pas xy, et ensuite on observe que λ^2 divise z car $\pm 1\pm 1\equiv uz^3\bmod\lambda^4$, donc $z^3\equiv 0\bmod\lambda^4$ donc $\mathrm{ord}_\lambda(z)\geq 4/3$. On montre alors l'énoncé de descente :

Si $x^3+y^3=uz^3$ avec $x,y,z\in A$, $u\in A^*$ et $\mathrm{ord}_\lambda(z)=m\geq 2$ alors il existe $x_1,y_1,z_1\in A$ et $u'\in A^*$ avec $x_1^3+y_1^3=u'z_1^3$ et $\mathrm{ord}_\lambda(z_1)=m-1$.

On commence bien sûr par factoriser :

$$(x+y)(x+jy)(x+j^2y)=uz^3$$

On voit que λ^2 doit diviser l'un des facteurs de gauche (car $\mathrm{ord}_\lambda(z^3)=3m\geq 6$), disons $x+y$ et alors $\mathrm{ord}_\lambda(x+jy)=\mathrm{ord}_\lambda(x+j^2y)=1$; en effet par exemple $x+jy=x+y-\lambda y$ et λ ne divise pas y. Ainsi le pgcd de deux des facteurs est exactement λ. Comme A est factoriel, on en tire

$$\begin{cases} x+y &= u_1X^3\lambda^{3m-2}\\ x+jy &= u_2Y^3\lambda\\ x+j^2y &= u_3Z^3\lambda\end{cases}$$

avec X,Y,Z premiers entre eux et avec λ et u_1,u_2,u_3 unités.

En multipliant les équations respectivement par 1, j et j^2 et en les additionnant on obtient $0=u_1X^3\lambda^{3m-2}+u_2jY^3\lambda+u_3j^2Z^3\lambda$. En simplifiant par λ et en posant $u_4:=ju_3/u_2$ et $u_5:=-j^2u_1/u_2$ on obtient

$$Y^3+u_4Z^3=u_5\left(\lambda^{m-1}X\right)^3.$$

On termine en remarquant que $\pm 1\pm u_4\equiv 0\bmod\lambda^2$ et donc $u_4=\pm 1$. On pose alors $x_1=Y$, $y_1=u_4Z$, $z_1=\lambda^{m-1}X$ et $u'=u_5$ et on a bien $x_1^3+y_1^3=u'z_1^3$ et $\mathrm{ord}_\lambda(z_1)=m-1$. □

Exercice : Ecrire les détails de la preuve.

C. Equation de Pell-Fermat $x^2-dy^2=1$.

Dans ce paragraphe on décrit les solutions de l'équation citée, en expliquant le lien avec les unités de l'anneau $\mathbf{Z}[\sqrt{d}]$ et les "bonnes" approximations rationnelles de $\sqrt{d}$.

Notons que l'équation possède toujours les solutions $(x,y)=(\pm 1,0)$ que l'on appellera triviales. Notons aussi que si d est un carré, disons $d=a^2$ alors $(x-ay)(x+ay)=1$ entraînerait $x+ay=x-ay=1$ (ou $=-1$) et donc $2ay=0$ et donc il n'y a aucune solution non triviale. Le cas intéressant est donc celui où d n'est pas un carré d'entier (et donc $\sqrt{d}\notin\mathbf{Q}$). Le théorème principal peut s'énoncer :

(*) Cette remarque est un cas très particulier du "lemme de Kummer" qui dit qu'une unité congrue modulo λ^{p-1} à un nombre rationnel est en fait la puissance p-ième d'une unité de $\mathbf{Z}[\exp(2\pi i/p)]$, pourvu que p soit "régulier" (en particulier quand l'anneau $\mathbf{Z}[\exp(2\pi i/p)]$ est factoriel, ce qui est vrai pour $p=3$).

Théorème. *Soit d entier positif qui n'est pas un carré, alors il existe une solution non triviale appelée* solution fondamentale $(x_1, y_1) \in \mathbf{N}^* \times \mathbf{N}^*$ *à l'équation $x^2 - dy^2 = 1$ telle que toutes les solutions en entiers positifs soient données par (x_n, y_n) où $x_n + y_n\sqrt{d} := (x_1 + y_1\sqrt{d})^n$ et les solutions générales par $(\pm x_n, \pm y_n)$.*

Le lien avec les approximations rationnelles de $\sqrt{d}$ est le suivant. Supposons que (x, y) soit une solution non triviale de l'équation (avec disons $x, y > 1$) alors

$$0 < \frac{x}{y} - \sqrt{d} = \frac{1}{y^2\left(\frac{x}{y} + \sqrt{d}\right)} < \frac{1}{2\sqrt{d}y^2}$$

Inversement, si $x/y \in \mathbf{Q}$ est une approximation qui vérifie l'inégalité précédente alors

$$0 < x^2 - dy^2 = y^2\left(\frac{x}{y} - \sqrt{d}\right)\left(\frac{x}{y} + \sqrt{d}\right) < \frac{1}{2\sqrt{d}}\left(2\sqrt{d} + \frac{1}{2\sqrt{d}y^2}\right) < 2$$

donc $x^2 - dy^2 = 1$ (puisque c'est un entier). Ainsi une solution positive (x, y) de l'équation de Pell-Fermat correspond à une approximation rationnelle x/y de $\sqrt{d}$ vérifiant $0 < \frac{x}{y} - \sqrt{d} < \frac{1}{2\sqrt{d}y^2}$.

Introduisons l'anneau $\mathbf{Z}[\sqrt{d}]$, l'homomorphisme $\sigma(a+b\sqrt{d}) = a-b\sqrt{d}$ (pourquoi est-ce un homomorphisme?) et la norme d'un élément $\alpha = a + b\sqrt{d}$ comme

$$\mathrm{N}(\alpha) = \alpha\sigma(\alpha) = a^2 - db^2.$$

La norme est multiplicative et on a, comme dans $\mathbf{Z}[i]$, le lemme suivant dont la preuve similaire est omise :

Lemme. *Dans l'anneau $\mathbf{Z}[\sqrt{d}]$ un élément est inversible si et seulement si sa norme vaut ± 1.*

Si l'on note $A^* = \mathbf{Z}[\sqrt{d}]^*$ et $U_1 = \{\alpha \mid \mathrm{N}(\alpha) = 1\}$, on voit que $(A^* : U_1) = 2$ ou 1 suivant qu'il existe ou non une unité de norme -1. Bien sûr les solutions (x, y) de l'équation de Pell-Fermat correspondent aux unités $x + y\sqrt{d} \in U_1$ et on voit que le théorème sur ces solutions se traduit en l'énoncé suivant :

Théorème. *Il existe une unité $\epsilon_1 \in \mathbf{Z}[\sqrt{d}]^*$ telle que*

$$\mathbf{Z}[\sqrt{d}]^* = \{\pm\epsilon_1^n \mid n \in \mathbf{Z}\} \cong \{\pm 1\} \times \mathbf{Z};$$

si $\mathrm{N}(\epsilon_1) = +1$ alors $U_1 = \mathbf{Z}[\sqrt{d}]^*$ et si $\mathrm{N}(\epsilon_1) = -1$ alors on a

$$U_1 = \left\{\pm\epsilon_1^{2n} \mid n \in \mathbf{Z}\right\} \cong \{\pm 1\} \times \mathbf{Z}.$$

Pour démontrer ce théorème introduisons l'application "logarithme" $L : \mathbf{Z}[\sqrt{d}]^* \to \mathbf{R}^2$ définie par la formule $L(\alpha) = (\log|\alpha|, \log|\sigma(\alpha)|)$.

Proposition. *L'application $L : \mathbf{Z}[\sqrt{d}]^* \to \mathbf{R}^2$ vérifie les propriétés suivantes :*
(i) L'application L est un homomorphisme $L(\alpha\beta) = L(\alpha) + L(\beta)$.
(ii) Le noyau est réduit à ± 1.
(iii) L'image est un sous-groupe discret.
(iv) L'image est contenue dans la droite $x + y = 0$.

Preuve. La propriété (i) est immédiate. La propriété (iv) provient de ce que $\log|\alpha| + \log|\sigma(\alpha)| = \log|\mathrm{N}(\alpha)| = 0$. Pour montrer (ii) et (iii) montrons que l'image réciproque par L d'une boule de $\mathbf{R}^2$ est finie, on en déduira d'une part que l'image est discète et d'autre part que le noyau de L est fini donc composé des racines de l'unité donc de ± 1 puisque $\mathbf{Z}[\sqrt{d}] \subset \mathbf{R}$. Or un élément $\alpha \in \mathbf{Z}[\sqrt{d}]^*$ est racine de $P := X^2 - t(\alpha)X + \mathrm{N}(\alpha) \in \mathbf{Z}[X]$ avec $t(\alpha) = \alpha + \sigma(\alpha)$ (la "trace") et $\mathrm{N}(\alpha) = \pm 1$, donc si $L(\alpha)$ est dans une boule de rayon C, on a $|\alpha| = \exp(\log|\alpha|) \leq \exp(C)$ et de même pour $|\sigma(\alpha)|$ donc $|t(\alpha)| \leq 2\exp(C)$. Il n'y a donc qu'un nombre fini de polynômes possibles et donc qu'un nombre fini de α. □

On utilise maintenant un lemme classique

Lemme. *Un sous-groupe discret G de $\mathbf{R}$ est de la forme $G = \mathbf{Z}\omega$.*

Preuve. (esquisse) Si $G = \{0\}$ on peut choisir $\omega = 0$. Sinon on choisit $\omega := \inf\{x \in G \mid x > 0\}$; comme G est discret on a $\omega > 0$ et $\omega \in G$ (sinon il y aurait une suite d'éléments de G convergeant vers ω, ce qui contredirait la discrétion de G). Enfin si $x \in G$, on choisit $m \in \mathbf{Z}$ tel que $m\omega \leq x < (m+1)\omega$, alors $0 \leq x - m\omega < \omega$ et $x - m\omega \in G$ donc $x = m\omega$. □

Ce lemme s'applique à $L(\mathbf{Z}[\sqrt{d}]^*)$ et fournit la preuve du théorème à condition de prouver l'existence d'une unité $\neq \pm 1$ ou d'une solution non triviale de l'équation de Pell-Fermat. Ces considérations montrent qu'il suffit de démontrer l'énoncé suivant :

Proposition. *Soit d entier positif qui n'est pas un carré, alors il existe une solution non triviale (x_1, y_1) (i.e. avec $y_1 \neq 0$) à l'équation $x^2 - dy^2 = 1$.*

Une bonne méthode pratique pour la construction de cette solution est la méthode des fractions continuées (ou fractions continues) esquissée en appendice. Nous allons démontrer l'existence de la solution en démontrant l'existence de bonnes approximations rationnelles de $\sqrt{d}$ sans les construire explicitement.

Lemme. *Soit $\alpha \in \mathbf{R}$ et $N \geq 1$, alors il existe un nombre rationnel $p/q \in \mathbf{Q}$ tel que*

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{qN} \qquad \text{et} \qquad 1 \leq q \leq N.$$

Preuve. Découpons l'intervalle $[0,1]$ en N intervalles de longueur $1/N$, alors parmi les $N+1$ nombres $j\alpha - [j\alpha]$ (pour $j = 0, \ldots, N$), il y en a deux dans un même petit intervalle donc distants d'au plus $1/N$; en d'autres termes, il existe $0 \leq j < \ell \leq N$ tels que $|(j\alpha - [j\alpha]) - (\ell\alpha - [\ell\alpha])| \leq 1/N$. On en déduit

$$\left|\alpha - \frac{[\ell\alpha] - [j\alpha]}{\ell - j}\right| \leq \frac{1}{(\ell - j)N}$$

d'où le résultat en posant $p := [\ell\alpha] - [j\alpha]$ et $q := \ell - j$. □

Remarquons qu'en particulier, l'approximation fournie par le lemme vérifie $|\alpha - p/q| \leq 1/q^2$.

Corollaire. *Soit $\alpha \in \mathbf{R} \setminus \mathbf{Q}$, alors il existe une infinité de nombres rationnels $p/q \in \mathbf{Q}$ tels que*

$$\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{q^2}.$$

Preuve. Soit $N_1 \geq 1$ et p_1/q_1 fourni par le lemme précédent tel que $\left|\alpha - \frac{p_1}{q_1}\right| \leq \frac{1}{q_1 N_1}$. Comme $\alpha \notin \mathbf{Q}$, le membre de gauche de l'inégalité est non nul, donc on peut choisir N_2 tel que $1/N_2 < \left|\alpha - \frac{p_1}{q_1}\right|$. Soit alors p_2/q_2 fourni par le lemme précédent tel que $\left|\alpha - \frac{p_2}{q_2}\right| \leq \frac{1}{q_2 N_2}$, on a

$$\left|\alpha - \frac{p_2}{q_2}\right| \leq \frac{1}{q_2 N_2} \leq 1/N_2 < \left|\alpha - \frac{p_1}{q_1}\right|$$

donc $p_2/q_2 \neq p_1/q_1$. Il est maintenant clair qu'on peut itérer indéfiniment ce processus. □

Remarques. 1) Si l'on supprimait l'hypothèse $\alpha \notin \mathbf{Q}$ dans l'énoncé du corollaire, le résultat serait faux. En effet si $\alpha = a/b$ et $a/b \neq p/q$ avec $\left|\alpha - \frac{p}{q}\right| \leq \frac{1}{q^2}$, alors

$$\frac{1}{bq} \leq \frac{|aq - bp|}{bq} = \left|\alpha - \frac{p}{q}\right| \leq \frac{1}{q^2}$$

et donc $q \leq b$ et il n'existe qu'un nombre fini de p/q. 2) Prenons l'exemple de $\alpha = \sqrt{d}$ avec d non carré ; on peut montrer que le corollaire est optimal au sens suivant : il existe une constante $C > 0$ telle que pour tout $p/q \in \mathbf{Q}$ on ait

$$\left|\sqrt{d} - \frac{p}{q}\right| \geq \frac{C}{q^2}.$$

Pour cela considérons $P(X) = X^2 - d = (X - \sqrt{d})(X + \sqrt{d})$, on a $|P(p/q)| \geq 1/q^2$. Or, si disons $|\sqrt{d} - p/q| \leq 1$ on a $|p/q| \leq \sqrt{d} + 1$ donc $|p/q + \sqrt{d}| \leq 2\sqrt{d} + 1$ et ainsi

$$\left|\sqrt{d} - \frac{p}{q}\right| = \frac{|P(p/q)|}{\left|p/q + \sqrt{d}\right|} \geq \frac{1}{(2\sqrt{d}+1)q^2}.$$

3) Si α est un nombre algébrique de degré $d \geq 3$ la même démonstration donnera $\left|\alpha - \frac{p}{q}\right| \geq \frac{C}{q^d}$. Roth a démontré (1955), mais la preuve est beaucoup plus difficile, que l'on a encore pour tout $\epsilon > 0$ une constante C (dépendant de α et ϵ) telle que :

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{C}{q^{2+\epsilon}}.$$

Appliquons le corollaire à $\sqrt{d} \notin \mathbf{Q}$; on obtient ainsi une infinité d'entiers (x, y) tels que $|\sqrt{d} - x/y| \leq 1/q^2$ et donc tels que $|\sqrt{d} + x/y| \leq 2\sqrt{d} + 1$ et enfin $|x^2 - dy^2| \leq 2\sqrt{d} + 1$. En particulier il existe un entier c avec une infinité de solutions à l'équation $x^2 - dy^2 = c$. Comme il n'y a qu'un nombre fini de classes modulo c, il existe même une infinité de solutions congrues deux à deux modulo c. Prenons donc (x_1, y_1) et (x_2, y_2) solutions de $x^2 - dy^2 = c$ et vérifiant $x_1 \equiv x_2 \bmod c$ et $y_1 \equiv y_2 \bmod c$. Posons :

$$u + v\sqrt{d} := \frac{x_1 + y_1\sqrt{d}}{x_2 + y_2\sqrt{d}}.$$

On a donc

$$u^2 - dv^2 = \mathrm{N}(u + v\sqrt{d}) = \frac{\mathrm{N}(x_1 + y_1\sqrt{d})}{\mathrm{N}(x_2 + y_2\sqrt{d})} = \frac{c}{c} = 1$$

et il suffit de voir que u, v sont entiers. On calcule donc :

$$u + v\sqrt{d} = \frac{(x_1 + y_1\sqrt{d})(x_2 - y_2\sqrt{d})}{x_2^2 - dy_2^2} = \frac{x_1x_2 - dy_1y_2}{c} + \frac{y_1x_2 - x_1y_2}{c}\sqrt{d}$$

et on observe que $x_1x_2 - dy_1y_2 \equiv x_1^2 - dy_1^2 \equiv 0 \bmod c$ et $y_1x_2 - x_1y_2 \equiv y_1x_1 - x_1y_1 \equiv 0 \bmod c$, ce qui achève la démonstration.

Compléments.

L'équation un peu plus générale $x^2 - dy^2 = m$ n'a pas toujours de solution. Par exemple si $m = -1$ et si p premier congru à 3 modulo 4 divise d alors une solution donnerait $x^2 \equiv -1 \bmod p$ ce qui est impossible. Plus généralement pour chaque p impair divisant d sans diviser m, on doit avoir $x^2 \equiv m \bmod p$ d'où $\left(\frac{m}{p}\right) = 1$. Inversement, si il existe une solution, il en existe une infinité puisque si $\mathrm{N}(\alpha) = m$ et $\mathrm{N}(u) = 1$ alors $\mathrm{N}(u\alpha) = m$. On peut énoncer

Proposition. *Soit $m \in \mathbf{Z} \setminus \{0\}$, et d non carré, il existe $\alpha_1, \ldots \alpha_r \in \mathbf{Z}[\sqrt{d}]$ tels que :*

$$\left\{\alpha \in \mathbf{Z}[\sqrt{d}] \mid \mathrm{N}(\alpha) = m\right\} = \alpha_1 U_1 \cup \ldots \cup \alpha_r U_1.$$

Preuve. Le fait que l'ensemble des solutions soit une réunion de classes modulo U_1 est clair, montrons qu'on peut prendre une réunion finie. Si $\mathrm{N}(\alpha) = m$, on a donc α divise m et ou encore $m\mathbf{Z}[\sqrt{d}] \subset \alpha\mathbf{Z}[\sqrt{d}]$. Mais

l'ensemble des idéaux contenant $m\mathbf{Z}[\sqrt{d}]$ est en bijection avec les idéaux du quotient $\mathbf{Z}[\sqrt{d}]/m\mathbf{Z}[\sqrt{d}]$ et par conséquent est un ensemble fini. Mais $\alpha\mathbf{Z}[\sqrt{d}] = \alpha'\mathbf{Z}[\sqrt{d}]$ équivaut à dire que α et α' sont égaux à une unité près. L'ensemble des solutions est donc fini modulo le groupe des unités donc également modulo le sous-groupe U_1. □

Exhibons maintenant un procédé pour calculer les bonnes approximations rationnelles d'un nombre réel : l'algorithme des *fractions continuées*.

Notation. Soit a_0 réel et $a_1, \ldots, a_n$ une suite de réels > 0, on pose

$$[a_0, a_1, \ldots, a_n] := a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ldots \cfrac{}{} + \cfrac{1}{a_n}}}$$

Définition. Si x est un nombre réel, on lui associe une suite d'entiers a_n et une suite auxiliaire de réels x_n définis ainsi : $a_0 := [x]$ et $x_0 := x$ puis $x_{n+1} = 1/(x_n - a_n)$ et $a_{n+1} = [x_{n+1}]$. On convient que la suite s'arrête quand x_n est entier (ce qui ne se produit que si $x \in \mathbf{Q}$). On définit la n-ième *réduite* (en anglais "*convergent*") comme

$$\frac{p_n}{q_n} := [a_0, a_1, \ldots, a_n].$$

Lemme. *On a les formules suivantes :*
(i) $x = [a_0, a_1, \ldots, a_{n-1}, x_n]$.
(ii) $p_{n+1} = a_{n+1}p_n + p_{n-1}$ *(avec* $p_0 = a_0$ *et* $p_1 = a_1a_0 + 1$*, alors que* $q_{n+1} = a_{n+1}q_n + q_{n-1}$ *(avec* $q_0 = 1$ *et* $q_1 = a_1$*).*
(iii) Si $p_n/q_n = [a_0, \ldots, a_n]$ *on a* $[a_0, \ldots, a_n, y] = (p_n y + p_{n-1})/(q_n y + q_{n-1})$.
(iv) $q_n p_{n-1} - p_n q_{n-1} = (-1)^n$.
(v) $q_n p_{n-2} - p_n q_{n-2} = (-1)^{n-1} a_n$.

Preuve. Observons tout d'abord que pour tous *réels* a_i on a $[a_0, \ldots, a_{n-1}, a_n] = [a_0, \ldots, a_{n-1} + 1/a_n]$. La première formule se montre par récurrence (le cas $n = 0$ étant vérifié par construction). Supposons donc $x = [a_0, \ldots, a_{n-1}, x_n]$ alors $[a_0, \ldots, a_n, x_{n+1}] = [a_0, \ldots, a_{n-1}, a_n + 1/x_{n+1}] = [a_0, \ldots, a_{n-1}, x_n] = x$. Ensuite

$$[a_0, \ldots, a_{n-1}, a_n, a_{n+1}] = [a_0, \ldots, a_{n-1}, a_n + 1/a_{n+1}] = \frac{p'_n}{q'_n}$$

où on peut supposer (par récurrence) que les p'_n, q'_n sont donnés par les formules $p'_{m+1} = a'_{m+1}p'_m + p'_{m-1}$ avec $a'_m = a_m$ pour $m \leq n-1$ et $a'_n = a_n + 1/a_{n+1}$. Ainsi $p'_n = (a_n + 1/a_{n+1})p'_{n-1} + p'_{n-2} = (a_n + 1/a_{n+1})p_{n-1} + p_{n-2}$ et $q'_n = (a_n + 1/a_{n+1})q'_{n-1} + q'_{n-2} = (a_n + 1/a_{n+1})q_{n-1} + q_{n-2}$ d'où

$$\frac{p'_n}{q'_n} = \frac{(a_n + 1/a_{n+1})p_{n-1} + p_{n-2}}{(a_n + 1/a_{n+1})q_{n-1} + q_{n-2}} = \frac{a_{n+1}(a_n p_{n_1} + p_{n-2}) + p_{n-1}}{a_{n+1}(a_n q_{n_1} + q_{n-2}) + q_{n-1}} = \frac{a_{n+1}p_n + p_{n-1}}{a_{n+1}q_n + q_{n-1}}.$$

La formule (iii) se démontre comme les précédentes. Les formules (iv) et (v) se démontrent également par récurrence, par exemple

$$p_{n+1}q_n - q_{n+1}p_n = (a_{n+1}p_n + p_{n-1})q_n - (a_{n+1}q_n + q_{n-1})p_n = -(p_n q_{n-1} - q_n p_{n-1})$$

et de même pour (v). □

Remarque : ces formules de récurrence permettent de calculer p_n et q_n, à partir du calcul des a_n; comme $a_n \geq 1$, on voit également que q_n croit au moins comme une suite de Fibonacci et qu'on a l'estimation $q_n \geq \left(\frac{1+\sqrt{5}}{2}\right)^{n-1}$. On peut également tirer de ces formules

Théorème. *La suite p_n/q_n converge vers x ; plus précisément la suite des p_{2n}/q_{2n} est croissante et converge vers x, la suite des p_{2n+1}/q_{2n+1} est décroissante et converge vers x. On a l'estimation :*

$$\frac{1}{q_n(q_n+q_{n+1})} < \left|x - \frac{p_n}{q_n}\right| < \frac{1}{q_n q_{n+1}}$$

De plus les réduites définissent les meilleures approximations de x au sens suivant. Si $q \leq q_n$ et $p/q \neq p_n/q_n$, alors :

$$q_n \left|x - \frac{p_n}{q_n}\right| < q \left|x - \frac{p}{q}\right|$$

et en fait si $\left|x - \frac{p}{q}\right| < 1/2q^2$ alors il existe n tel que $p/q = p_n/q_n$.

Preuve. On a $a_n = [x_n] \leq x_n$, or la fonction $[a_0, \ldots, a_n]$ est clairement une fonction croissante (resp. décroissante) de a_m pour m pair (resp. m impair) donc si n est pair $[a_0, \ldots, a_n] \leq [a_0, \ldots, x_n] = x$ et l'inverse si n impair. D'après le lemme on a

$$\frac{p_{n-1}}{q_{n-1}} - \frac{p_n}{q_n} = \frac{(-1)^n}{q_n q_{n-1}} \qquad \text{et} \qquad \frac{p_{n-2}}{q_{n-2}} - \frac{p_n}{q_n} = \frac{(-1)^{n-1} a_n}{q_n q_{n-2}}.$$

Ainsi on peut ordonner par exemple

$$\frac{p_{2n}}{q_{2n}} < \frac{p_{2n+2}}{q_{2n+2}} < x < \frac{p_{2n+1}}{q_{2n+1}} < \frac{p_{2n-1}}{q_{2n-1}}$$

et ainsi $|x - p_n/q_n| \leq |p_n/q_n - p_{n+1}/q_{n+1}| = 1/q_n q_{n+1}$ alors que $|x - p_n/q_n| \geq |p_n/q_n - p_{n+2}/q_{n+2}| = a_{n+2}/q_n q_{n+2} = a_{n+2}/q_n(a_{n+2}q_{n+1} + q_n) \geq 1/q_n(q_{n+1} + q_n)$. Ces estimations montrent clairement que la suite p_n/q_n converge vers x. Observons au passage que $1/q_{n+2} < |p_n - xq_n| < 1/q_{n+1}$ donc la suite $|p_n - xq_n|$ est strictement décroissante. Soit maintenant p/q une fraction avec $q \leq q_n$ et $p/q \neq p_n/q_n$. On peut supposer $q_{n-1} < q$; résolvons le système de Cramer $up_n + vp_{n-1} = p$ et $uq_n + vq_{n-1} = q$, on obtient $u = \pm(pq_n - qp_{n-1})$ et $v = \pm(pq_n - qp_n)$ et en particulier u, v sont entiers et non nuls. Comme $q = uq_n + vq_{n-1} \leq q_n$ on voit que u et v sont de signes opposés et donc les deux quantités $u(p_n - q_n x)$ et $v(p_{n-1} - q_{n-1}x)$ sont de même signe. Or $p - qx = u(p_n - q_n x) + v(p_{n-1} - q_{n-1}x)$ donc

$$|p - qx| = |u(p_n - q_n x)| + |v(p_{n-1} - q_{n-1}x)| \geq |p_n - q_n x| + |p_{n-1} - q_{n-1}x|.$$

Enfin si $|x - p/q| < 1/2q^2$, posons $x - p/q = \epsilon\theta/q^2$ avec $\epsilon = \pm 1$ et $0 < \theta < 1/2$. Développons $p/q = [a_0, \ldots, a_m]$ en fraction continuée finie, en remarquant que si $a_m > 1$ on a $[a_0, \ldots, a_m] = [a_0, \ldots, a_m - 1, 1]$, on voit qu'on peut choisir la parité de m (*). On choisit cette parité de sorte que $p_{m-1}q - pq_{m-1} = (-1)^m = \epsilon$. Définissons maintenant y par l'égalité $x = (yp_m + p_{m-1})/(yq_m + q_{m-1})$, après calcul on en tire $y = (q - \theta q_{m-1})/\theta q$. En utilisant $q_{m-1} < q$ et $\theta < 1/2$ on voit que $y > 1$ donc on peut écrire $y = [a_{m+1}, \ldots]$ avec $a_{m+1} \geq 1$ et on obtient le développement en fraction continuée $x = [a_0, \ldots, a_m, a_{m+1}, \ldots]$ qui montre que $p/q = [a_0, \ldots, a_m]$ est bien une réduite. □

Remarques. 1) Lorsque $x \in \mathbf{Q}$, le développement en fraction continuée est fini (i.e. il existe n tel que $a_n = 0$). Lorsque $x \in \mathbf{R} \setminus \mathbf{Q}$ on a donc $x = \lim_n [a, \ldots, a_n]$, ce que l'on note aussi $x = [a, \ldots, a_n, a_{n+1}, \ldots]$ et on dit que la suite des a_n donne le développement de x en fraction continuée. 2) Lorsque $x \notin \mathbf{Q}$ on a donc $x = \lim_{n\to\infty}[a_0, \ldots, a_n]$, ce que, par convention on écrit $x = [a_0, \ldots, a_n, \ldots]$ et on dit que l'on a écrit le développement en fraction continuée de x. 2) D'après ce que nous avons vu, une solution de l'équation de Pell-Fermat $p^2 - dq^2 = 1$ fournit une bonne approximation p/q de $\sqrt{d}$ qui doit donc apparaître comme une réduite du développement en fraction continuée de $\sqrt{d}$ et on pourra donc la trouver ainsi.

Exemples. Le développement en fraction continuée de $x = \sqrt{2}$ (resp. $y = \sqrt{7}$) s'écrit

$$\sqrt{2} = [1, 2, 2, 2, \ldots] \qquad \text{et} \qquad \sqrt{7} = [2, 1, 1, 1, 4, 1, 1, 1, 4, \ldots]$$

(*) On peut d'ailleurs démontrer que c'est le seul cas d'ambigüité d'écriture en fraction continuée.

et on constate que le développement est périodique ; en fait un théorème de Lagrange indique qu'il en est toujours ainsi pour x quadratique ; c'est même une condition nécessaire et suffisante. Dans le cas de $\sqrt{2}$, la réduite initiale p_0/q_0 donne $p_0^2 - 2q_0^2 = -1$ et $p_1/q_1 = 3/2$ donne $p_1^2 - 2q_1^2 = +1$; dans le cas de $\sqrt{7}$, la réduite $p_3/q_3 = 8/3$ donne $p_3^2 - 7q_3^2 = +1$. Le fait que le développement en fraction continuée soit périodique est un cas particulier du théorème de Lagrange cité ci-dessus qui affirme que le développement en fraction continuée d'un réel x est périodique si et seulement si x est quadratique, i.e. racine d'une équation à coeeficients entiers de degré deux.

Donnons un exemple illustrant la qualité de l'algorithme des fractions continuées : la recherche de solutions de l'équation $x^2 - 61y^2 = 1$ (essayer de trouver une solution à tâtons!). Le développement en fraction continuée de $x = \sqrt{61}$ s'écrit

$$\sqrt{61} = [7, 1, 4, 3, 1, 2, 2, 1, 3, 4, 1, 14, 1, \ldots]$$

et le développement devient périodique à partir de $a_{12} = a_1 = 1$. Les premières réduites sont

$$\frac{7}{1}, \frac{8}{1}, \frac{39}{5}, \frac{125}{16}, \frac{164}{21}, \frac{453}{58}, \frac{1070}{137}, \frac{1523}{195}, \frac{5639}{722}, \frac{24079}{3083}, \frac{29718}{3805}, \frac{440131}{56353}, \frac{469849}{60158}$$

La dixième réduite $p_{10}/q_{10} = 29718/3805$ fournit la première solution de $x^2 - 61y^2 = -1$ et l'on en déduit que la solution fondamentale de $x^2 - 61y^2 = 1$ est donnée par $x + y\sqrt{61} = (p_{10} + q_{10}\sqrt{61})^2$ soit :

$$(x_1, y_1) = (1766319049, 226153980).$$

D. Anneaux d'entiers algébriques.

On donne ici quelques indications sur les propriétés générales des anneaux extensions de $\mathbf{Z}$ *qui nous amène à la notion d'anneau de Dedekind ; les outils sont l'algèbre et un peu de géométrie des nombres.*

On connait la notion d'élément algébrique ("algébrique" est toujours entendu ici au sens de "algébrique sur le corps de rationnels") ; c'est une notion de théorie des corps, la notion correspondante pour les anneaux est la suivante.

Définition. Un *entier algébrique* est un nombre complexe α vérifiant une équation polynomiale unitaire à coefficients entiers. Plus généralement un élément α est dit *entier* ou *entier algébrique* sur un anneau A s'il vérifie une équation polynomiale unitaire à coefficients dans A.

Exemple. Un nombre rationnel $\alpha = a/b$ est racine de $bX - a \in \mathbf{Z}[X]$ et on voit que α est un entier algébrique si et seulement si c'est un entier. Remarquons également que si α est algébrique sur $\mathbf{Q}$, alors il existe un entier $d \in \mathbf{Z}$ (un "dénominateur") tel que $d\alpha$ soit un entier algébrique. Cette propriété de $\mathbf{Z}$ se généralise ainsi.

Définition. Un anneau intègre A est *intégralement clos* si les seuls éléments de $K := \mathrm{Frac}(A)$ qui sont entiers algébriques sur A sont les éléments de A.

Exemples. On montre facilement qu'un anneau principal ou factoriel est intégralement clos. Par contre l'anneau $A = \mathbf{Z}[\sqrt{5}]$ n'est pas intégralement clos puisque $\alpha := \frac{1+\sqrt{5}}{2}$ est dans $\mathbf{Q}(\sqrt{5})$, est racine de $X^2 - X - 1$ donc est entier sur A (ou même $\mathbf{Z}$) sans être dans A.

Lemme. *L'élément α est entier algébrique sur A si et seulement si l'anneau $A[\alpha]$ est un A-module de type fini ou encore si et seulement si l'anneau $A[\alpha]$ est contenu dans un A-module de type fini.*

Corollaire. *La somme, la différence, le produit de deux entiers algébriques est un entier algébrique. Si α est entier sur B et chaque élément de B entier sur A alors α est entier sur A. En particulier, si K est une extension de* $\mathbf{Q}$*, l'ensemble :*

$$\mathcal{O}_K := \{\alpha \in K \mid \alpha \text{ est un entier algébrique}\}$$

est un anneau.

Preuve. Si α est entier sur A, il vérifie une relation $\alpha^n = a_{n-1}\alpha^{n-1} + \ldots + a_0$ avec $a_i \in A$ donc $A[\alpha] = A + A\alpha + \ldots + A\alpha^{n-1}$ est bien un A-module de type fini. Inversement si $A[\alpha]$ est contenu dans un A-module de type fini $Au_1 + \ldots + Au_m$, on peut écrire $\alpha u_i = \sum_{j=1}^m a_{i,j}u_j$ avec $a_{i,j} \in A$; notons M la matrice $m \times m$ de coefficients $a_{i,j}$ alors le polynôme $P(X) := \det(XId - A)$ est unitaire à coefficients dans A et $P(\alpha) = 0$ (penser au théorème de Cayley-Hamilton, ou refaire sa démonstration) donc α est entier sur A. Pour le corollaire, on observe que si α, β sont entiers algébriques, alors $\mathbf{Z}[\alpha, \beta]$ est un $\mathbf{Z}$-module de type fini (une partie génératrice est donnée par un nombre fini de $\alpha^i\beta^j$) donc tous ses éléments sont entiers sur $\mathbf{Z}$. □

On appellera *corps de nombres* un K extension finie de $\mathbf{Q}$ et $\mathcal{O}_K$ l'anneau des entiers de K. On peut toujours supposer (théorème de l'élément primitif) qu'il existe α tel que $K = \mathbf{Q}(\alpha)$.

Définition. Soit K un corps de nombres et $\alpha \in K$; on définit la *norme* $\mathrm{N}(\alpha) = \mathrm{N}_{\mathbf{Q}}^K(\alpha)$ (resp. la *trace* $\mathrm{Tr}(\alpha) = \mathrm{Tr}_{\mathbf{Q}}^K(\alpha)$) comme le déterminant (resp. la trace) de la multiplication par α, vue comme application $\mathbf{Q}$-linéaire de K vers K.

On a $\mathrm{N}(\alpha\beta) = \mathrm{N}(\alpha)\,\mathrm{N}(\beta)$ et $\mathrm{Tr}(\alpha + \beta) = \mathrm{Tr}(\alpha) + \mathrm{Tr}(\beta)$; d'autre part, si $\alpha \in \mathcal{O}_K$ alors $\mathrm{N}(\alpha)$ et $\mathrm{Tr}(\alpha)$ sont dans $\mathbf{Z}$ car ce sont à la fois des rationnels et des entiers algébriques. En effet on peut donner une expression plus concrète de la trace et de la norme :

Lemme. *Soit α algébrique sur $\mathbf{Q}$ et $K = \mathbf{Q}(\alpha)$, soit $P(X) = X^d + a_{d-1}X^{d-1} + \ldots + a_0 = (X - \alpha_1)\ldots(X - \alpha_d)$ le polynôme minimal de α sur $\mathbf{Q}$, alors*

$$\mathrm{N}_{\mathbf{Q}}^K(\alpha) = \alpha_1 \ldots \alpha_d \qquad \text{et} \qquad \mathrm{Tr}_{\mathbf{Q}}^K(\alpha) = \alpha_1 + \ldots + \alpha_d.$$

Plus généralement si $\alpha \in K$ et $m = [K : \mathbf{Q}(\alpha)]$, on a $\mathrm{N}_{\mathbf{Q}}^K(\alpha) = (\alpha_1 \ldots \alpha_d)^m$ et $\mathrm{Tr}_{\mathbf{Q}}^K(\alpha) = m(\alpha_1 + \ldots + \alpha_d)$.

Preuve. Traitons le cas $K = \mathbf{Q}(\alpha)$ et laissons en exercice le cas général. Il suffit de remarquer que le polynôme caractéristique de la multiplication par α, vue comme application $\mathbf{Q}$-linéaire de K vers K n'est autre que le polynôme minimal de α. C'est clair si l'on choisit comme $\mathbf{Q}$-base de K les éléments $1, \alpha, \ldots, \alpha^{d-1}$. □

Exemples. 1) Si $K = \mathbf{Q}(\sqrt{d})$ avec d sans facteurs carrés, alors $\mathcal{O}_K = \mathbf{Z}[\sqrt{d}]$ si $d \equiv 2$ ou $3 \bmod 4$ mais $\mathcal{O}_K = \mathbf{Z}\left[\frac{1+\sqrt{d}}{2}\right]$ si $d \equiv 1 \bmod 4$. En effet si $\alpha \in \mathcal{O}_K$, écrivons $\alpha = x + y\sqrt{d}$ avec, a priori, $x, y \in \mathbf{Q}$. On sait que la trace et la norme sont dans $\mathbf{Z}$ et en fait, comme α est racine de $X^2 - \mathrm{Tr}(\alpha)X + \mathrm{N}(\alpha)$, cela équivaut même à $\alpha \in \mathcal{O}_K$. Or $\mathrm{Tr}(\alpha) = 2x$ et $\mathrm{N}(\alpha) = x^2 - dy^2$, donc $x = a/2$, $y = b/2$ avec $a, b \in \mathbf{Z}$ et $a^2 - db^2 \in 4\mathbf{Z}$. Si a est pair, b également et inversement ; si a et b sont impairs on obtient $d \equiv 1 \bmod 4$, d'où le résultat. 2) Si $K = \mathbf{Q}(\zeta)$ avec $\zeta =: \exp(2\pi i/p)$ alors $\mathcal{O}_K = \mathbf{Z}[\zeta]$. On a vu que $\lambda\mathbf{Z}[\zeta] \cap \mathbf{Z} = p\mathbf{Z}$ (rappelons que $\lambda := 1 - \zeta$). Si $\alpha = a_0 + a_1\zeta + \ldots + a_{p-2}\zeta^{p-2}$ on vérifie que $\mathrm{Tr}(\lambda\alpha) = pa_0$ et donc $a_0 \in \mathbf{Z}$. On recommence avec $\alpha' := \zeta^{-1}(\alpha - a_0)$ et on conclut que $a_1 \in \mathbf{Z}$ et ainsi de suite.

Proposition. *Si $[K : \mathbf{Q}] = n$ alors il existe $e_1, \ldots, e_n \in \mathcal{O}_K$ tels que $\mathcal{O}_K = \mathbf{Z}e_1 \oplus \ldots \oplus \mathbf{Z}e_n$ (comme groupe abélien ou $\mathbf{Z}$-module).*

Remarquer qu'il n'est pas toujours vrai qu'il existe un entier algébrique α tel que $\mathcal{O}_K = \mathbf{Z}[\alpha]$.

Preuve. Nous allons utiliser l'observation que la forme $\mathbf{Q}$-bilinéaire $(x, y) := \mathrm{Tr}(xy)$ de $K \times K$ vers $\mathbf{Q}$ est non dégénérée. Soit $f_1, \ldots, f_n$ une base de K sur $\mathbf{Q}$, quitte à les multiplier par un dénominateur commun $d \in \mathbf{Z}$, on peut supposer $f_i \in \mathcal{O}_K$. Définissons $f_1^*, \ldots, f_n^*$ la base duale (i.e. telle que $\mathrm{Tr}(f_i f_j^*) = \delta_{ij}$) et notons d un dénominateur commun des f_j^*. Soit donc $x \in \mathcal{O}_K$, écrivons $x = x_1 f_1 + \ldots + x_n f_n$ avec $x_i \in \mathbf{Q}$. On remarque que $\mathrm{Tr}(x(df_i^*)) = d\,\mathrm{Tr}(xf_i^*) = dx_i$ est dans $\mathbf{Z}$ et on obtient ainsi

$$\mathbf{Z}f_1 \oplus \ldots \oplus \mathbf{Z}f_n \subset \mathcal{O}_K \subset \frac{1}{d}\left(\mathbf{Z}f_1 \oplus \ldots \oplus \mathbf{Z}f_n\right)$$

d'où l'on tire l'énoncé voulu. □

On en tire facilement que, si I est idéal non nul, alors il existe $e_i' \in I$ tels que I s'écrit $\mathbf{Z}e_1' \oplus \ldots \oplus \mathbf{Z}e_n'$; on voit alors que $\mathrm{card}(\mathcal{O}_K/I)$ est fini. On note $\mathrm{N}(I)$ ce cardinal (c'est la *norme* de l'idéal) ; on a alors

Proposition. *Soit $\alpha \in \mathcal{O}_K$ alors*

$$\mathrm{N}(\alpha\mathcal{O}_K) = \left|\mathrm{N}_{\mathbf{Q}}^K(\alpha)\right|$$

De plus la norme est multiplicative sur les idéaux : $\mathrm{N}(IJ) = \mathrm{N}(I)\,\mathrm{N}(J)$.

Preuve. Si M est une application $\mathbf{Z}$-linéaire de $\mathbf{Z}^n$ vers $\mathbf{Z}^n$ de déterminant non nul, on a $\mathrm{card}(\mathbf{Z}^n/M\mathbf{Z}^n) = |\det(M)|$. Si l'on considère maintenant $M(\alpha)$ la multiplication par α de $\mathcal{O}_K$ vers $\mathcal{O}_K$ on obtient

$$\mathrm{N}(\alpha\mathcal{O}_K) = \mathrm{card}\,(\mathcal{O}_K/\alpha\mathcal{O}_K) = |\det(M(\alpha))| = \left|\mathrm{N}_{\mathbf{Q}}^K(\alpha)\right|.$$

Pour la deuxième propriété, il suffit, à cause du théorème de décomposition des idéaux énoncé ci-dessous, de prouver la formule $\mathrm{N}(IJ) = \mathrm{N}(I)\,\mathrm{N}(J)$ avec J premier (non nul) donc maximal. Comme $IJ \subset I$ on a

$$\mathrm{card}\,(\mathcal{O}_K/IJ) = \mathrm{card}\,(\mathcal{O}_K/I)\,\mathrm{card}\,(I/IJ).$$

Comme J est maximal, $k := \mathcal{O}_K/J$ est un corps (fini) ; d'autre part I/IJ est un $\mathcal{O}_K$-module annulé par J donc on peut le voir comme un k-module ou k-espace vectoriel ; montrons qu'il est de dimension 1, ce qui montrera que $\mathrm{card}(I/IJ) = \mathrm{card}\,(\mathcal{O}_K/J)$ et achèvera la preuve. Si on avait un k-sous-espace vectoriel $\{0\} \subset L \subset I/IJ$, ce serait également un A-module et il correspondrait donc à un idéal I' tel que $L = I'/IJ$ et $IJ \subset I' \subset J$. Comme J est maximal on doit avoir $I' = IJ$ ou $I' = J$. □

Exemple d'anneau non principal. L'anneau $\mathbf{Z}[i\sqrt{3}]$ n'est pas principal, ni factoriel car il n'est pas intégralement clos : $\mathbf{Z}[i\sqrt{3}] \overset{\subset}{\neq} \mathbf{Z}[(1+i\sqrt{3})/2]$ cependant on peut justement l'inclure dans l'anneau principal $\mathbf{Z}[j]$. Plus fondamentalement les anneaux $\mathbf{Z}[\sqrt{10}]$ et $\mathbf{Z}[i\sqrt{5}]$ ne sont ni principal ni factoriel ; en effet

$$9 = 3^2 = (\sqrt{10}+1)(\sqrt{10}-1) \qquad \text{et} \qquad 6 = 2.3 = (1+i\sqrt{5})(1-i\sqrt{5})$$

donnent deux décompositions essentiellement différentes en produits d'éléments irréductibles. En fait on peut montrer directement que l'idéal engendré par 3 et $\sqrt{10}+1$ dans $\mathbf{Z}[\sqrt{10}]$ (resp. l'idéal engendré par 2 et $i\sqrt{5}+1$ dans $\mathbf{Z}[i\sqrt{5}]$) n'est pas principal car le quotient par l'idéal est $\mathbf{Z}/3\mathbf{Z}$ (resp. $\mathbf{Z}/2\mathbf{Z}$) et il n'y a aucun élément de norme 3 dans $\mathbf{Z}[\sqrt{10}]$ (resp. de norme 2 dans $\mathbf{Z}[i\sqrt{5}]$).

Définition. Un anneau A est un *anneau de Dedekind* s'il est noethérien, intégralement clos et si tout idéal premier non nul est maximal.

Exemple fondamental : l'anneau des entiers $\mathcal{O}_K$ d'un corps de nombres est un anneau de Dedekind. En effet il est intégralement clos par construction et, comme le quotient par un idéal non nul est fini, les deux autres conditions sont aisément vérifiées : l'ensemble des idéaux contenant un idéal non nul donné est fini et un anneau fini est intègre si et seulement c'est un corps.

La propriété fondamentale des anneaux de Dedekind – celle qui remplace en quelque sorte la factorialité – peut se formuler ainsi

Théorème. *Soit A un anneau de Dedekind, tout idéal non nul se décompose en produit d'idéaux premiers ; de plus on a unicité de la décomposition (à l'ordre près).*

Pour la preuve je renvoie au livre de Samuel *théorie algébrique des nombres* (chapitre III). Une autre propriété importante, qui peut d'ailleurs servir de définition pour les anneaux de Dedekind est que les idéaux fractionnaires sont inversibles et en particulier si l'on considère le monoïde des idéaux et qu'on le quotiente par les idéaux principaux, on obtient un groupe que nous précisons ci-dessous (un idéal fractionnaire I est un sous-$\mathcal{O}_K$-module de K tel qu'il existe $d \in \mathcal{O}_K$ tel que $dI \subset \mathcal{O}_K$; il est *inversible* s'il existe I' tel que $II' = \mathcal{O}_K$).

Nous allons maintenant décrire un peu plus les idéaux premiers. Tout d'abord remarquons que si $\wp$ est un idéal premier non nul de $\mathcal{O}_K$, alors $\wp \cap \mathbf{Z}$ est un idéal premier non nul de $\mathbf{Z}$, donc de la forme $p\mathbf{Z}$ pour un certain nombre premier p ; ainsi tout idéal premier $\wp$ est associé à un p qui est aussi la caractéristique du corps résiduel $\mathcal{O}_K/\wp$. Inversement, si p est un nombre premier, alors on peut considérer l'idéal qu'il engendre

dans $\mathcal{O}_K$; celui-ci n'a aucune raison d'être encore premier en général et s'écrira donc, au vu du théorème ci-dessus :

$$p\mathcal{O}_K = \wp_1^{e_1} \dots \wp_s^{e_s} \quad \text{avec } \wp_i \text{ idéaux premiers distincts et } e_i \geq 1.$$

Notons $f_i = [\mathcal{O}_K/\wp_i : \mathbf{F}_p]$ de sorte que $\mathrm{N}\,\wp_i = p^{f_i}$, en prenant les normes on obtient

$$\mathrm{N}(p\mathcal{O}_K) = p^n = \mathrm{N}\,\wp_1^{e_1} \dots \mathrm{N}\,\wp_s^{e_s} = p^{e_1 f_1 + \dots + e_s f_s}$$

d'où l'on tire la relation

$$\sum_{i=1}^{s} e_i f_i = n.$$

En utilisant le théorème des restes chinois, on peut aussi observer que :

$$\mathcal{O}_K/p\mathcal{O}_K \cong (\mathcal{O}_K/\wp_1^{e_1}) \times \dots \times (\mathcal{O}_K/\wp_s^{e_s})\,.$$

Décrire les idéaux premiers de K revient ainsi à décrire la décomposition des premiers de $\mathbf{Z}$ dans $\mathcal{O}_K$.

Exemple (décomposition des premiers dans un corps quadratique). Dans le cas où $K = \mathbf{Q}(\sqrt{d})$ et $[K : \mathbf{Q}] = 2$ (on peut supposer d sans facteurs carrés), on a trois possibilités de décomposition :
(i) On a $p\mathcal{O}_K = \wp_1\wp_2$ avec $\mathrm{N}\,\wp_i = p$; on dit que p est *décomposé* dans K.
(i) On a $p\mathcal{O}_K = \wp_1$ avec $\mathrm{N}\,\wp_1 = p^2$; on dit que p est *inerte* dans K.
(i) On a $p\mathcal{O}_K = \wp_1^2$ avec $\mathrm{N}\,\wp_1 = p$; on dit que p est *ramifié* dans K.

Ce qui correspond respectivement à $s = 2$, $e_1 = e_2 = 1$ et $f_1 = f_2 = 1$, ou bien $s = 1$, $e_1 = 1$ et $f_1 = 2$, ou bien $s = 1$, $e_1 = 2$ et $f_1 = 1$. On peut caractériser ces cas à l'aide du symbole de Legendre.

Théorème. *Soit $K = \mathbf{Q}(\sqrt{d})$ un corps quadratique avec d sans facteurs carrés. Si p est un nombre premier impair on a*
(i) p est décomposé dans K si et seulement si $\left(\frac{d}{p}\right) = +1$.
(ii) p est inerte dans K si et seulement si $\left(\frac{d}{p}\right) = -1$.
(iii) p est ramifié dans K si et seulement si $\left(\frac{d}{p}\right) = 0$, c'est-à-dire si p divise d.

Pour le nombre $p = 2$ la loi de décomposition est donnée par
(i) 2 est décomposé dans K si et seulement si $d \equiv 1 \bmod 8$.
(ii) 2 est inerte dans K si et seulement si $d \equiv 5 \bmod 8$.
(ii) 2 est ramifié dans K si et seulement si $d \equiv 2$ ou $3 \bmod 4$.

Preuve. Si p est premier impair, remarquons que $\mathcal{O}_K/p\mathcal{O}_K \cong \mathbf{Z}[\sqrt{d}]/p\mathbf{Z}[\sqrt{d}]$; c'est trivial si $d \equiv 2$ ou $3 \bmod 4$ et si $d \equiv 1 \bmod 4$ il suffit de remarquer que, si b est un entier impair, $a + b\left(\frac{1+\sqrt{d}}{2}\right) = a + \left(\frac{b-p}{2}\right)(1 + \sqrt{d}) + p\left(\frac{1+\sqrt{d}}{2}\right)$ donc $\mathcal{O}_K = \mathbf{Z}[\sqrt{d}] + p\mathcal{O}_K$. Ensuite on obtient les isomorphismes

$$A := \mathcal{O}_K/p\mathcal{O}_K \cong \mathbf{Z}[\sqrt{d}]/p\mathbf{Z}[\sqrt{d}] \cong \mathbf{Z}[X]/(p, X^2 - d)\mathbf{Z}[X] \cong \mathbf{F}_p[X]/(X^2 - d)\mathbf{F}_p[X].$$

On voit alors que, ou bien $X^2 - d$ se décompose dans $\mathbf{F}_p[X]$ en deux facteurs distincts, ce qui correspond à $\left(\frac{d}{p}\right) = +1$ et alors $A \cong \mathbf{F}_p \times \mathbf{F}_p$ et p est décomposé, ou bien $X^2 - d$ est irréductible dans $\mathbf{F}_p[X]$, ce qui correspond à $\left(\frac{d}{p}\right) = -1$ et alors $A \cong \mathbf{F}_{p^2}$ et p est inerte, ou bien $X^2 - d$ a une racine double dans $\mathbf{F}_p[X]$, ce qui correspond à $d = 0$ dans $\mathbf{F}_p$ ou encore $\left(\frac{d}{p}\right) = 0$ et alors $A \cong \mathbf{F}_p[X]/X^2\mathbf{F}_p[X]$ et p est ramifié.

Si $p = 2$ et $d \equiv 2$ ou $3 \bmod 4$ on a alors

$$\mathcal{O}_K/2\mathcal{O}_K \cong \mathbf{Z}[\sqrt{d}]/2\mathbf{Z}[\sqrt{d}] \cong \mathbf{Z}[X]/(2, X^2 - d)\mathbf{Z}[X] \cong \mathbf{F}_2[X]/(X^2 - d)\mathbf{F}_2[X] \cong \mathbf{F}_2[X]/(X - d)^2\mathbf{F}_2[X]$$

donc 2 est ramifié. Si maintenant $d \equiv 1 \bmod 4$, comme le polynôme minimal de $\frac{1+\sqrt{d}}{2}$ est $X^2 - X - \frac{d-1}{4}$, on a

$$\mathcal{O}_K/2\mathcal{O}_K \cong \mathbf{Z}\left[\frac{1+\sqrt{d}}{2}\right]/2\mathbf{Z}\left[\frac{1+\sqrt{d}}{2}\right] \cong \mathbf{Z}[X]/(2, X^2 - X - \frac{d-1}{4})\mathbf{Z}[X] \cong \mathbf{F}_2[X]/(X^2 - X - \frac{d-1}{4})\mathbf{F}_2[X].$$

Ainsi si $\frac{d-1}{4} \equiv 0 \bmod 2$, c'est-à-dire $d \equiv 1 \bmod 8$, on $X^2 - X - \frac{d-1}{4} = X(X-1)$ dans $\mathbf{F}_2[X]$ donc $\mathcal{O}_K/2\mathcal{O}_K \cong \mathbf{F}_2 \times \mathbf{F}_2$ et 2 est décomposé alors que si $\frac{d-1}{4} \equiv 1 \bmod 2$, c'est-à-dire $d \equiv 5 \bmod 8$, on $X^2 - X - \frac{d-1}{4} = X^2 + X + 1$ irréductible dans $\mathbf{F}_2[X]$ donc $\mathcal{O}_K/2\mathcal{O}_K \cong \mathbf{F}_4$ et 2 est inerte. □

On a vu que les anneaux $\mathcal{O}_K$ ne sont pas en général principaux mais on peut montrer qu'ils sont "presque principaux" au sens suivant. On a une notion de produit d'idéaux et on peut définir une relation d'équivalence sur les idéaux non nuls par $I \sim J$ si il existe $\alpha, \beta \in \mathcal{O}_K \setminus \{0\}$ tels que $\alpha I = \beta J$. On démontre que tout idéal est inversible modulo cette relation et on peut donc parler du *groupe des classes d'idéaux* noté $C\ell_K$ ou $\mathrm{Pic}(\mathcal{O}_K)$. Le second théorème central de la théorie est le théorème de finitude.

Théorème. *Soit K un corps de nombres alors le groupe des classes d'idéaux $C\ell_K$ est fini.*

Corollaire. *Soit $h_K := \mathrm{card}(C\ell_K)$ alors pour tout idéal non nul I de $\mathcal{O}_K$, l'idéal I^{h_K} est principal. Inversement, si $\mathrm{pgcd}(h_K, m) = 1$ et I^m est principal, alors I est principal.*

Nous expliquons ci-dessous les grandes lignes de la démonstration de cet énoncé ainsi que la structure du groupe des unités de $\mathcal{O}_K$. Ces deux dernières propriétés (finitudes du groupe des classes, génération finie du groupe des unités) ne sont pas purement algébriques (elles sont d'ailleurs fausses pour les anneaux de Dedekind en général) et leur preuve nécessite des considérations de géométrie des nombres.

Géométrie des nombres.

Si $K = \mathbf{Q}(\alpha)$ et P est le polynôme minimal de α, on a $n = [K : \mathbf{Q}] = \deg(P)$ et P possède r_1 racines réelles $\alpha_1, \ldots, \alpha_{r_1}$ et r_2 paires de racines complexes conjuguées $\alpha_{r_1+1}, \bar{\alpha}_{r_1+1}, \ldots, \alpha_{r_1+r_2}, \bar{\alpha}_{r_1+r_2}$ (donc $n = r_1 + 2r_2$). Les plongements de K dans $\mathbf{R}$ sont donc donnés par $\sigma_i(\alpha) = \alpha_i$ (pour $1 \leq i \leq r_1$) et les plongements complexes (non réels) par $\sigma_{r_1+i}(\alpha) = \alpha_{r_1+i}, \bar{\sigma}_{r_1+i}(\alpha) = \bar{\alpha}_{r_1+i}$ (pour $1 \leq i \leq r_2$). Remarquons qu'on peut montrer que ces plongements, définis ici à partir du choix d'un élément primitif, n'en dépendent en fait pas.

Théorème. (Théorème des unités de Dirichlet) *Soit K un corps de nombres avec r_1 plongements réels et r_2 paires de plongements complexes conjugués, alors le groupe des unités $\mathcal{O}_K^*$ est isomorphe au groupe fini des racines de l'unité de K fois $\mathbf{Z}^r$ avec $r := r_1 + r_2 - 1$.*

Preuve. (Sauf pour les exemples ci-dessous, nous allons seulement prouver le théorème avec $r \leq r_1 + r_2 - 1$). On procède comme pour l'étude des unités de $\mathbf{Q}(\sqrt{d})$: on introduit l'application $L : \mathcal{O}_K^* \to \mathbf{R}^{r_1+r_2}$ définie par :

$$L(\alpha) := (\log|\sigma_1(\alpha)|, \ldots, \log|\sigma_{r_1}(\alpha)|, 2\log|\sigma_{r_1+1}(\alpha)|, \ldots, 2\log|\sigma_{r_1+r_2}(\alpha)|),$$

on montre qu'il n'y a qu'un nombre fini d'élément de $\alpha \in \mathcal{O}_K^*$ tels que $L(\alpha)$ soit dans une boule donnée de $\mathbf{R}^{r_1+r_2}$, qu'ainsi le noyau est fini et est donc constitué des racines de l'unité contenues dans K et que l'image $L(\mathcal{O}_K^*)$ est discrète. On observe que cette image est contenue dans l'hyperplan $x_1 + \ldots + x_{r_1+r_2} = 0$ puisque $\log|\mathrm{N}_{\mathbf{Q}}^K(\alpha)| = 0$. Enfin si on sait qu'un sous-groupe discret de $\mathbf{R}^m$ est isomorphe à $\mathbf{Z}^r$ avec $r \leq m$ on en déduit l'énoncé avec $r \leq r_1 + r_2 - 1$. □

Exemples. Un corps quadratique imaginaire vérifie $r_1 = 0$, $r_2 = 1$ donc $r = 0$, i.e. $\mathcal{O}_K^*$ est fini. Un corps quadratique réel vérifie $r_1 = 2$, $r_2 = 0$ donc $r = 1$, donc $\mathcal{O}_K^* \cong \{\pm 1\} \times \mathbf{Z}$ comme nous l'avons vu. Dans le cas $K = \mathbf{Q}(\sqrt[3]{2})$, on a $r_1 = r_2 = 1$ donc $r \leq 1$, or, en posant $\alpha := \sqrt[3]{2}$ pour alléger, on voit que $(1 + \alpha + \alpha^2)(\alpha - 1) = 1$ donc $1 + \alpha + \alpha^2$ est une unité et $r = 1$. Dans le cas de $K = \mathbf{Q}(\zeta)$ avec $\zeta := \exp(2\pi i/p)$, on voit que $r_1 = 0$, $r_2 = (p-1)/2$ donc $r = (p-3)/2$. On peut montrer directement que les unités $\eta_k := \sin(k\pi/p)/\sin(\pi/p)$ pour $k = 2, \ldots, (p-1)/2$ sont indépendantes et engendrent donc un sous-groupe de rang $(p-3)/2$ donc d'indice fini dans $\mathcal{O}_K^*$.

Pour démontrer la finitude du groupe des classes on utilise le plongement de K dans $E := \mathbf{R}^{r_1} \times \mathbf{C}^{r_2} \cong \mathbf{R}^{[K:\mathbf{Q}]}$ donné par

$$\Phi(\alpha) := (\sigma_1(\alpha), \ldots, \sigma_{r_1}(\alpha), \sigma_{r_1+1}(\alpha), \ldots, \sigma_{r_1+r_2}(\alpha)) .$$

On montre que l'image de $\mathcal{O}_K$ est discrète et, comme $\mathcal{O}_K \cong \mathbf{Z}^n$, c'est donc un réseau de volume disons $D = D_K$. Joint au théorème de Minkowski, ceci permet de montrer :

Lemme. *Il existe $c_1 > 0$ (dépendant de K) tel que si I est un idéal non nul de $\mathcal{O}_K$, il existe un élément non nul $\alpha \in I$ tel que $\left|\mathrm{N}_{\mathbf{Q}}^K(\alpha)\right| \leq c_1 \,\mathrm{N}(I)$.*

Preuve. Soit K_t le compact convexe symétrique de E défini par $|x_i| \leq t$, son volume est proportionnel à t^n (avec $n = [K : \mathbf{Q}]$) ; le réseau $\Phi(I)$ a pour volume $D_K \,\mathrm{N}(I)$ donc si $t^n \geq c2^n D_K \,\mathrm{N}(I)$, on aura $\Phi(I) \cap K_t \neq \{0\}$. On peut donc choisir t^n proportionnel à $\mathrm{N}(I)$ et il existe donc $\alpha \in I$ non nul tel que $\Phi(\alpha) \in K_t$ et donc $\left|\mathrm{N}_{\mathbf{Q}}^K(\alpha)\right| \leq t^n \leq c_1 \,\mathrm{N}(I)$. □

Corollaire. *Il existe $c_1 > 0$ (dépendant de K) tel que toute classe d'idéaux de $\mathcal{O}_K$ contienne un idéal de norme inférieure à c_1.*

Preuve. Soit $\mathcal{C}$ une classe d'idéaux et I un idéal entier dans la classe inverse, d'après le lemme, il existe $\alpha \in I$ tel que $\left|\mathrm{N}_{\mathbf{Q}}^K(\alpha)\right| \leq c_1 \,\mathrm{N}(I)$. Considérons $J := \alpha(I)^{-1}$; c'est un idéal entier dans la classe de $(I)^{-1}$ c'est-à-dire $\mathcal{C}$ et l'on a $\mathrm{N}(J) = \left|\mathrm{N}_{\mathbf{Q}}^K(\alpha)\right| \mathrm{N}(I)^{-1} \leq c_1$. □

Le corollaire entraîne la finitude du groupe des classes en observant qu'il n'y a qu'un nombre fini d'idéaux de norme donnée.

Troisième partie : Théorie analytique des nombres.

A. Enoncés et estimations "élémentaires".
B. Fonctions holomorphes (résumé/rappels).
C. Séries de Dirichlet, fonction $\zeta(s)$.
D. Caractères et théorème de Dirichlet.
E. Le théorème des nombres premiers.

A. Enoncés et estimations "élémentaires".

Donnons tout d'abord quelques énoncés précisant la proposition d'Euclide : "*il existe une infinité de nombres premiers*".

Proposition. *La série de terme* $\log(p)p^{-1}$ *ou* p^{-1} *est divergente ; plus précisément*

$$\sum_{p\leq x}\frac{\log(p)}{p}=\log x+O(1)\qquad\text{et}\qquad\sum_{p\leq x}\frac{1}{p}=\log\log x+C+O\left(\frac{1}{\log x}\right)$$

Ces énoncés peuvent être raffinés avec le

Théorème. (théorème des nombres premiers) *Lorsque* x *tend vers l'infini, on a le comportement asymptotique :*

$$\pi(x):=\mathrm{card}\{p\text{ premier },\ p\leq x\}\sim\frac{x}{\log x}.$$

On peut donner diverses formes équivalentes de ce théorème :

$$\theta(x)\sim x;\qquad\psi(x)\sim x\qquad\text{ou encore}\qquad p_n\sim n\log n$$

où l'on a posé $\theta(x)=\sum_{p\leq x}\log p$, $\psi(x)=\sum_{p^m\leq x}\log p$ et p_n désigne le n-ième nombre premier. Nous allons démontrer également l'énoncé suivant

Théorème. (théorème de la progression arithmétique de Dirichlet) *Soient* $a,b\geq 1$ *premiers entre eux, alors il existe une infinité de nombres premiers* p *de la forme* $a+bn$.

On peut préciser cet énoncé en montrant que les nombres premiers se répartissent grosso modo de manière uniforme entre les classes de congruences modulo b.

Théorème. *Sous les mêmes hypothèses, on a :*

$$\sum_{\substack{p\leq x\\ p\equiv a\bmod b}}\frac{1}{p}=\frac{\log\log x}{\phi(b)}+C+O\left(\frac{1}{\log x}\right)$$

Nous ne démontrerons pas mais énonçons un énoncé un peu plus fin.

Théorème. *Lorsque* x *tend vers l'infini, on a le comportement asymptotique :*

$$\pi(x;a,b):=\mathrm{card}\{p\text{ premier },\ p\leq x,\ p\equiv a\bmod b\}\sim\frac{x}{\phi(b)\log x}.$$

Nous développons maintenant dans ce paragraphe les méthodes dites "élémentaires" (c'est-à-dire dans ce contexte n'utilisant pas la variable complexe) permettant de prouver les assertions précédentes hormis le théorème de la progression arithmétique et le théorème des nombres premiers. On obtient néanmoins une version partielle sous la forme : il existe deux constantes $c_1,c_2>0$ *telles que* $c_1x/\log x\leq\pi(x)\leq c_2x/\log x$.

Lemme. *On a l'estimation* $n\log 2\leq\log C_{2n}^n=\log\binom{2n}{n}\leq n\log 4$.

Preuve. D'abord, par la formule du binôme de Newton, $C_{2n}^n \leq \sum_{k=0}^{2n} C_{2n}^k = (1+1)^{2n} = 4^n$. Ensuite, on peut minorer $C_{2n}^n = \frac{(2n)!}{(n!)^2} = \frac{2n(2n-1)\ldots(n+1)}{n(n-1)\ldots 1} \geq 2^n$. □

Lemme. *On a la formule* $\mathrm{ord}_p(n!) = \sum_{m\geq 1} \left[\frac{n}{p^m}\right]$*; de plus la somme peut être restreinte à* $m \leq \log n/\log p$.

Preuve. Ecrivons $n! = 1.2.3\ldots n = \prod_{k=1}^n k$. Le nombre d'entiers $\leq n$ divisibles par p est $[n/p]$; le nombre d'entiers $\leq n$ divisibles par p^2 est $[n/p^2]$; etc. L'ordre en p de $n!$ est donc bien la somme des $[n/p^m]$. Enfin, $p^m \leq n$ équivaut à $m \leq \log n/\log p$, d'où la dernière affirmation. □

On peut ainsi écrire

$$\log C_{2n}^n = \sum_{p\leq 2n} \mathrm{ord}_p\left(C_{2n}^n\right) \log p = \sum_{p\leq 2n} \left(\sum_{m\geq 1} \left[\frac{2n}{p^m}\right] - 2\left[\frac{n}{p^m}\right] \right) \log p$$

Pour écrire une minoration, on ne garde que les termes avec $n < p \leq 2n$; en effet on observe qu'un tel p divise clairement $C_{2n}^n = (2n)!/(n!)^2$ et on obtient ainsi

$$n\log 4 \geq \log C_{2n}^n \geq \sum_{n<p\leq 2n} \log p = \theta(2n) - \theta(n).$$

On en tire une majoration du type $\theta(x) \leq Cx$. En effet

$$\theta(2^m) = \sum_{k=0}^{m-1} \theta(2^{k+1}) - \theta(2^k) \leq \sum_{k=0}^{m-1} 2^k \log 4 = (2^m - 1)\log 4$$

donc, si $2^m \leq x < 2^{m+1}$ on obtient

$$\theta(x) \leq \theta(2^{m+1}) \leq 2^{m+1}\log 4 \leq x2\log 4.$$

Pour écrire une minoration on remarque que $[2u] - 2[u]$ vaut toujours 1 ou 0 et vaut zéro dès que $u < 1/2$. Ainsi

$$n\log 2 \leq \log C_{2n}^n = \sum_{p\leq 2n} \left(\sum_{m\geq 1} \left[\frac{2n}{p^m}\right] - 2\left[\frac{n}{p^m}\right] \right) \log p \leq \sum_{p\leq 2n} \left(\frac{\log(2n)}{\log p}\right) \log p = \log(2n)\pi(2n).$$

On en tire une minoration du type $\pi(x) \geq Cx/\log x$. En effet, si $2n \leq x < 2(n+1)$ alors

$$\pi(x) \geq \pi(2n) \geq \frac{n\log 2}{\log(2n)} \geq \left(\frac{x}{2} - 1\right)\frac{\log 2}{\log x}.$$

Par ailleurs on a facilement

$$\theta(x) = \sum_{p\leq x} \log p \leq \log x \sum_{p\leq x} 1 = \pi(x)\log x.$$

Ensuite notons que pour $2 \leq y < x$

$$\pi(x) - \pi(y) = \sum_{y<p\leq x} 1 \leq \frac{1}{\log y} \sum_{y<p\leq x} \log p = \frac{1}{\log y}\left(\theta(x) - \theta(y)\right)$$

On en tire

$$\pi(x) \leq \frac{\theta(x)}{\log y} + \pi(y) \leq \frac{\theta(x)}{\log y} + y.$$

En choisissant $y = x/(\log x)^2$ on obtient donc en rappelant l'inégalité précédente

$$\frac{\theta(x)}{\log x} \leq \pi(x) \leq \frac{\theta(x)}{\log x + 2\log\log x} + \frac{x}{(\log x)^2}$$

En résumé on tire facilement des inégalités précédentes que $\theta(x) \sim x$ équivaut à $\pi(x) \sim x/\log x$ et que $C_1 x \leq \theta(x) \leq C_2 x$ et $C_3 x/\log x \leq \pi(x) \leq C_4/\log x$. Par ailleurs la comparaison entre la fonction $\theta(x)$ et la fonction $\psi(x)$ est aisée :

$$\theta(x) \leq \psi(x) := \sum_{p^m \leq x} \log p = \theta(x) + \theta(\sqrt{x}) + \theta(\sqrt[3]{x}) + \ldots \leq \theta(x) + \log(x)\theta(\sqrt{x}) \leq \theta(x) + C\log x\sqrt{x}.$$

Enfin, si l'on note p_n le n-ième nombre premier, on a $\pi(p_n) = n$ par définition. Le théorème des nombres premiers implique donc que $n \sim p_n \log(p_n)$ ou encore que $p_n \sim n\log n$. On vérifie que ce dernier énoncé est en fait équivalent au théorème des nombres premiers.

Lemme. (Formule d'Abel) Soit $A(x) := \sum_{n\leq x} a_n$ et f une fonction de classe $\mathcal{C}^1$, on a :

$$\sum_{y<n\leq x} a_n f(n) = A(x)f(x) - A(y)f(y) - \int_y^x A(t)f'(t)dt.$$

Preuve. On remarque que $\int_n^{n+1} A(t)f'(t)dt = A(n)\int_n^{n+1} f'(t)dt = A(n)\left(f(n+1) - f(n)\right)$. Ainsi, si on pose $N = [x]$ et $M = [y]$, on a

$$\begin{aligned}
-\int_M^N A(t)f'(t)dt = -\sum_{n=M}^{N-1}\int_n^{n+1} A(t)f'(t)dt &= \sum_{n=M}^{N-1} A(n)\left(f(n) - f(n+1)\right) \\
&= \sum_{n=M+1}^{N} f(n)\left(A(n) - A(n-1)\right) - f(N)A(N) + A(M)f(M)\,. \\
&= \sum_{n=M+1}^{N} f(n)a_n - f(N)A(N) + A(M)f(M)
\end{aligned}$$

D'où la formule, quand x et y sont entiers. Pour la formule générale on observe que $-\int_{[x]}^x A(t)f'(t)dt = A([x])\left(f(x) - f([x])\right) = A(x)\left(f(x) - f([x])\right)$. □

Applications. 1) La formule permet une comparaison assez précise entre "somme" et "intégrale" ; de façon plus précise, si on prend $a_n = 1$ et si on fait une intégration par parties on obtient :

$$\sum_{n=M+1}^{N} f(n) = \int_M^N f(t)dt + \int_M^N (t - [t])f'(t)dt.$$

Pour le choix de $f(t) = 1/t$ on obtient

$$\sum_{n=1}^{N}\frac{1}{n} = 1 + \int_1^N \frac{dt}{t} - \int_1^N (t-[t])\frac{dt}{t^2} = \log N + \left(1 - \int_1^\infty (t-[t])\frac{dt}{t^2}\right) + \int_N^\infty (t-[t])\frac{dt}{t^2} = \log N + \gamma + O\left(\frac{1}{N}\right),$$

où $\gamma := 1 - \int_1^\infty (t-[t])\frac{dt}{t^2}$ est appelée *constante d'Euler*.
2) Prenons $a_n = 1$ donc $A(t) = [t]$, $y = 1$ et $f(t) = \log t$, on obtient alors

$$\log\left([x]!\right) = [x]\log(x) - \int_1^x \frac{[t]dt}{t} = x\log x - \int_1^x dt + ([x] - x)\log x - \int_1^x \frac{[t]-t}{t}dt = x\log x - x + O(\log x).$$

Remarque. La formule de Stirling donne un énoncé légèrement plus précis, à savoir : $n! \sim n^n e^{-n}\sqrt{2\pi n}$ et donc $\log(n!) = n\log n - n + \frac{1}{2}\log n + \frac{1}{2}\log(2\pi) + \epsilon(n)$ avec $\lim_{n\to\infty}\epsilon(n) = 0$.

Par ailleurs on voit que

$$\begin{aligned}
\log\left([x]!\right) &= \sum_{p\leq x} \mathrm{ord}_p\left([x]!\right)\log p \\
&= \sum_{p\leq x}\sum_{m\geq 1}\left[\frac{x}{p^m}\right]\log p \\
&= x\sum_{p\leq x}\frac{\log p}{p} + \sum_{p\leq x}\log p\left(\left[\frac{x}{p}\right] - \frac{x}{p}\right) + \sum_{p\leq x}\sum_{m\geq 2}\left[\frac{x}{p^m}\right]\log p \\
&= x\sum_{p\leq x}\frac{\log p}{p} + O(x)
\end{aligned}$$

où la dernière estimation provient de la majoration $\theta(x) = \sum_{p\leq x}\log p = O(x)$ et de l'estimation

$$\sum_{p\leq x}\sum_{m\geq 2}\left[\frac{x}{p^m}\right]\log p \leq x\sum_{p\leq x}\sum_{m\geq 2}\frac{\log p}{p^m} = x\sum_{p\leq x}\frac{\log p}{p(p-1)} = O(x).$$

On en tire la première formule cherchée

$$\sum_{p\leq x}\frac{\log p}{p} = \log x + O(1).$$

Pour la deuxième on applique la formule d'Abel avec $f(t) = 1/\log t$ et $a_n = \log p/p$ si $n = p$ est premier et $a_n = 0$ sinon; on obtient, en posant $A(x) = \sum_{p\leq x}\frac{\log p}{p}$, que :

$$\begin{aligned}
\sum_{p\leq x}\frac{1}{p} &= \sum_{n\leq x} a_n f(n) \\
&= \frac{A(x)}{\log x} + \int_2^x \frac{A(t)dt}{t(\log t)^2} \\
&= 1 + O(1/\log x) + \int_2^x \frac{dt}{t\log t} + O\left(\int_2^x \frac{dt}{t(\log t)^2}\right) \\
&= \log\log x + C + O(1/\log x).
\end{aligned}$$

B. Fonctions holomorphes (résumé/rappels).

(Paragraphe sans preuve)

Concernant les séries, nous utiliserons les règles de calcul du produit de deux séries *absolument* convergente :

$$\left(\sum_{n=0}^{\infty} a_n\right)\left(\sum_{n=0}^{\infty} b_n\right) = \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} a_k b_{n-k}\right),$$

ainsi que la règle d'interversion de sommation avec des termes $a_{m,n}$ *positifs* :

$$\sum_{n=0}^{\infty}\left(\sum_{m=0}^{\infty} a_{m,n}\right) = \sum_{m=0}^{\infty}\left(\sum_{n=0}^{\infty} a_{m,n}\right)$$

On sait qu'une série entière $S(z) = \sum_{n=0}^{\infty} a_n z^n$ possède un rayon de convergence, disons $R \geq 0$ (éventuellement $R = 0$ ou $R = +\infty$) tel que la série converge pour tout $|z| < R$ et diverge pour tout $|z| > R$; de plus la convergence est absolue à l'intérieur du disque de convergence et la fonction est de classe $\mathcal{C}^\infty$ avec $S^{(k)}(z) = \sum_{n=k}^{\infty} n(n-1)\dots(n-k+1)a_n z^{n-k}$. En fait la fonction S est de nouveau développable en série entière autour de chaque point $z_0 \in D(0,R)$, c'est-à-dire que pour $z \in D(z_0, r) \subset D(0,R)$ on a $S(z) = \sum_{n=0}^{\infty} b_n (z - z_0)^n$ (avec en fait $b_n = S^{(n)}(z_0)/n!$). Une telle fonction ne possède qu'un nombre fini de zéros dans chaque disque fermé (ou compact) contenu dans $D(0,R)$. On peut définir la multiplicité d'un zéro z_0 comme l'entier k tel que $S(z) = (z - z_0)^k \sum_{n=0}^{\infty} b_n (z - z_0)^n$ avec $b_0 \neq 0$. Une fonction représentable par une série entière au voisinage de chaque point est dite *analytique*.

On définit une fonction *holomorphe* comme une fonction admettant une dérivée (complexe) en chaque point, i.e.

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} = f'(z_0) \in \mathbf{C} \quad \text{existe.}$$

Bien sûr les notions de "dérivable" et de "analytique" sont très différentes en variable réelle; en variable complexe, elles sont par contre équivalentes :

Proposition. *Soit $f : U \to \mathbf{C}$ une fonction holomorphe, et supposons que $D(z_0, r) \subset U$ alors pour tout $z \in D(z_0, r)$, on a $f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n$, avec $a_n = f^{(n)}(z_0)/n!$.*

Proposition. *Soit $f : U \to \mathbf{C}$ une fonction holomorphe, supposons U connexe et f non identiquement nulle, alors l'ensemble des zéros de f est discret dans U.*

Corollaire. *Soit $f, g : U \to \mathbf{C}$ deux fonctions holomorphes, supposons U connexe, alors, si l'ensemble $\{z \in U \mid f(z) = g(z)\}$ n'est pas discret dans U on a $f = g$. En particulier, une fonction holomorphe sur un disque $D(z_0, r) \subset U$ admet au plus un prolongement holomorphe à U.*

On définit ensuite les fonctions *méromorphes* comme des fonctions holomorphes sur un ouvert U sauf en les *pôles* disons z_0 où elles ont le comportement suivant : il existe un entier m, appelé *l'ordre du pôle* tel que la fonction $(z - z_0)^m f(z)$ admet un prolongement holomorphe au voisinage de z_0 et non nul en z_0. Il revient au même de demander que $f(z)$ s'écrive, au voisinage de z_0 :

$$f(z) = \frac{a_m}{(z - z_0)^m} + \frac{a_{m-1}}{(z - z_0)^{m-1}} + \dots + \frac{a_1}{z - z_0} + \text{ fonction holomorphe en } z_0.$$

Le coefficient a_1 s'appelle le *résidu* de f en z_0 et sera noté $\text{Rés}(f; z_0)$. Son importance provient de son utilité dans le calcul d'intégrales.

On définit l'intégrale le long d'un contour ainsi : pour $\gamma : [a, b] \to \mathbf{C}$ de classe $\mathcal{C}^1$ on pose

$$\int_\gamma f(z)dz := \int_a^b f(\gamma(t))\gamma'(t)dt.$$

La formule de changement de variables montre que la valeur de l'intégrale ne dépend pas de la paramétrisation du chemin mais dépend du sens selon lequel on parcourt le contour. On appellera par commodité *contour simple* un chemin $\gamma : [a, b] \to \mathbf{C}$ tel que $\gamma(a) = \gamma(b)$ mais γ injectif sur $[a, b[$ et tournant dans le sens trigonométrique. On notera qu'un tel contour partage le plan en deux parties connexes (l'intérieur et l'extérieur).

Théorème. (théorème des résidus) *Soit f une fonction méromorphe sur U; soit γ un contour simple évitant les pôles de f et soit S l'ensemble des pôles de f situés à l'intérieur de γ alors :*

$$\int_\gamma f(z)dz = 2\pi i \sum_{a \in S} \text{Rés}(f; a).$$

En particulier, on voit que si U est *simplement connexe*, i.e. "sans trous", alors, si f est holomorphe sur U et γ_1, γ_2 sont deux chemins dans U joignant a et b alors $\int_{\gamma_1} f(z)dz = \int_{\gamma_2} f(z)dz$. Ainsi on peut définir

une primitive d'une fonction holomorphe $f(z)$ sur un tel ouvert par la formule $F(b) = \int_\gamma f(z)dz$ où γ est un chemin dans U joignant a et b.

Proposition. *Soient $f_n(z)$ une suite de fonctions holomorphe de U vers $\mathbf{C}$; supposons que la suite converge uniformément sur tout compact de U vers une fonction f, alors f est holomorphe et les dérivées $f_n^{(k)}$ converge uniformément sur tout compact de U vers la fonction $f^{(k)}$.*

Explicitons ceci sur l'exemple des séries et produits infinis. Soit $u_n(z)$ une suite de fonction holomorphes telles que la série $S(z) := \sum_{n=0}^{\infty} u_n(z)$ soit convergente; supposons de plus la convergence uniforme sur tout compact - c'est-à-dire que $\left|\sum_{n=M}^{N} u_n(z)\right| \leq \epsilon(M)$ avec $\epsilon(M)$ tendant vers 0 quand M tend vers l'infini. Alors la fonction $S(z)$ est holomorphe et

$$S^{(k)}(z) = \sum_{n=0}^{\infty} u_n^{(k)}(z).$$

De même si maintenant on suppose que $\prod_{n=0}^{\infty} u_n(z)$ converge uniformément vers $P(z)$ sur tout compact de U - c'est-à-dire que $\left|P(z) - \prod_{n=0}^{N} u_n(z)\right| \leq \epsilon(N)$avec $\epsilon(N)$ tendant vers 0 quand N tend vers l'infini - alors $P(z)$ est holomorphe sur U.

Exemple. (logarithme complexe). La fonction $\exp(z) = e^z = \sum_{n=0}^{\infty} z^n/n!$ est holomorphe sur $U = \mathbf{C}$ (par exemple la convergence de la série est uniforme sur tout disque de centre 0 et de rayon R). Définissons

$$F(z) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{(z-1)^n}{n} = -\sum_{n=1}^{\infty} \frac{(1-z)^n}{n}.$$

La série est (absolument) convergente sur le disque $D(1,1) = \{z \in \mathbf{C} \mid |1-z| < 1\}$ et la convergence uniforme sur le disque de centre 1 et rayon $r < 1$ donc $F(z)$ est holomorphe. Si z est réel dans l'intervalle $]0,2[$ on voit que $F(z) = \log z$ (logarithme usuel) et en particulier on a

$$\exp(F(z)) = z.$$

La formule précédente indique que les deux fonctions identité et $\exp \circ F$, analytiques sur le disque $D(1,1)$ coïncident sur le segment $]0,2[$ et donc sur le disque entier. Ainsi F définit un logarithme complexe sur le disque $|z-1| < 1$.

Définition. Soit $f(z)$ une fonction holomorphe sur U, on dit que $F(z)$ est une *détermination du logarithme* de f sur U (on écrira alors un peu abusivement $F(z) = \log f(z)$) si, d'une part $F(z)$ est holomorphe et, d'autre part $\exp(F(z)) = z$.

Remarques. Si $F(z)$ est un logarithme, alors f ne s'annule pas sur U et on a $|\exp(F(z))| = \exp(\Re F(z)) = |z|$ donc

$$\Re \log f(z) = \log |f(z)|.$$

De même $f'(z)/f(z) = F'(z) \exp(F(z)) / \exp(F(z)) = F'(z)$ ou encore

$$\frac{d}{dz} \log f(z) = \frac{f'(z)}{f(z)}.$$

Enfin, si F_1 et F_2 sont deux logarithmes, alors $F_2(z) = F_1(z) + 2k\pi i$ sur U connexe.

Ces remarques suggèrent de construire le logarithme de $f(z)$ comme une primitive de $f'(z)/f(z)$, sous la condition que f ne s'annule pas. On a vu que cela est possible si U est simplement connexe.

Proposition. *Soit U ouvert simplement connexe du plan complexe et $f(s)$ holomorphe sans zéro sur U alors il existe une détermination holomorphe $F(s) = \log f(s)$ sur U; deux telles déterminations diffèrent d'un multiple entier de $2\pi i$.*

C. Séries de Dirichlet, fonction $\zeta(s)$.

On appelle *série de Dirichlet* une série de la forme $F(s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$. La première propriété importante est donnée par

Proposition. *Soit $F(s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$ une série de Dirichlet, supposons qu'elle soit convergente en s_0 alors elle converge uniformément sur tout ensemble de type $\{s \in \mathbf{C} \mid \Re(s-s_0) \geq 0, |s-s_0| \leq C\Re(s-s_0)\}$.*

Preuve. Soit $M \geq 1$ et posons $A(x) := \sum_{M \leq n \leq x} a_n n^{-s_0}$; d'après les hypothèses on a donc $|A(x)| \leq \epsilon(M)$ avec $\epsilon(M)$ tendant vers zéro quand M tend vers l'infini. Appliquons la formule d'Abel

$$\sum_{M<n\leq N} a_n n^{-s} = \sum_{M<n\leq N} a_n n^{-s_0} n^{-(s-s_0)} = A(N)N^{-(s-s_0)} - A(M)M^{-(s-s_0)} + (s-s_0)\int_M^N A(t)t^{-(s-s_0+1)}dt$$

et majorons l'intégrale ainsi

$$\left|\int_M^N A(t)t^{-(s-s_0+1)}dt\right| \leq \epsilon(M)\int_M^N t^{-(\sigma-\sigma_0+1)}dt = \epsilon(M)\frac{M^{-(\sigma-\sigma_0)} - N^{-(\sigma-\sigma_0)}}{(\sigma-\sigma_0)}$$

En se plaçant sur un secteur angulaire délimité par $\sigma - \sigma_0 = \Re(s) - \Re(s_0) \geq 0$ et $|s-s_0| \leq C(\sigma-\sigma_0)$ on obtient

$$\left|\sum_{M<n\leq N} a_n n^{-s}\right| \leq \epsilon(M)(2+C)$$

ce qui suffit à montrer la convergence uniforme sur ce secteur (Cf dessin ci-dessous). □

En appliquant les théorèmes généraux du paragraphe précédent on obtient :

Corollaire. *Toute série de Dirichlet $F(s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$ possède une abcisse de convergence, disons σ_0 telle que la série converge pour $\Re(s) > \sigma_0$ et diverge pour $\Re(s) < \sigma_0$. De plus la fonction F définie par la série est holomorphe dans le demi-plan de convergence $\Re(s) > \sigma_0$ et ses dérivées s'expriment comme $F^{(k)}(s) = \sum_{n=1}^{\infty} a_n(-\log n)^k n^{-s}$.*

Preuve. Il suffit de poser $\sigma_0 = \inf\{\sigma \in \mathbf{R}$ la série converge en $\sigma\}$, puis d'observer que tout compact du demi-plan (ouvert) de convergence est contenu dans un secteur comme ci-dessus, où la convergence est uniforme. □

On peut aussi tirer de la démonstration précédente la formule (où l'on pose $A(t) := \sum_{n\leq t} a_n$) :

$$\sum_{n=1}^{\infty} a_n n^{-s} = s\int_1^{\infty}\left(\sum_{n\leq t} a_n\right)t^{-s-1}dt = s\int_1^{\infty} A(t)t^{-s-1}dt$$

qui montre en particulier la convergence pour $\Re(s) > 0$ si $A(t) = \sum_{n\leq t} a_n$ est bornée.

La plus célèbre série de Dirichlet est la *fonction zêta* de Riemann définie par la série $\sum_{n=1}^{\infty} \frac{1}{n^s}$. Il est bien connu, au moins pour les valeurs réelles, mais le cas général en découle, que l'abscisse de convergence est +1.

Théorème. (Produit d'Euler) *Pour $\Re(s) > 1$, on a la formule*

$$\zeta(s) = \sum_{n=1}^{\infty}\frac{1}{n^s} = \prod_p\left(1-\frac{1}{p^s}\right)^{-1}$$

Preuve. Pour $\Re(s) > 0$, la convergence d'une série géométrique donne $(1-p^{-s})^{-1} = \sum_{m=0}^{\infty} p^{-ms}$ et donc en faisant le produit pour $p_1, \dots, p_r$ les nombres premiers $\leq T$ on obtient

$$\prod_{p \leq T} (1-p^{-s})^{-1} = \prod_{p \leq T} \left(\sum_{m=0}^{\infty} p^{-ms} \right) = \sum_{m_1, \dots, m_r \geq 1} (p_1^{m_1} \dots p_r^{m_r})^{-s} = \sum_{n \in \mathcal{N}(T)} n^{-s}$$

où l'on a noté $\mathcal{N}(T)$ l'ensemble des entiers dont tous les facteurs premiers sont $\leq T$. Ainsi, lorsque $\Re(s) > 1$, on a

$$\left| \sum_{n=1}^{\infty} \frac{1}{n^s} - \prod_{p \leq T} \left(1 - \frac{1}{p^s}\right)^{-1} \right| = \left| \sum_{n \notin \mathcal{N}(T)} \frac{1}{n^s} \right| \leq \sum_{n > T} \left| \frac{1}{n^s} \right| = \sum_{n > T} \frac{1}{n^\sigma}.$$

La dernière somme est la queue d'une série réelle convergente (quand $\sigma := \Re(s) > 1$) et tend donc vers zéro, ce qui prouve à la fois la convergence du produit et la formule d'Euler. □

Corollaire. *La fonction $\zeta(s)$ ne s'annule pas dans le demi-plan $\Re(s) > 1$. On peut construire une détermination holomorphe de $\log \zeta(s)$ pour $\Re(s) > 1$ en posant :*

$$\log \zeta(s) = \sum_p \sum_{m \geq 1} \frac{p^{-ms}}{m}$$

En posant

$$\Lambda(n) = \begin{cases} \log p & \text{si } n = p^m \\ 0 & \text{sinon} \end{cases}$$

on peut aussi écrire

$$-\frac{\zeta'(s)}{\zeta(s)} = \sum_p \sum_{m \geq 1} \frac{\log p}{p^{ms}} = \sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s}$$

On a $1 - p^{-s} \neq 0$ et le produit est convergent donc la première affirmation est claire. La première formule se déduit de la formule d'Euler en prenant les logarithmes et en utilisant le développement en série

$$\log\left((1-x)^{-1}\right) = \sum_{m=1}^{\infty} \frac{x^m}{m}.$$

La deuxième s'obtient en dérivant la première. □

Intermède (I). Ces formules se généralisent en remplaçant $\mathbf{Z}, \mathbf{Q}$ par $\mathcal{O}_K, K$ et l'unicité de la décomposition en facteurs premiers par l'unicité de la décomposition en idéaux premiers. Notons $\mathcal{I}_K$ l'ensemble des idéaux non nuls et $\mathcal{P}_K$ l'ensemble des idéaux premiers non nuls (maximaux) de $\mathcal{O}_K$ alors on peut introduire la fonction zêta de Dedekind et démontrer :

$$\zeta_K(s) := \sum_{I \in \mathcal{I}_K} \mathrm{N}(I)^{-s} = \prod_{\wp \in \mathcal{P}_K} \left(1 - \mathrm{N}(\wp)^{-s}\right)^{-1} \qquad \text{pour } \Re(s) > 1.$$

Proposition. *La fonction $\zeta(s)$ se prolonge en une fonction méromorphe sur le demi-plan $\Re(s) > 0$ avec un unique pôle en $s = 1$ et résidu $+1$.*

Preuve. L'énoncé affirme que la fonction $\zeta(s) - 1/(s-1)$, définie au départ pour $\Re(s) > 1$, admet un prolongement holomorphe au demi-plan $\Re(s) > 0$. Pour cela écrivons à l'aide de $[t] = \sum_{n \leq t} 1$:

$$\begin{aligned} \zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} &= s \int_1^{\infty} [t] t^{-s-1} dt \\ &= s \int_1^{\infty} t^{-s} dt + s \int_1^{\infty} ([t] - t)\, t^{-s-1} dt \\ &= \frac{1}{s-1} + 1 + s \int_1^{\infty} ([t] - t)\, t^{-s-1} dt \end{aligned}$$

Or $|[t]-t| \leq 1$ donc la dernière intégrale est convergente et définit une fonction holomorphe pour $\Re(s) > 0$. □

Remarque. En fait la fonction $\zeta(s) - 1/(s-1)$ peut se prolonger à tout le plan complexe et $\zeta(s)$ vérifie de plus la célèbre équation fonctionnelle démontrée par Riemann, qui s'écrit $\xi(s) = \xi(1-s)$, en posant $\xi(s) := \pi^{-s/2}\Gamma(s/2)\zeta(s)$.

Exercice. Rappelons que $(f * g)(n) = \sum_{d|n} f(d)g(n/d)$, montrer que, si les deux séries de Dirichlet $F(s) = \sum_{n=1}^{\infty} f(n)n^{-s}$ et $G(s) = \sum_{n=1}^{\infty} g(n)n^{-s}$ convergent absolument pour $\Re(s) > \sigma_0$, alors, dans le même demi-plan, on a :

$$F(s)G(s) = \left(\sum_{n=1}^{\infty} f(n)n^{-s}\right)\left(\sum_{n=1}^{\infty} g(n)n^{-s}\right) = \sum_{n=1}^{\infty}(f*g)(n)n^{-s}.$$

En particulier montrer que, pour $\Re(s) > 1$ on a

$$\zeta(s)^{-1} = \sum_{n=1}^{\infty} \mu(n)n^{-s}, \qquad \text{où } \mu \text{ est la fonction de Moebius}$$

(i.e. $\mu(1) = 1$, $\mu(p_1 \dots p_k) = (-1)^k$ et $\mu(n) = 0$ si n possède un facteur carré).

D. Caractères et théorème de Dirichlet.

Définition Si G est un groupe fini abélien, on appellera *caractère* un homomorphisme de G vers $\mathbf{C}^*$. L'ensemble des caractères de G forme un groupe qui sera noté $\hat{G}$.

Proposition. *Le groupe $\hat{G}$ est isomorphe (non canoniquement) au groupe G.*

Preuve. Si $G = \mathbf{Z}/n\mathbf{Z}$, un caractère vérifie $\chi(1) \in \mu_n$ (où μ_n désigne comme d'habitude le groupe des racines n-ièmes de l'unité) et l'application $\chi \mapsto \chi(1)$ fournit un isomorphisme entre $\hat{G}$ et μ_n, or ce dernier est isomorphe à $\mathbf{Z}/n\mathbf{Z}$ donc à G. Montrons ensuite que $\widehat{G_1 \times G_2} \cong \hat{G_1} \times \hat{G_2}$; en effet un caractère χ de $G_1 \times G_2$ s'écrit $\chi(g_1, g_2) = \chi(g_1, e_2)\chi(e_1, g_2)$ et, en posant $\chi_1 = \chi(., e_2)$ et $\chi_2 = \chi(e_1, .)$ on obtient un isomorphisme $\chi \mapsto (\chi_1, \chi_2)$ de $\widehat{G_1 \times G_2}$ vers $\hat{G_1} \times \hat{G_2}$. Le cas général est maintenant facile : on a $G \cong \mathbf{Z}/n_1\mathbf{Z} \times \ldots \times \mathbf{Z}/n_r\mathbf{Z}$ donc

$$\hat{G} \cong \widehat{(\mathbf{Z}/n_1\mathbf{Z} \times \ldots \times \mathbf{Z}/n_r\mathbf{Z})} \cong \widehat{\mathbf{Z}/n_1\mathbf{Z}} \times \ldots \times \widehat{\mathbf{Z}/n_r\mathbf{Z}} \cong \mathbf{Z}/n_1\mathbf{Z} \times \ldots \times \mathbf{Z}/n_r\mathbf{Z} \cong G.$$

□

Lemme. *Soit x un élément d'ordre r dans G alors pour chaque ξ racine r-ième de l'unité, il existe $|G|/r$ caractères tels que $\chi(x) = \xi$. En particulier on a la formule :*

$$\prod_{\chi \in \hat{G}} (1 - \chi(x)T) = (1 - T^r)^{|G|/r} \qquad \textit{dans } \mathbf{C}[T].$$

Preuve. On voit immédiatement que $\chi(x) \in \mu_r$ puisque $\chi(x)^r = \chi(x^r) = \chi(e_G) = 1$. Considérons l'application $\chi \mapsto \chi(x)$ de $\hat{G}$ vers μ_r; c'est un homomorphisme dont nous allons identifier le noyau. Soit H le sous-groupe engendré par x (donc $H \cong \mathbf{Z}/r\mathbf{Z}$); le noyau de l'homomorphisme précédent est constitué des caractères tels que $\chi(x) = 1$ ou encore des caractères triviaux sur H. Ces derniers sont en bijection avec les caractères de G/H et leur cardinal est donc $\mathrm{card}(G/H) = |G|/r$. On voit que l'image a pour cardinal r donc que l'homomorphisme est surjectif ce qui achève la preuve de la première partie du lemme. Pour la dernière formule on remarque que

$$\prod_{\chi \in \hat{G}} (1 - \chi(x)T) = \left(\prod_{\xi \in \mu_r} (1 - \xi T)\right)^{|G|/r} = (1 - T^r)^{|G|/r}$$

□

Proposition. *Soit G un groupe fini commutatif, on a les relations suivantes :*

$$\forall g \in G \setminus \{e\},\ \sum_{\chi \in \hat{G}} \chi(g) = 0 \qquad \text{et} \qquad \forall \chi \in G \setminus \{1\},\ \sum_{g \in G} \chi(g) = 0.$$

Preuve. Si $g = e$ on a clairement $\sum_{\chi \in \hat{G}} \chi(g) = |G|$; si $g \neq e$ d'après le lemme précédent, il existe χ_1 avec $\chi_1(g) \neq 1$ alors

$$\sum_{\chi \in \hat{G}} \chi(g) = \sum_{\chi \in \hat{G}} (\chi\chi_1)(g) = \chi_1(g) \sum_{\chi \in \hat{G}} \chi(g),$$

d'où le résultat. On procède de même avec l'autre somme en observant que, si $\chi = 1$ alors $\sum_{g \in G} \chi(g) = |G|$ et, si $\chi \neq 1$, il existe g_1 tel que $\chi(g_1) \neq 1$ donc

$$\sum_{g \in G} \chi(g) = \sum_{g \in G} \chi(gg_1) = \chi(g_1) \sum_{g \in G} \chi(g)$$

d'où la deuxième formule. □

Corollaire. *Soit $a \in G$ alors*

$$\frac{1}{|G|} \sum_{\chi \in \hat{G}} \overline{\chi}(a)\chi(x) = \begin{cases} 1 & \text{si } x = a \\ 0 & \text{sinon} \end{cases}$$

Preuve. Cela découle des relations précédentes en observant que, puisque $\chi(a)$ est une racine de l'unité, $\overline{\chi}(a) = \chi(a)^{-1} = \chi(a^{-1})$ et donc $\sum_{\chi \in \hat{G}} \overline{\chi}(a)\chi(x) = \sum_{\chi \in \hat{G}} \chi(a^{-1}x)$ vaut $|G|$ si $x = a$ et 0 sinon. □

Définition Soit $\chi : \mathbf{Z}/m\mathbf{Z}^* \to \mathbf{C}^*$ un caractère de $\mathbf{Z}/m\mathbf{Z}^*$, on appellera *caractère de Dirichlet* modulo m (et on notera encore χ) l'application de $\mathbf{Z}$ vers $\mathbf{C}$ définie par

$$\chi(n) = \begin{cases} \chi(n \bmod m) & \text{si } \operatorname{pgcd}(m,n) = 1 \\ 0 & \text{si } \operatorname{pgcd}(m,n) > 1 \end{cases}$$

Remarque. On garde la propriété de mutiplicativité, c'est-à-dire que, $\forall n, n' \in \mathbf{Z},\ \chi(nn') = \chi(n)\chi(n')$.

On va utiliser ces caractères de la façon suivante : on a l'égalité (au moins formellement)

$$\sum_{p \equiv a \bmod m} f(p) = \frac{1}{\phi(m)} \sum_{\chi} \sum_{p} \overline{\chi}(a)\chi(p)f(p) = \frac{1}{\phi(m)} \sum_{p \not| m} f(p) + \frac{1}{\phi(m)} \sum_{\chi \neq 1} \overline{\chi}(a) \left(\sum_{p} \chi(p)f(p) \right).$$

On a déjà traité des sommes comme le premier terme $\sum_p f(p)$; pour pouvoir traiter les sommes du type $\sum_p \chi(p)f(p)$, on va intoduire les séries suivantes.

Définition Soit χ, un caractère de Dirichlet modulo m, on définit la série "L" de Dirichlet par

$$L(\chi, s) := \sum_{n=1}^{\infty} \chi(n)n^{-s}.$$

Remarque. Si χ_0 est le caractère unité modulo n, on a $\chi_0(n) = 1$ ou 0 suivant que n est premier avec n ou non. On en déduit aisément que $L(\chi_0, s)$ est presque égal à la fonction $\zeta(s)$; plus précisément :

$$L(\chi_0, s) = \sum_{\operatorname{pgcd}(n,m)=1} n^{-s} = \prod_{p \not| m} (1 - p^{-s}) = \prod_{p|m} (1 - p^{-s})\zeta(s).$$

Proposition. *L'abscisse de convergence de la série $L(\chi, s)$ est $\sigma = 0$ sauf si χ est le caractère unité, auquel cas $\sigma = 1$.*

Preuve. D'après la remarque précédente, la série $L(\chi_0, s)$ pour χ_0 égal au caractère unité a la même abscisse de convergence que la série définissant la fonction zêta, c'est-à-dire 1. Le terme général de la série ne tend pas vers zéro si $\Re(s) \leq 0$ donc la série ne peut pas converger. Si χ est un caractère modulo m qui n'est pas le caractère unité, alors $\sum_{n=r+1}^{r+m} \chi(n) = 0$ et donc

$$\left|\sum_{n\leq x} \chi(n)\right| = \left|\sum_{m\left[\frac{x}{m}\right]<n\leq x} \chi(n)\right| \leq m$$

et on a vu qu'alors la série de Dirichlet converge pour $\Re(s) > 0$. Remarquons que l'abscisse de convergence absolu est 1 et est strictement supérieur à l'abscisse de convergence dans ce cas. □

Théorème. *On a la formule d'Euler généralisée :*

$$L(\chi, s) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} = \prod_p \left(1 - \frac{\chi(p)}{p^s}\right)^{-1} \qquad \text{pour } \Re(s) > 1.$$

Pour $\Re(s) > 0$, la convergence d'une série géométrique donne $(1 - \chi(p)p^{-s})^{-1} = \sum_{m=0}^{\infty} \chi(p)^m p^{-ms}$ et donc en faisant le produit pour $p_1, \ldots, p_r$ les nombres premiers $\leq T$ on obtient

$$\prod_{p\leq T} (1 - \chi(p)p^{-s})^{-1} = \prod_{p\leq T} \left(\sum_{m=0}^{\infty} \chi(p)^m p^{-ms}\right) = \sum \chi(p_1)^{m_1} \ldots \chi(p_r)^{m_r} (p_1^{m_1} \ldots p_r^{m_r})^{-s} = \sum_{n\in\mathcal{N}(T)} \chi(n) n^{-s}$$

où l'on a noté $\mathcal{N}(T)$ les entiers dont tous les facteurs premiers sont $\leq T$. Ainsi, lorsque $\Re(s) > 1$, on a

$$\left|\sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} - \prod_{p\leq T} \left(1 - \frac{\chi(p)}{p^s}\right)^{-1}\right| = \left|\sum_{n\notin\mathcal{N}(T)} \frac{\chi(n)}{n^s}\right| \leq \sum_{n>T} \left|\frac{\chi(n)}{n^s}\right| \leq \sum_{n>T} \frac{1}{n^\sigma}.$$

La dernière somme est la queue d'une série réelle convergente (quand $\sigma := \Re(s) > 1$) et tend donc vers zéro quand T tend vers l'infini, ce qui prouve à la fois la convergence du produit et la formule d'Euler généralisée. □

Corollaire. *Pour $\Re(s) > 1$, on a $L(\chi, s) \neq 0$.*

Preuve. C'est clair puisque le produit d'Euler est convergent et que $1 - \chi(p)p^{-s} \neq 0$. □

Corollaire. *On a également les formules*

$$\log L(\chi, s) = \sum_p \sum_{m\geq 1} \frac{\chi(p)^m}{m} p^{-ms} \qquad \text{et} \qquad -\frac{L'(\chi, s)}{L(\chi, s)} = \sum_p \sum_{m\geq 1} \chi(p)^m \frac{\log p}{p^{ms}} = \sum_p \chi(n) \frac{\Lambda(n)}{n^s}.$$

Preuve. La preuve est semblable à celle donnée pour la fonction $\zeta(s)$. □

Intermède (II). Choisissons $K = \mathbf{Q}(\sqrt{d})$ avec d sans facteurs carrés. La loi de décomposition des premiers de $\mathbf{Z}$ dans $\mathcal{O}_K$ permet d'écrire la contribution en p à la fonction de Dedekind $\zeta_K(s)$

$$\begin{cases} (1 - p^{-s})^{-2} = (1 - p^{-s})^{-1}(1 - \left(\frac{d}{p}\right) p^{-s})^{-1} & p \text{ est décomposé dans } K \\ (1 - p^{-2s})^{-1} = (1 - p^{-s})^{-1}(1 - \left(\frac{d}{p}\right) p^{-s})^{-1} & p \text{ est inerte dans } K \\ (1 - p^{-s})^{-1} = (1 - p^{-s})^{-1}(1 - \left(\frac{d}{p}\right) p^{-s})^{-1} & p \text{ est ramifié dans } K \end{cases}$$

(pour p impair avec un énoncé similaire pour $p = 2$). On en tire donc que

$$\zeta_K(s) = \zeta(s)L(\chi_d, s)$$

avec χ le caractère défini par $\chi_d(p) = \left(\frac{d}{p}\right)$ pour p impair et $\chi_d(2) = 1$ (resp. $\chi_d(2) = -1$, $\chi_d(2) = 0$) si $d \equiv 1 \bmod 8$ (resp. $d \equiv 5 \bmod 8$, $d \equiv 2$ ou $3 \bmod 4$). On pourra montrer en exercice que, si $D := |d|$ pour $d \equiv 1 \bmod 4$ (resp. $D := 4|d|$ pour $d \equiv 2$ ou $3 \bmod 4$) χ_d est un caractère modulo D. On va montrer si-dessous que $L(\chi_d, 1) \neq 0$ et donc la fonction $\zeta_K(s)$ possède, tout comme $\zeta(s)$ un pôle d'ordre 1 en $s = 1$ et le résidu vaut $L(\chi_d, 1)$. Un des plus jolis résultats de théorie analytique, la *formule des classes* s'exprime ainsi dans ce cas :

$$\text{Rés}(\zeta_K, 1) = \begin{cases} \frac{2\pi h_K}{w\sqrt{D}} & \text{si } K \text{ imaginaire} \\ \frac{2h_K \log \epsilon}{\sqrt{D}} & \text{si } K \text{ réel} \end{cases}$$

où h_K est le nombre de classes, w le nombre de racines de l'unité (égal à 2 si $d < -4$ et resp. 4 et 6 pour $d = -1$ et $d = -3$) et ϵ désigne l'unité fondamentale > 1 (générateur de $\mathcal{O}_K$ modulo ± 1). Cette formule, jointe au calcul explicite de $L(\chi, 1)$ est très utile pour étudier notamment h_K.

Pour la démonstration du théorème de la progession arithmétique, on a besoin du résultat clef suivant :

Théorème. *Soit χ un caractère de Dirichlet différent du caractère unité, alors :*

$$L(\chi, 1) \neq 0.$$

Remarque. Si on savait déjà que le produit d'Euler convergeait, on pourrait conclure immédiatement, puisque $1 - \chi(p)p^{-1} \neq 0$.

Avant de démontrer le théorème, voyons comment en déduire le théorème de Dirichlet. On écrit, pour $\Re(s) > 1$ la formule

$$\sum_{p \equiv a \bmod m} p^{-s} = \frac{1}{\phi(m)} \sum_{p \nmid m} p^{-s} + \frac{1}{\phi(m)} \sum_{\chi \neq 1} \overline{\chi}(a) \left(\sum_p \chi(p) p^{-s} \right).$$

La formule d'Euler généralisée nous permet d'écrire

$$\sum_p \chi(p) p^{-s} = \log L(\chi, s) + \text{fonction holomorphe pour } \Re(s) > 1/2$$

Par conséquent, pour χ différent du caractère unité, au voisinage de $s = 1$, sachant que $L(\chi, 1) \neq 0$, on en déduit $\sum_p \chi(p)p^{-s} = O(1)$. Par contre $\sum_p p^{-s} = -\log(s-1) + O(1)$ donc on conclut que

$$\sum_{p \equiv a \bmod m} p^{-s} = -\frac{\log(s-1)}{\phi(m)} + O(1)$$

ce qui démontre bien le théorème de la progression arithmétique. □

Remarque. Si $\mathcal{Q}$ désigne un sous-ensemble de l'ensemble $\mathcal{P}$ des nombres premiers, on peut définir plusieurs notions de densité. La preuve précédente suggère d'introduire la notion de *densité analytique* :

$$d_{\text{an}}(\mathcal{Q}) := \lim_{s \to 1} \frac{\sum_{p \in \mathcal{Q}} p^{-s}}{\sum_{p \in \mathcal{P}} p^{-s}}.$$

Nous venons donc de démontrer que la densité analytique des premiers congrus à a modulo m est $1/\phi(m)$. On peut aussi définir la densité "naturelle" comme

$$d(\mathcal{Q}) := \lim_{x \to \infty} \frac{\text{card}\{p \in \mathcal{Q} \mid p \leq x\}}{\text{card}\{p \in \mathcal{P} \mid p \leq x\}}.$$

On peut montrer, mais nous ne le ferons pas, que la densité naturelle des premiers congrus à a modulo m est $1/\phi(m)$.

Pour prouver que $L(\chi, 1) \neq 0$, nous utiliserons le lemme suivant concernant les séries de Dirichlet à coefficients réels positifs.

Lemme. *Soient $a_n \geq 0$ réels, supposons que la série $F(s) = \sum_{n=1}^{\infty} a_n n^{-s}$ converge pour $\Re(s) > \sigma_0$ et que la fonction se prolonge analytiquement au voisinage de σ_0, alors l'abscisse de convergence de la série définissant $F(s)$ est strictement inférieure à σ_0.*

Preuve. Choisissons $r > 0$ et $\sigma < \sigma_0$ et $\sigma_1 > \sigma_0$ de sorte que $\sigma \in D(\sigma_1, r)$ avec ce disque contenu dans le domaine d'holomorphie de $F(s)$. Le point σ_1 est dans le demi-plan de convergence, donc

$$F^{(k)}(\sigma_1) = \sum_{n=1}^{\infty} a_n (-\log n)^k n^{-\sigma_1}.$$

En écrivant le développement en série entière de F dans le disque $D(\sigma_1, r)$ au point σ on obtient :

$$\begin{aligned}
F(\sigma) &= \sum_{k=0}^{\infty} \frac{F^{(k)}(\sigma_1)}{k!} (\sigma - \sigma_1)^k \\
&= \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{n=1}^{\infty} a_n (-\log n)^k n^{-\sigma_1} (\sigma - \sigma_1)^k \\
&= \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{n=1}^{\infty} a_n (\log n)^k n^{-\sigma_1} (\sigma_1 - \sigma)^k \\
&= \sum_{n=1}^{\infty} a_n n^{-\sigma_1} \sum_{k=0}^{\infty} \frac{1}{k!} (\log n)^k (\sigma_1 - \sigma)^k \\
&= \sum_{n=1}^{\infty} a_n n^{-\sigma_1} \exp\left(\log n(\sigma_1 - \sigma)\right) \\
&= \sum_{n=1}^{\infty} a_n n^{-\sigma_1} n^{\sigma_1 - \sigma} \\
&= \sum_{n=1}^{\infty} a_n n^{-\sigma}
\end{aligned}$$

où l'interversion des sommations est justifiée par la positivité des termes. Ceci montre que la série converge en σ. □

Lemme. *Soit $\hat{G}$ l'ensemble des caractères modulo m, soit p premier ne divisant pas m et soit f_p l'ordre de $p \bmod m$ et $g_p := \phi(m)/f_p$, on a l'identité :*

$$\prod_{\chi \in \hat{G}} (1 - \chi(p)T) = \left(1 - T^{f_p}\right)^{g_p}.$$

Preuve. Cela découle du lemme général sur les valeurs en un point des caractères. □

Corollaire. *La fonction $F(s) := \prod_{\chi \in \hat{G}} L(\chi, s)$ est une série de Dirichlet à coefficients positifs sur le demi-plan $\Re(s) > 1$ et possède un pôle simple en $s = 1$.*

Preuve. Pour la première affirmation, on calcule

$$\prod_{\chi} L(\chi, s) = \prod_{p \nmid m} \prod_{\chi} \left(1 - \frac{\chi(p)}{p^s}\right) = \prod_{p \nmid m} \left(1 - \frac{1}{p^{s f_p}}\right)^{-g_p} = \prod_{p \nmid m} \left(\sum_{r=0}^{\infty} p^{-r f_p s}\right)^{g_p}$$

qui est clairement une série de Dirichlet à coefficients positifs. Par ailleurs pour $\sigma \in \mathbf{R}$, en remarquant que $g_p \geq 1$ et $f_p \leq \phi(m)$, on a

$$\prod_{\chi} L(\chi, \sigma) = \prod_{p \not| m} \left(\sum_{r=0}^{\infty} p^{-r f_p \sigma} \right)^{g_p} \geq \prod_{p \not| m} \left(1 + p^{-\sigma \phi(m)} \right).$$

Ainsi la série et le produit divergent pour $\sigma = 1/\phi(m)$.

Par ailleurs la fonction $L(\chi_0, s)$, tout comme $\zeta(s)$ est méromorphe sur $\Re(s) > 0$ avec un unique pôle simple en $s = 1$; les autres $L(\chi, s)$ sont holomorphes sur $\Re(s) > 0$, donc le produit de ces fonctions est méromorphe sur $\Re(s) > 0$ avec un pôle simple en $s = 1$ si $\prod_{\chi \neq \chi_0} L(\chi, 1) \neq 1$ et aucun pôle si l'un des $L(\chi, 1)$ est nul. Montrons que ce dernier cas ne peut se produire. En effet si la fonction produit est holomorphe jusqu'à $\Re(s) > 0$ alors le lemme sur les séries de Dirichlet à coefficients positifs entraînerait que l'abscisse de convergence serait ≤ 0, ce qui serait contradictoire. □

On déduit de l'argument précédent que $L(\chi, 1)$ est non nul pour chaque χ différent du caractère unité et on termine ainsi la preuve du théorème de la progression arithmétique.

Remarque. On peut montrer que (aux facteurs correspondant aux premiers p divisant m près) le produit $\prod_{\chi} L(\chi, \sigma)$ est égal à la fonction zêta de Dedekind du corps $\mathbf{Q}(\exp(2\pi i/m))$, ce qui explique de manière plus conceptuelle pourquoi les coefficients sont positifs.

E. Le théorème des nombres premiers.

Nous allons démontrer le théorème des nombres premiers sous la forme :

Théorème. *L'intégrale $\int_1^{\infty} (\theta(t) - t) t^{-2} dt$ est convergente.*

Montrons que la convergence de l'intégrale implique $\theta(x) \sim x$ et donc $\pi(x) \sim x/\log x$. Supposons en effet que $\limsup \theta(x) x^{-1} > 1$, alors il existe $\epsilon > 0$ et x_n tendant vers l'infini tels que $\theta(x_n) x_n^{-1} \geq 1 + \epsilon$. Pour $t \in [x_n, (1 + \epsilon/2) x_n]$ on a donc $(\theta(t) - t) t^{-2} \geq (\theta(x_n) - (1 + \epsilon/2) x_n)\, x_n^2 \geq \epsilon/2 x_n$ et par conséquent

$$\int_{x_n}^{(1+\epsilon/2)x_n} (\theta(t) - t) t^{-2} dt \geq \frac{\epsilon^2}{4},$$

ce qui contredit la convergence de l'intégrale. On conclut que $\limsup \theta(x) x^{-1} \leq 1$. Un argument symétrique montre que $\liminf \theta(x) x^{-1} \geq 1$ et donc $\lim \theta(x) x^{-1} = 1$.

Pour démontrer le théorème nous allons utiliser le résultat suivant de variables complexes (dû à Newman).

Théorème. ("théorème analytique") *Soit $h(t)$ une fonction bornée et continue par morceaux, alors l'intégrale*

$$F(s) = \int_0^{+\infty} h(u) e^{-su} du$$

est convergente et définit une fonction homorphe sur le demi-plan $\Re(s) > 0$. Supposons que cette fonction se prolonge en une fonction holomorphe sur le demi-plan fermé *$\Re(s) \geq 0$, alors l'intégrale pour $s = 0$ converge et*

$$F(0) = \int_0^{+\infty} h(u) du.$$

Admettons provisoirement ce résultat, et voyons comment l'appliquer à la fonction

$$F(s) = \int_1^{+\infty} \frac{\theta(t) - t}{t^{s+2}} dt = \int_0^{+\infty} \frac{\theta(e^u) - e^u}{e^{u(s+2)}} e^u du = \int_0^{+\infty} (\theta(e^u) e^{-u} - 1) e^{-us} du$$

La fonction $h(u) := \theta(e^u)e^{-u} - 1$ est bien bornée et continue par morceaux donc, si l'on vérifie l'hypothèse de prolongement holomorphe, on pourra conclure que $F(0) = \int_0^{+\infty}(\theta(e^u)e^{-u} - 1)du = \int_1^{+\infty} \frac{\theta(t)-t}{t^2}dt$ est bien convergente.

On peut transformer l'intégrale définissant $F(s)$ (pour $\Re(s) > 0$) ainsi :

$$\begin{aligned} F(s) &= \int_1^{+\infty} \frac{\theta(t) - t}{t^{s+2}}dt \\ &= \sum_{n=1}^{\infty} \int_n^{n+1} \theta(t)t^{-s-2}dt - \int_1^{+\infty} t^{-s-1}dt \\ &= \sum_{n=1}^{\infty} \theta(n)\frac{n^{-s-1} - (n+1)^{-s-1}}{s+1} - \frac{1}{s} \\ &= \frac{1}{s+1}\sum_{n=1}^{\infty} n^{-s-1}\left(\theta(n) - \theta(n-1)\right) - \frac{1}{s} \\ &= \frac{1}{s+1}\sum_{p} p^{-s-1}\log(p) - \frac{1}{s}. \end{aligned}$$

Or on a vu que

$$-\frac{\zeta'(s)}{\zeta(s)} = \sum_{p,m\geq 1} \log(p)p^{-ms} = \sum_p \log(p)p^{-s} + \sum_{p,m\geq 2} \log(p)p^{-ms}$$

et la deuxième somme dans le dernier terme est convergente et donc holomorphe pour $\Re(s) > 1/2$. On en tire que $\sum_p \log(p)p^{-s} = -\frac{\zeta'(s)}{\zeta(s)} +$ fonction holomorphe sur $\Re(s) > 1/2$ et enfin que

$$F(s) = -\frac{\zeta'(s+1)}{\zeta(s+1)} - \frac{1}{s} + \text{ fonction holomorphe sur } \Re(s) > -1/2$$

Le point clef de la démonstration est alors le résultat suivant :

Théorème. (Hadamard, De la Vallée-Poussin) *La fonction $\zeta(s)$ ne s'annule pas sur la droite $\Re(s) = 1$.*

Preuve. On part de la formule

$$4\cos(x) + \cos(2x) + 3 = 2(1 + \cos(x))^2 \geq 0.$$

On rappelle que

$$\log\zeta(\sigma + it) = \sum_{p,m} \frac{p^{-m\sigma - mit}}{m}, \qquad \text{et} \qquad \log|\zeta(\sigma + it)| = \sum_{p,m} \frac{p^{-m\sigma}}{m}\cos(-mt\log p).$$

On obtient ainsi

$$\log\left(|\zeta(\sigma + it)|^4|\zeta(\sigma + 2it)|\zeta(\sigma)^3\right) = \sum_{p,m\geq 1} \frac{p^{-m\sigma}}{m}\left(4\cos(-mt\log p) + \cos(-2mt\log p) + 3\right) \geq 0$$

Ainsi on peut conclure que, toujours pour $\sigma > 1$, on a

$$|\zeta(\sigma + it)|^4|\zeta(\sigma + 2it)|\zeta(\sigma)^3 \geq 1. \qquad (*)$$

Si maintenant $\zeta(s)$ a un zéro d'ordre k en $1 + it$, d'ordre ℓ en $1 + 2it$, on a $|\zeta(\sigma + it)| \sim a(\sigma - 1)^k$, $|\zeta(\sigma + 2it)| \sim b(\sigma - 1)^\ell$ et $\zeta(\sigma) \sim (\sigma - 1)^{-1}$ (quand σ tend vers 1 par valeurs supérieures) donc le membre de gauche de l'inégalité $(*)$ est équivalent à $c(\sigma - 1)^{4k+\ell-3}$, ce qui entraîne $4k + \ell - 3 \leq 0$ et donc $k = 0$. □

Corollaire. *La fonction définie pour* $\Re(s) > 1$ *par*

$$G(s) := -\frac{\zeta'(s)}{\zeta(s)} - \frac{1}{s-1}$$

s'étend en une fonction holomorphe sur $\Re(s) \geq 1$.

Preuve. Le théorème précédent montre que la fonction $\zeta'(s)/\zeta(s)$ est holomorphe sur la droite $\Re(s) = 1$ sauf en $s = 1$, par conséquent la fonction $G(s)$ l'est également. Pour étudier $G(s)$ au voisinage de $s = 1$ on utilise que $\zeta(s)$ a un pôle simple en $s = 1$ et par conséquent $\zeta'(s)/\zeta(s) = -1/(s-1) + g(s)$ avec $g(s)$ holomorphe au voisinage de 1. Ainsi $G(s)$ est bien holomorphe au voisinage de 1 et donc sur la droite $\Re(s) = 1$. □

Ce dernier résultat achève la preuve du théorème des nombres premiers, au résultat analytique près qui est démontré en appendice.

Appendice : Preuve du "théorème analytique"

Démontrons maintenant le résultat analytique utilisé dans la preuve du théorème des nombres premiers dont nous rappelons l'énoncé :

Théorème. *Soit* $h(t)$ *une fonction bornée et continue par morceaux, alors l'intégrale*

$$F(s) = \int_0^{+\infty} h(u)e^{-su}du$$

est convergente et définit une fonction homorphe sur le demi-plan $\Re(s) > 0$. *Supposons que cette fonction se prolonge en une fonction holomorphe sur le demi-plan* fermé $\Re(s) \geq 0$, *alors l'intégrale pour* $s = 0$ *converge et*

$$F(0) = \int_0^{+\infty} h(u)du.$$

Preuve. La première partie est immédiate, démontrons donc la deuxième affirmation. Pour T réel (grand), définissons $F_T(s) := \int_0^T h(t)e^{-st}dt$; ce sont des fonctions holomorphes pour tout $s \in \mathbf{C}$. On veut montrer que $\lim_{T\to\infty} F_T(0)$ existe et vaut $F(0)$. Pour cela considérons R grand et le contour $\gamma = \gamma(R, \delta)$ délimitant la région $S := \{s \in \mathbf{C} \mid \Re(z) \geq -\delta \text{ et } |s| \leq R\}$. Une fois fixé R, on peut choisir $\delta > 0$ suffisamment petit pour que $F(s)$ soit analytique sur cette région. L'astuce réside dans le choix de la fonction suivante, on pose

$$G_T(s) := (F(s) - F_T(s))\, e^{sT} \left(1 + \frac{s^2}{R^2}\right)$$

de sorte que $G_T(0) = F(0) - F_T(0)$ et on veut donc montrer que $\lim_{T\to\infty} G_T(0) = 0$. Pour cela utilisons le théorème des résidus une première fois en notant que

$$G_T(0) = F(0) - F_T(0) = \frac{1}{2\pi i} \int_\gamma (F(s) - F_T(s))\, e^{sT} \left(1 + \frac{s^2}{R^2}\right) \frac{ds}{s}.$$

Pour majorer cette intégrale on la découpe en deux morceaux : γ_1 qui est la partie de γ située dans le demi-plan $\Re(s) > 0$ et γ_2 celle située dans le demi-plan $\Re(s) < 0$. On utilise le calcul suivant :

Soit s avec $|s| = R$ ou encore $s = Re^{i\theta}$, alors on a :

$$\left| e^{sT} \left(1 + \frac{s^2}{R^2}\right) \frac{1}{s} \right| = e^{\Re(s)T} \left| e^{-i\theta} + e^{i\theta} \right| \frac{1}{R} = e^{\Re(s)T} \frac{2\Re(s)}{R^2}.$$

Ensuite on a la majoration

$$|F(s) - F_T(s)| = \int_T^\infty h(t)e^{-st}dt \leq M \int_T^\infty \left|e^{-st}\right| dt = \frac{Me^{-\Re(s)T}}{\Re(s)}.$$

On obtient maintenant

$$\left|\frac{1}{2\pi i}\int_{\gamma_1}(F(s)-F_T(s))\,e^{sT}\left(1+\frac{s^2}{R^2}\right)\frac{ds}{s}\right|\leq\frac{M}{R}.$$

Ainsi, à condition d'avoir choisi R très grand, ce morceau de l'intégrale sera arbitrairement petit. Découpons maintenant l'intégrale sur γ_2 en deux morceaux $I_1 - I_2$ avec :

$$I_1 := \frac{1}{2\pi i}\int_{\gamma_2} F(s)e^{sT}\left(1+\frac{s^2}{R^2}\right)\frac{ds}{s} \qquad \text{et} \qquad I_2 := \frac{1}{2\pi i}\int_{\gamma_2} F_T(s)e^{sT}\left(1+\frac{s^2}{R^2}\right)\frac{ds}{s}$$

Pour majorer I_2 on observe que, par le théorème des résidus (ou ici la formule de Cauchy), on peut remplacer le contour γ_2 par un arc de cercle de rayon R et en utilisant les mêmes majorations conclure que $|I_2| \leq M/R$. Pour majorer I_1 on observe simplement que la fonction $F(s)e^{sT}\left(1+\frac{s^2}{R^2}\right)\frac{ds}{s}$ converge vers 0 quand T tend vers $+\infty$, et ceci uniformément sur tout compact contenu dans $\Re(s) < 0$. Par conséquent on a :

$$\lim_{T\to\infty}\frac{1}{2\pi i}\int_{\gamma_2} F(s)e^{sT}\left(1+\frac{s^2}{R^2}\right)\frac{ds}{s} = 0$$

En mettant ensemble les trois majorations obtenues, on voit que

$$|F_T(0) - F(0)| \leq \frac{2M}{R} + \epsilon(T)$$

avec $\epsilon(T)$ tendant vers zéro (de façon dépendant de R), ceci suffit à prouver que $\lim F_T(0) = F(0)$, ce qu'il fallait démontrer. □

Compléments sans preuves.

1) En lieu et place du résultat analytique, on peut aussi utiliser le théorème de Ikehara, plus puissant mais aussi plus délicat à démontrer, que nous nous contenterons d'énoncer :

Théorème. (Ikehara) *Soit $A(t)$ une fonction croissante telle que l'intégrale $F(s) = \int_1^{+\infty} A(t)t^{-s-1}dt$ soit convergente pour $\Re(s) > 1$ et se prolonge continûment à la droite $\Re(s) = 1$ sauf un pôle simple en $s = 1$ de résidu λ (c'est-à-dire que la fonction $F(s) - \lambda/(s-1)$ se prolonge continûment à $\Re(s) \geq 1$) alors*

$$A(x) \sim \lambda x.$$

Si $\lambda = 0$, c'est-à-dire s'il n'y a pas de pôle en $s = 1$, on obtient $A(x) = o(x)$.

Le théorème des nombres premiers s'en déduit en remarquant que les hypothèses s'appliquent avec $A(x) = \psi(x)$ puisque

$$-\frac{\zeta'(s)}{\zeta(s)} = s\int_1^{+\infty}\psi(t)t^{-s-1}dt.$$

2) Le résultat de non annulation de la fonction ζ sur la droite $\Re(s) = 1$ peut être considérablement renforcé, au moins conjecturalement :

Conjecture. ("Hypothèse de Riemann"[(*)]) *Soit $s \in \mathbf{C}$ avec $\Re(s) > 1/2$, alors $\zeta(s) \neq 0$.*

Si on savait démontrer cette hypothèse, on pourrait en déduire que $\psi(x) = x + O(x^\alpha)$ pour tout $\alpha > 1/2$ et de même $\theta(x) = x + O(x^\alpha)$. On en tirerait alors un équivalent beaucoup plus précis pour $\pi(x)$ en utilisant la formule (déduite en appliquant la formule d'Abel) :

$$\pi(x) = \frac{\theta(x)}{\log(x)} + \int_2^x \frac{\theta(t)dt}{t(logt)^2}.$$

[(*)] L'hypothèse de Riemann est un des problèmes ouverts majeurs parmi les mathématiques ; la fondation Clay offre également un million de dollars pour sa solution.

On peut transformer cette formule en

$$\pi(x) = \int_2^x \frac{dt}{\log t} + 2\log 2 + \frac{\theta(x) - x}{\log(x)} + \int_2^x \frac{\theta(t) - t}{t(logt)^2} dt.$$

Si l'on introduit la fonction "*logarithme intégral*"

$$Li(x) := \int_2^x \frac{dt}{\log t}$$

on voit que l'hypothèse de Riemann entraîne que $\pi(x) = Li(x) + O(x^\alpha)$ pour tout $\alpha > 1/2$. En observant que

$$Li(x) = \frac{x}{\log x} + \frac{1}{2}\frac{x}{(\log x)^2} + O\left(\frac{x}{(\log x)^3}\right)$$

on voit que $Li(x)$ constitue une approximation bien plus précise que $x/\log(x)$. Hélas le meilleur résultat démontré est loin de ces espoirs, on sait néanmoins prouver des énoncés comme

$$\pi(x) = Li(x) + O\left(x \exp\left(-c\sqrt{\log x}\right)\right).$$