Table des matières

0 Introduction

Un étudiant en mathématiques est souvent, d'abord et avant tout, quelqu'un qui se passionne pour les mystères entourant les nombres. La science la plus pure qui englobe ce sujet se nomme la théorie des nombres et elle est vieille de plus de 4500 ans. Cette science fait elle-même partie de ce que nous appelons les mathématiques et la frontière entre les deux est pratiquement inexistante. Personnellement, je ne me souviens même pas depuis combien de temps on m'a dit que chacun des entiers se factorise de façon unique en produit de nombres premiers. Cette assertion qui semble très évidente nécessite tout de même une preuve rigoureuse. Lorsque nous amorçons une étude plus appronfondie de la théorie des nombres, nous sommes très vite confrontés à comprendre différentes structures qui, en quelque sorte, sont comparables à la suite des entiers. Par exemple, très tôt nous voyons la définition de groupe, qui à première vue semble être trop abstraite pour être intéressante. Ensuite, nous rencontrons les anneaux $\mathbb{Z}_n$ et finalement les corps. Chacune de ces structures hérite naturellement de la commutativité de $\mathbb{Z}$. On rencontre aussi plusieurs autres sortes de structures en allégeant, renforçant ou modifiant les conditions et elles ont leur importance, mais il n'en sera pas question dans ce travail. Donc, après avoir défini ce que sont un idéal et une classe d'équivalence, nous voyons $\mathbb{Z}$ comme une immense source d'anneaux quotients. Les anneaux quotients de $\mathbb{Z}$ les plus simples, ceux qui apparaissent en quotientant par des idéaux qui sont maximaux, jouissent même d'une structure de corps. Les autres, la majorité, ne sont pas des domaines d'intégrité. Cette dernière assertion se quantifie à l'aide du théorème des nombres premiers.

Dans ce travail, nous allons souvent nous restreindre aux corps finis à p éléments, bien que la majorité des techniques employées seront valides tantôt dans tout corps fini, tantôt dans tout groupe commutatif, les détails techniques et la beauté des résultats favorisant cette restriction. Cependant, il sera toujours clairement indiqué pour quelle sorte d'objets chaque assertion est valide.

Comme le titre l'indique, ce mémoire a pour but de faire ressortir un ensemble de structures, d'idées et de principes qui sont très utiles pour étudier certains problèmes relatifs aux corps finis. Les questions qui y sont traitées font parties de plusieurs domaines en même temps. Par exemple, une technique pour caractériser les suites à distribution uniforme dans un groupe fini est décrite au chapitre 1. Il s'agit du fameux critère de Weyl qui est très similaire au critère pour la distribution uniforme dans $[0,1]$; cependant, la preuve de la version quantitative est plus courte. Dans le but d'appliquer ce principe au plus grand nombre de suites possibles, nous sommes amenés à introduire une technique qui fut inventée par Stepanov vers la fin des années soixante. Cette technique est basée sur l'idée très simple suivante : si un ensemble de M éléments peut se faire annuler au complet avec une mutiplicité d'au moins D par un polynôme P non identiquement nul, alors on peut en déduire que $M < \frac{degP}{D}$. Un des résultats spectaculaires est le fait que $\frac{n^p - n}{p}$ est distribué uniformément modulo p lorsque n parcourt $\mathbb{Z}_p$.

Le chapitre 3 est voué à une preuve de l'hypothèse de Riemann pour les corps finis.

Cette assertion est en fait un résultat sur le nombre de zéros qu'a un polynôme $f(x,y)$ absolument irréductible sur $\mathbb{F}_q[x,y]$. La preuve, bien qu'extrêmement astucieuse, est en fait basée sur la simple idée de Stepanov. Cette dernière est lourdement inspirée de celle de Axel Thue, du début du siècle dernier, de son résultat sur l'approximation diophantienne des nombres algébriques qui fait d'ailleurs l'objet du chapitre 4. Ce dernier chapitre peut être vu comme un appendice qui est là dans le but de mettre en évidence le vrai lien entre ces techniques.

Le chapitre 2 se distingue des autres par son contenu. En effet, c'est le seul qui ne contient aucune construction de polynômes. Il est consacré au complet à la preuve de la conjecture originale de Kurepa qui fut démontrée en 2004 par Daniel Barsky et Bénali Benzaghou dans [1]. Cet énoncé stipule que le nombre

$$!n := \sum_{i=0}^{n-1} i!$$

n'est pas divisible par n lorsque $n > 2$. En fait, il existe une liste très impressionnante d'assertions qui y sont équivalentes (voir [23]). Premièrement, il est clair qu'il est suffisant de se restreindre au cas où n est un nombre premier p. Ensuite, il n'est pas trop difficile de voir que le même énoncé est équivalent au fait que $(n!, !n) = 2$ pour $n \geq 2$. La preuve de ce résultat utilise plusieurs idées provenant de l'algèbre linéaire et une foule de calculs astucieux dans une extension de type Artin-Schreier. Le fil conducteur qui me laisse introduire ce chapitre est un résultat, qui à ma connaissance, semble nouveau même si les idées pour le produire ont au moins 10 ans. En effet, il y a beaucoup de travaux qui se font sur le genre de question suivante : Étant donnée une suite S de M termes dans $\mathbb{F}_p$, combien d'éléments de $\mathbb{F}_p$ sont effectivement représentés par cette suite ? Les exemples qui me viennent à l'esprit sont les récents articles de Moubariz Z. Garaev, Florian Luca et Igor E. Shparlinski sur les factoriels (voir [15] et [16]).

En fait, dès que la suite en question n'a aucune raison d'avoir de structure, un argument probabiliste très simple nous laisse penser qu'il devrait y avoir

$$\left(1 - \left(1 - \frac{1}{p}\right)^M + o_S(1)\right) p$$

termes différents de $\mathbb{F}_p$ dans la suite en question, et c'est ce qu'on observe en pratique. Il est à noter que pour un polynôme fixé $f(x) \in \mathbb{F}_p[x]$, on peut calculer l'exacte proportion des éléments de $\mathbb{F}_p$ que prend $f(n)$, lorsque n parcourt $\mathbb{F}_p$, avec p qui tend vers l'infini ; voir par exemple [2]. Pour le cas qui nous concerne, si on définit le polynôme

$$\kappa_p(x) := \sum_{i=0}^{p-1} i! x^i,$$

alors en utilisant les techniques de Stepanov [32], Heath-Brown et Konyagin [22], nous montrons au chapitre 1 que le nombre de valeurs prises par $\kappa_p(n)$, lorsque n parcourt $\mathbb{F}_p$, est au moins $(1/14)p^{1/2}$ et qu'en particulier ce polynôme a au plus $4p^{2/3}$ zéros dans $\mathbb{F}_p$. Ces résultats sont loin du fait presque certain que ce polynôme doit prendre $(1-1/e+o(1))p$ valeurs de $\mathbb{F}_p$. Le chapitre 2 est l'étude de $\kappa_p(x)$ spécialisé en $x = 1$.

0.1 Remarques préliminaires.

Comme il est coutume, p représente toujours un nombre premier et q, une puissance d'un nombre premier. Dans un premier temps, il n'est pas clair qu'un corps de q éléments existe. Pour le voir, il suffit de considérer les racines du polynôme $x^q - x \in \mathbb{Z}_p[x]$, où $q = p^n$ pour un certain $n > 0$. Comme les nombres

$$\binom{q}{j} \quad \text{avec} \quad 1 \leq j \leq q-1$$

sont tous divisibles par p, il s'ensuit que la somme et la différence de deux racines est aussi une racine. Il en est de même pour le produit de deux racines. Clairement ce polynôme est à racines simples, car $\partial(x^q - x) = -1 \neq 0$. On voit aussi que 0 et 1 sont des racines. Si $\mathbb{A}_q$ est l'ensemble des racines distinctes, alors on a montré que $|\mathbb{A}_q| = q$ et qu'il possède une structure d'anneau. Prenons un élément non nul b de $\mathbb{A}_q$ et considérons l'ensemble $b\mathbb{A}_q$. Si deux éléments $b\beta_1$ et $b\beta_2$ avec $\beta_1 \neq \beta_2$ de ce nouvel ensemble sont identiques, alors c'est que $b(\beta_1 - \beta_2) = 0$ et comme $\mathbb{Z}_p$ est un domaine d'intégrité, on en déduit que $\beta_1 = \beta_2$ ce qui est une contradiction. On vient donc de montrer que $b\mathbb{A}_q$ est une permutation de $\mathbb{A}_q$ et en particulier $1 \in b\mathbb{A}_q$ et donc b est inversible et finalement $\mathbb{A}_q =: \mathbb{F}_q$ est un corps.

Une façon de construire explicitement un tel corps est de considérer l'anneau quotient $\mathbb{F}_p[x]/I$ où I est l'idéal engendré par un polynôme $P(x)$ irréductible de degré n. Comme $P(x)$ est irréductible, il s'ensuit que $\mathbb{F}_p[x]/I$ est un corps et il est évident qu'il possède q éléments. Ce qui n'est pas clair a priori, c'est qu'un tel polynôme existe. Soit donc R l'ensemble des polynômes irréductibles unitaires de $\mathbb{F}_p[x]$ et $R(d)$ le nombre de ceux qui sont de degré d, clairement on a $R(d) \leq p^d$ pour $d \geq 0$. Comme $\mathbb{F}_p[x]$ est un domaine à factorisation unique, on peut écrire

$$\begin{aligned}\sum_{\substack{f(x)\in\mathbb{F}_p[x]\\ f(x)\ unitaire}} t^{degf} &= \sum_{d\geq 0}(pt)^d = \frac{1}{1-pt} \\ &= \prod_{P(x)\in R} \frac{1}{1-t^{degP}} = \prod_{d\geq 1}\left(\frac{1}{1-t^d}\right)^{R(d)}.\end{aligned}$$

On a donc

$$\ln(1-pt) = \sum_{d\geq 1} R(d)\ln(1-t^d),$$

d'où on obtient

$$\sum_{m\geq 1}\frac{(pt)^m}{m} = \sum_{k\geq 1}\sum_{d\geq 1} dR(d)\frac{t^{kd}}{kd}.$$

En comparant les coefficients, on obtient la relation

$$p^m = \sum_{d|m} dR(d) \equiv$$

valide pour $m > 0$, qui nous fournit finalement que

$$R(m) = \frac{1}{m} \sum_{d|m} \mu(d) p^{m/d},$$

où μ est la fonction de Moebius. On peut obtenir cette relation un peu plus rapidement en utilisant directement le fait que $x^q - x$ est exactement le produit des polynômes irréductibles de degré d qui divise n (ce qu'on a montré implicitement), et donc l'existence d'au moins un candidat est presqu'une trivialité. Cependant, la méthode exposée ici est plus esthétique et dans le fond elle revient au même : les calculs font le raisonnement pour nous en utilisant seulement le fait que $\mathbb{F}_p[x]$ est à factorisation unique. Néanmoins, nous avons montré que

$$\left| R(n) - \frac{p^n}{n} \right| < \frac{2p^{n/2}}{n}$$

pour $n > 0$. Le même raisonnement peut s'appliquer si le corps de base est $\mathbb{F}_{p^a}$ et qu'on veut y compter les polynômes irréductibles de degré e.

Une propriété des corps finis qui est particulièrement importante est le fait que le groupe multiplicatif $\mathbb{F}_q^*$ est cyclique. Pour le voir on commence par remarquer, par induction, que dans un corps fini les polynômes de degré d ont au plus d racines. Ensuite, comme les éléments de $\mathbb{F}_q^*$ satisfont $x^{q-1} = 1$, il s'ensuit que les ordres des éléments de $\mathbb{F}_q^*$ sont des diviseurs de $q - 1$, i.e. chaque élément de $\mathbb{F}_q^*$ satisfait une équation de la forme $x^d = 1$ avec $d|q-1$. Alors supposons qu'aucun ordre ne vaut exactement $q-1$ et posons $\eta(d)$ pour le nombre d'éléments de $\mathbb{F}_q^*$ qui ont un ordre égal à d. Si $\eta(d) \neq 0$, alors il existe un élément x d'ordre d. On considère alors x^e avec $e = 1, \ldots, d$. Ces valeurs sont toutes des solutions distinctes de l'équation $y^d = 1$. Il n'y en a donc pas d'autres. Soit maintenant z un élément d'ordre d. Un tel élément est donc de la forme x^e et de plus $(d, e) = 1$. On en déduit que $\eta(d) \leq \phi(d)$ et on peut alors écrire

$$q - 1 = \sum_{d|q-1} \eta(d) \leq \sum_{d|q-1} \phi(d) = q - 1,$$

ce qui nous permet de constater que $\eta(d) = \phi(d)$ pour tout d qui divise $q - 1$ et donc que le nombre de racines primitives est exactement $\phi(q-1)$ qui est non nul.

Une autre propriété fondamentale est la connaissance du groupe de Galois de $\mathbb{F}_q/\mathbb{F}_p$. Comme $\sigma(y) := y^p$ laisse fixe $\mathbb{F}_p$, il s'ensuit que σ génère un sous-groupe de $Gal(\mathbb{F}_q/\mathbb{F}_p)$. Comme $y^q = y$, il est naturel de considérer σ^i avec $0 \leq i \leq n-1$. Premièrement, on remarque que ces éléments sont distincts car si $y^{p^i} = y^{p^j}$ avec $0 \leq i < j \leq n-1$, alors $y^{p^j - p^i} = 1$ ce qui est une contradiction car $p^j - p^i \not\equiv 0$ modulo $q - 1$. On a donc montré que $|Gal(\mathbb{F}_q/\mathbb{F}_p)| \geq n$. La démonstration de l'inégalité inverse se fait en deux étapes. Commençons par montrer que l'existence d'un autre élément $\tau \neq \sigma^i$ pour un $0 \leq i \leq n-1$ entraîne une dépendance linéaire. Pour ce faire, on fixe un polynôme irréductible $f(x)$ sur $\mathbb{F}_p$ de degré n avec $f(\beta) = 0$ et on considère le système d'équations

$$0 = \sum_{i=0}^{n-1} x_i \sigma^i(\beta^j) + x_n \tau(\beta^j) \quad \text{avec} \quad 0 \leq j \leq n-1$$

en les variables x_i avec $0 \leq i \leq n$. Comme on a n équations linéaires et $n+1$ inconnues, on déduit qu'il existe une solution non triviale en nombres α_i avec $0 \leq i \leq n$ dans $\mathbb{F}_p[\beta]$ qui est isomorphe à $\mathbb{F}_q$. On peut alors choisir des éléments a_j de $\mathbb{F}_p$ avec $0 \leq j \leq n-1$ et obtenir la solution générale

$$0 = \sum_{i=0}^{n-1} \alpha_i \sigma^i(a_j \beta^j) + \alpha_n \tau(a_j \beta^j) \quad \text{avec} \quad 0 \leq j \leq n-1.$$

En additionnant les n équations, on obtient

$$0 = \sum_{i=0}^{n-1} \alpha_i \sigma^i \left(\sum_{j=0}^{n-1} a_j \beta^j \right) + \alpha_n \tau \left(\sum_{j=0}^{n-1} a_j \beta^j \right)$$

et comme les β^j avec $0 \leq j \leq n-1$ forment une base de $\mathbb{F}_q$ sur $\mathbb{F}_p$ et que les a_j avec $0 \leq j \leq n-1$ peuvent être choisis de façon arbitraire dans $\mathbb{F}_p$, il s'ensuit qu'on a trouvé la dépendance linéaire

$$0 = \sum_{i=0}^{n-1} \alpha_i \sigma^i + \alpha_n \tau.$$

On remarque qu'on a utilisé toutes les propriétés des automorphismes, sauf la structure de groupe. Pour la deuxième partie de la preuve de l'inégalité $|Gal(\mathbb{F}_q/\mathbb{F}_p)| \leq n$ on utilise seulement la structure multiplicative, donc le fait qu'un automorphisme est un caractère multiplicatif. On suppose ainsi qu'on a une combinaison linéaire avec un nombre minimal de coefficients non nuls. Comme

$$\sum_{i=0}^{n-1} \gamma_i \sigma^i(w) + \gamma_n \tau(w) = 0$$

pour tout $w \in \mathbb{F}_q$ et qu'on peut choisir un élément $z \in \mathbb{F}_q^*$ pour lequel $\sigma^0(z) \neq \tau(z)$, ceci étant dû au fait que les automorphismes sont distincts, on déduit que les équations

$$\sum_{i=0}^{n-1} \gamma_i \sigma^i(zw) + \gamma_n \tau(zw) = \sum_{i=0}^{n-1} \gamma_i \sigma^i(z)\sigma^i(w) + \gamma_n \tau(z)\tau(w) = 0$$

et

$$\sum_{i=0}^{n-1} \gamma_i \tau(z)\sigma^i(w) + \gamma_n \tau(z)\tau(w) = 0$$

sont satisfaites. En les soustrayant, on trouve

$$\sum_{i=0}^{n-1} \gamma_i [\sigma^i(z) - \tau(z)]\sigma^i(w) = 0,$$

ce qui nous fournit une solution non triviale qui contredit la minimalité de notre choix. En somme, on a montré que $Gal(\mathbb{F}_q/\mathbb{F}_p)$ est isomorphe au groupe cyclique $(\mathbb{Z}_n, +)$ et est engendré par le Frobenius σ.

Lorsqu'on s'intéresse aux calculs explicites dans $\mathbb{F}_q$, il est utile de savoir comment inverser rapidement un élément m quelconque de $\mathbb{F}_q^*$, i.e. résoudre l'équation $ml = 1$ en l dans $\mathbb{F}_q^*$. Plus généralement, on doit résoudre efficacement l'équation $ml = b$ avec $l \in \mathbb{F}_q$. Pour m et b fixés, on note la solution par m_b. Il est très facile de voir que $m_b = bm^{q-2}$ et que la complexité de calcul est de $O(\ln(q))$ étapes. Dans le cas où on ne s'intéresse qu'à $\mathbb{F}_p$ avec un p premier fixé, pour des entiers m compris entre 1 et $p-1$, on peut définir la fonction arithmétique $t_b(m)$ par

$$t_b(m) := \frac{mm_b - b}{p},$$

où m_b est le représentant modulo p choisi minimal et positif pour tout $1 \leq m \leq p-1$. Dans un premier temps, il est facile de voir qu'on a toujours $0 \leq t_b(m) \leq m-1$, $t_b(m) = t_b(m_b)$ et donc $0 \leq t_b(m) \leq \min\{m-1, m_b-1\}$. Ensuite, il est naturel de définir $N_{b,p}$ comme étant le nombre de valeurs distinctes prises par la fonction $t_b(m)$ lorsque m parcourt $\mathbb{F}_p^*$. Évidemment, on a que $N_{b,p} \leq (p-1)/2$. En fait ce nombre est exactement égal au nombre d'entiers de la forme $ap+b$ avec $a \in [0, p-2]$ qui ont la propriété de pouvoir s'écrire comme le produit de deux nombres dans $[a+1, p-1]$. Comme $\omega(m)$ et $\omega(ap+b)$, où ω est la fonction qui compte le nombre de facteurs premiers distincts, valent environ $\ln\ln(p)$ pour presque tout $m \in [1, p-1]$ et $a \in [0, p-2]$, il s'ensuit que les nombres de la forme $pt_b(m)+b$ sont de densité limite nulle, i.e. $N_{b,p} = o(p)$. Ceci signifie que ça prend "très peu" d'entiers pour inverser les éléments de $\mathbb{F}_p$ au complet. En fait, les résultats récents de Kevin Ford (voir [10], [11] et [12]) nous laissent supposer qu'on a

$$c\frac{p}{(\ln(p))^{\delta}(\ln\ln(p))^{3/2}} \leq N_{b,p} \leq C\frac{p}{(\ln(p))^{\delta}(\ln\ln(p))^{3/2}}$$

pour certaines constantes c et C avec $c \leq C$, où

$$\delta = 1 - \frac{1+\ln\ln(2)}{\ln(2)} \approx 0.08607133...$$

Ce qui nous apprend en particulier qu'il y a des entiers qui sont pris très souvent par $t_b(m)$.

Cette fonction traduit certaines propriétés arithmétiques intéressantes. En effet, il est facile de voir que

$$\prod_{m=1}^{p-1}(t_b(m)p+b) = (p-1)!^2,$$

ce qui est équivalent à

$$\prod_{m=1}^{p-1}(t_b(m)p+b) - 1 = ((p-1)!-1)((p-1)!+1).$$

On en déduit alors que p est un nombre premier de Wilson, i.e. $(p-1)! \equiv -1$ modulo p^2, si et seulement si

$$p \left| \left(\sum_{m=1}^{p-1} t_b(m)\right) + \frac{b^p - b}{p} \right.$$

pour un certain $b \not\equiv 0$ modulo p, ce qui entraîne la même propriété pour tout $b \not\equiv 0$ modulo p. Seulement trois nombres premiers de Wilson sont connus, il s'agit de 5, 13 et 563. Carl Pomerance, Karl Dilcher et Richard Crandall ont montré dans [3] qu'il n'y en a pas un autre inférieur à 500 millions.

Il est facile de déduire de ce qui précède que la moyenne géométrique vaut

$$\left(\prod_{m=1}^{p-1} t_b(m)\right)^{1/(p-1)} = \frac{p}{e^2} + O\left(\frac{b\ln^2(p)}{p}\right).$$

Il semble difficile de montrer le résultat correspondant pour la moyenne arithmétique. Cependant on obtient l'identité

$$\sum_{b,m=1}^{p-1} t_b(m) = \frac{(p-2)(p-1)^2}{4},$$

qui nous laisse supposer que pour toutes les valeurs de b on a

$$\sum_{m=1}^{p-1} t_b(m) = \frac{p^2}{4}(1+o(1)).$$

Il est possible de calculer les moments de la suite $(t_b(m)/m)_{1\leq m\leq p-1}$. Il suffit de considérer, pour un b fixé avec $1 \leq b \leq p-1$, l'expression

$$\sum_{m=1}^{p-1} \left(\frac{t_b(m)p+b}{m}\right)^k$$

pour montrer que

$$\sum_{m=1}^{p-1} \left(\frac{t_b(m)}{m}\right)^k = \frac{p}{k+1} + O(k\ln(p)),$$

ce qui fournit une valeur asymptotique lorsque $k = o\left(\left(\frac{p}{\ln(p)}\right)^{1/2}\right)$. Le cas $k=1$, à savoir

$$\sum_{m=1}^{p-1} \frac{t_b(m)}{m} = \frac{p-1}{2} - b\frac{H_{p-1}}{p}$$

est particulièrement remarquable, d'autant plus que le nombre harmonique H_{p-1} est divisible par p^2 pour $p \geq 5$. Les moments de cette suite nous permettent de montrer que la suite $(t_b(m)/m)_{1\leq m\leq p-1}$ est uniformément répartie modulo 1. En fait, on a la version explicite du théorème de Weyl (voir par exemple [14], [20], [29] et [30]), i.e. si $0 \leq \alpha < \beta \leq 1$, $(a_m)_{m\leq N}$ est une suite dans $[0,1]$ et $\alpha < \beta$, alors

$$\frac{1}{N}\left(|\{m \leq N | \alpha < a_m \leq \beta\}| - (\beta-\alpha)\right) \leq \frac{6}{M+1} + \frac{4}{\pi}\sum_{j=1}^{M}\left|\frac{1}{jN}\sum_{m=1}^{N} exp(2\pi i j a_m)\right|.$$

Ceci nous permet de montrer que

$$|\{1 \leq m \leq p-1 |\ \alpha < t_b(m)/m \leq \beta\}| = (\beta - \alpha)(p-1) + O\left(\frac{p}{\ln(p)}\right)$$

avec une constante explicite pour le terme de l'erreur qui peut être prise arbitrairement plus grande en valeur absolue que $12\pi \approx 37.6991118$ pour p suffisament grand.

Ces nombres et leurs généralisations offrent une foule d'autres questions intéréssantes pouvant avoir certaines implications en théorie des nombres.

1 Principes d'analyse de suites remarquables de $\mathbb{Z}_N$.

1.1 Notations et concepts de théorie additive.

On définit l'énergie entre deux sous-ensembles A et B d'éléments non nécessairement distincts de l'ensemble $\mathbb{A}$ par :

$$E^{\circ}_{\sim}(A,B) := |\{(a,b,a',b')|\ a \circ b \sim a' \circ b' \text{ avec } a,a' \in A \text{ et } b,b' \in B\}|.$$

Le symbole $\circ$ est une loi de composition quelconque et $\sim$ est une relation d'équivalence quelconque. Dans le cas où $\circ$ est l'addition on écrit $E_{\sim}(A,B)$ et on parle d'énergie additive. On écrit aussi $E^{\circ}_{\sim}(A,A) =: E^{\circ}_{\sim}(A)$ pour alléger. Clairement, on a que

$$E^{\circ}_{\sim}(A,B) \geq |A||B|,$$

de plus, si $a \circ b \sim x \circ b'$ et $a \circ b \sim a' \circ y$ on respectivement au plus X et Y solutions en x et y parmi l'ensemble des couples possibles lorsque les variables sont fixées dans les mêmes ensembles que la définition de l'énergie, alors

$$E^{\circ}_{\sim}(A,B) \leq |A||B| \min(Y|A|, X|B|)$$

Dans ce travail, on a souvent $\mathbb{A} = \mathbb{Z}$ et il est entièrement question de relation de congruence modulo N que nous allons noter $E^{\circ}_{N}(A,B)$. Nous allons aussi convenir naturellement que $E^{\circ}(A,B) := E^{\circ}_{\infty}(A,B)$ correspond au cas où $\sim$ est le $=$ ordinaire. Comme

$$|A||B| = |A+B| \Leftrightarrow |A||B| = |A-B| \Leftrightarrow (A-A) \cap (B-B) = \{0\},$$

on remarque aussi que $E_{\sim}(A,B) = |A||B|$ si et seulement si on est dans la situation $|A||B| = |A+B|$, i.e. $r_{A+B}(m) = 1$ pour tout élément $m \in A+B$ où on note en général

$$r_{A\circ B}(m) := |R_{A\circ B}(m)| = |R^{(1)}_{A\circ B}(m)| = |R^{(2)}_{A\circ B}(m)|$$

et où

$$R_{A\circ B}(m) := \{(a,b) \in A \times B|\ a \circ b \sim m\},$$

$$R^{(1)}_{A\circ B}(m) := \{a \in A|\ b \in B \text{ et } a \circ b \sim m\}$$

et

$$R^{(2)}_{A\circ B}(m) := \{b \in B|\ a \in A \text{ et } a \circ b \sim m\}.$$

Ces dernières définitions dépendent évidemment de $\sim$, cependant il est inutile de l'inclure dans la notation pour les cas qui nous concernent puisque le tout est intimement lié à la définition de l'énergie. Clairement, on a que

$$E_N(A,B) = E_N(B,A) = E_N(A+m, B+m) = E_N(-A,B)$$

et que

$$E^{\circ}_{d}(A,B) \geq E^{\circ}_{e}(A,B) \geq E^{\circ}(A,B)$$

si $d|e$. On remarque que

$$|A||B| = \sum_{m \in A+B} r_{A+B}(m) = \sum_{l \in A-B} r_{A-B}(l)$$

et que

$$E_N(A,B) = \sum_{m \in A+B} r_{A+B}(m)^2 = \sum_{l \in A-B} r_{A-B}(l)^2 = \sum_{k \in (A-A)\cap(B-B)} r_{A-A}(k) r_{B-B}(k).$$

On peut donc déduire, en utilisant l'inégalité de Cauchy-Schwarz, que

$$(|A||B|)^2 = \left(\sum_{m \in A+B} r_{A+B}(m) \right)^2 \leq |A+B| \; E_N(A,B)$$

et que

$$E_N(A,B)^2 = \left(\sum_{l \in (A-A)\cap(B-B)} r_{A-A}(l) r_{B-B}(l) \right)^2 \leq E_N(A) E_N(B).$$

On retrouve souvent ces identités fondamentales de base dans le cadre de la théorie additive des nombres. Un autres cas particulièrement important fait l'objet du lemme suivant.

Lemme 1.1. $\max_m r_{A+B}(m) \geq \dfrac{|A||B|}{|A+B|} \max \left\{ 1, \left(\dfrac{|A+B|}{|A-B|} \right)^{1/2} \right\}.$

Preuve. Comme

$$|A+B| \max_m \; r_{A+B}(m)^2 \geq \sum_{m \in A+B} r_{A+B}(m)^2 = E_N(A,B),$$

alors

$$E_N(A,B) \geq (|A||B|)^2 / \min\{|A+B|, |A-B|\}$$

et le résultat s'ensuit.□

Définition. Soit $N \in \mathbb{N}$ et $f : \mathbb{Z}_N \to \mathbb{C}$. La *transformée de Fourier discrète* $\hat{f}$ de f est définie par

$$\hat{f}(n) = \sum_{m=0}^{N-1} f(m) e\left(\frac{mn}{N} \right),$$

où $e(t) = \exp(2\pi i t)$. On vérifie facilement que son inverse est donné par

$$f(m) = \frac{1}{N} \sum_{n=0}^{N-1} \hat{f}(n) e\left(\frac{-mn}{N} \right),$$

ce qui s'écrit $\hat{\hat{f}}(-m) = N f(m)$. On définit la *convolution* $f * g$ de f et g par

$$(f * g)(n) := \sum_{i-j=n} f(i) \overline{g(j)}.$$

Une liste exaustive des propriétés que satisfait cette transformée prendrait un livre entier. Nous proposons ici seulement certains faits qui sont fondamentaux dans l'optique de notre étude.

1.2 Propriétés de la transformée de Fourier discrète.

Nous avons successivement

$$(1)\quad \sum_{n=0}^{N-1} \hat{f}(n)\bar{\hat{g}}(n) = N \sum_{n=0}^{N-1} f(n)\bar{g}(n);$$

$$(2)\quad \sum_{n=0}^{N-1} |\hat{f}(n)|^2 = N \sum_{n=0}^{N-1} |f(n)|^2;$$

$$(3)\quad \widehat{(f * g)} = \hat{f}\bar{\hat{g}};$$

$$(4)\quad \sum_{n=0}^{N-1} |\hat{f}(n)|^2 |\hat{g}(n)|^2 = N \sum_{n=0}^{N-1} |(f * g)(n)|^2;$$

$$(5)\quad \sum_{n=0}^{N-1} |\hat{f}(n)|^4 = N \sum_{a+b=c+d} f(a)f(b)\overline{f(c)f(d)}.$$

La seule idée fondamentale pour démontrer chacune de ces relations est l'identité bien connue

$$\sum_{i=0}^{N-1} e\left(\frac{ai}{N}\right) = \begin{cases} N & \text{si } a \equiv 0 \ (mod\ N), \\ 0 & \text{sinon,} \end{cases}$$

qui provient du fait que le membre de gauche reste inchangé si on le multiplie par $e(\frac{a}{N})$ et que ce dernier terme est différent de 1 pour $a \not\equiv 0$ modulo N.

L'identité (2) est souvent appelée *l'identité de Parseval* dans la littérature ; elle provient évidement de (1) dans le cas particulier où $f = g$. Le même phénomène est valide pour le passage de (4) à (5).

On définit le support d'une fonction $f : \mathbb{Z}_N \to \mathbb{C}$ par

$$supp(f) := \{x \in \mathbb{Z}_N | f(x) \neq 0\}.$$

En utilisant Cauchy-Schwarz et (2), on observe que

$$\begin{aligned} \max_n |f(n)| \leq \frac{1}{N} \sum_{m=0}^{N-1} |\hat{f}(m)| &\leq \frac{1}{N} (|supp(\hat{f})|)^{1/2} \left(\sum_{m=0}^{N-1} |\hat{f}(m)|^2 \right)^{1/2} \\ &\leq \left(\frac{|supp(f)||supp(\hat{f})|}{N} \right)^{1/2} \max_n |f(n)| \end{aligned}$$

et on en déduit que

$$|supp(f)||supp(\hat{f})| \geq N$$

si f n'est pas identiquement nul. Cette inégalité, qui est appelée le *principe d'incertitude*, ne peut être améliorée pour aucun N. Pour le voir, il s'agit de considérer la fonction caractéristique f d'un sous-groupe H de $\mathbb{Z}_N$. On remarque alors que $|supp(f)| = |H|$ et

$|supp(\hat{f})| = N/|H|$. Dans le cas où N est le nombre premier p, le dernier exemple nous fournit en fait

$$|supp(f)| + |supp(\hat{f})| = p + 1$$

et le prochain théorème, qui fut découvert par Terence Tao dans [34], affirme que cette version additive du principe d'incertitude est le pire cas possible, i.e.

$$|supp(f)| + |supp(\hat{f})| \geq p + 1,$$

pour toute fonction f non identiquement nulle. Il est fort possible que la même hypothèse entraîne que la borne

$$|supp(f)| + |supp(\hat{f})| \geq \min_{d|N} \left\{ d + \frac{N}{d} \right\}$$

soit vraie pour tout entier N, mais la preuve et le contre-exemple me sont inconnus. On peut montrer une certaine borne de ce genre pour un N général (voir [27]). En effet, soit $d_1 < d_2$ des diviseurs consécutifs de N. Si $d_1 \leq k := |supp(f)| \leq d_2$, alors on a

$$|supp(\hat{f})| \geq \frac{N}{d_1 d_2}(d_1 + d_2 - k).$$

Avant d'énoncer et de prouver le théorème de Tao, il faut deux propositions préliminaires.

Proposition 1.1. *Soit n un entier positif et $P(x_1, \ldots, x_n)$ un polynôme à coefficients entiers. Supposons qu'il existe n racines p-ièmes de l'unité $\beta_1, \ldots, \beta_n$ pas nécessairement distinctes pour lesquelles $P(\beta_1, \ldots, \beta_n) = 0$. Alors $P(1, \ldots, 1)$ est un multiple de p.*

Preuve. Écrivons $\beta = exp(2\pi i/p)$. Ainsi pour chaque $1 \leq i \leq n$ il existe $0 \leq k_i < p$ tel que $\beta_i = \beta^{k_i}$. Si on considère le polynôme $P(x) := P(x^{k_1}, \ldots, x^{k_n})$ à une variable, alors en le réduisant modulo $x^p - 1$ nous obtenons un polynôme de degré au plus $p - 1$. Puisque $\beta^p = 1$ et que par hypothèse $P(\beta) = 0$, il s'ensuit que $P(x)$ est un multiple du polynôme minimal de β qui est $(x^p - 1)/(x - 1)$ et la proposition est démontrée.□

Comme d'habitude, on définit l'espace $l^2(X)$ comme étant l'espace des fonctions $f : X \to \mathbb{C}$ qui satisfait

$$\sum_{x \in X} |f(x)|^2 < \infty.$$

Proposition 1.2. *Soit n une entier positif avec $1 \leq n \leq p$ et supposons que $A := \{a_1, \ldots, a_n\} \subseteq \mathbb{Z}_p$ et que $B := \{b_1, \ldots, b_n\} \subseteq \mathbb{Z}_p$. Supposons de plus que f est une fonction qui satisfait $supp(f) = A$. Soit $\vartheta : l^2(A) \to l^2(B)$ la transformation linéaire définie par $\vartheta(f) := \hat{f}|_B$, i.e., ϑ est la restriction de $\hat{f}$ à B. Alors ϑ est inversible.*

Preuve. Sans perte de généralité, les ensembles A et B sont ordonnés. Il est facile de voir qu'il s'agit simplement de montrer que la matrice $(\beta_{i,j})_{1 \leq i,j \leq n}$, où $\beta_{i,j} := exp(2\pi i a_i b_j/p)$, est à déterminant non nul. Posons $\beta_i = exp(2\pi i a_i/p)$; ainsi chaque β_i est une racine distincte de l'unité et on veut montrer que $det\ (\beta_i^{b_j})_{1 \leq i,j \leq n}$ est différent de zéro. Définissons le polynôme à n variables

$$D(x_1, \ldots, x_n) := det\ (x_i^{b_j})_{1 \leq i,j \leq n}.$$

On cherche un façon d'appliquer la proposition 1 à ce polynôme. Le problème, c'est que $D(1,\ldots,1)=0$. On remarque cependant la factorisation partielle

$$D(x_1,\ldots,x_n)=P(x_1,\ldots,x_n)\prod_{1\leq i<j\leq n}(x_j-x_i)$$

qui définit le polynôme $P(x_1,\ldots,x_n)$. On va montrer que $P(1,\ldots,1)$ n'est pas un multiple de p, ce qui va nous permettre de conclure, à l'aide de la proposition 1, que $P(\beta_1,\ldots,\beta_n)$ est non nul et par conséquent, $D(\beta_1,\ldots,\beta_n)$ est non nul . Pour ce faire, on considère l'expression

$$(x_1\partial_{x_1})^0(x_2\partial_{x_2})^1\ldots(x_n\partial_{x_n})^{n-1}D(x_1,\ldots,x_n)|_{x_1=\ldots=x_n=1}.$$

D'une part, on observe qu'on n'a jamais à dériver $P(x_1,\ldots,x_n)$ et en fait il n'est pas trop difficile de voir que la dernière expression est égale à

$$0!1!\ldots(n-1)!P(1,\ldots,1).$$

D'autre part, en utilisant le fait que $x_i\partial_{x_i}(x_i^{b_j})=b_jx_i^{b_j}$ et que le déterminant est une forme multilinéaire alternée, on observe que la dernière expression est en fait un déterminant de Vandermonde qui vaut exactement

$$\prod_{1\leq j<k\leq n}(b_k-b_j).$$

Comme les b_i sont tous distincts, il s'ensuit que $P(1,\ldots,1)$ n'est pas un multiple de p, et le résultat est démontré.□

Théorème (Principe additif d'incertitude). *Soit $f:\mathbb{Z}_p\to\mathbb{C}$ une fonction non identiquement nulle. Alors*

$$|supp(f)|+|supp(\hat{f})|\geq p+1.$$

Réciproquement, si A et B sont deux ensembles non vides avec $|A|+|B|\geq p+1$, alors il existe f tel que $supp(f)=A$ et $supp(\hat{f})=B$.

Preuve. Pour la première partie, on procède par contradiction. Supposons qu'il existe une fonction non identiquement nulle pour laquelle on ait $|supp(f)|+|supp(\hat{f})|\leq p$. On pose alors $supp(f)=A$. Par hypothèse, il existe un ensemble B disjoint de $supp(\hat{f})$ tel que $|A|=|B|$. On a alors $\vartheta(f)=0$, ce qui contredit la proposition 2.

Pour la seconde partie, il suffit de considérer le cas où $|A|+|B|=p+1$. Sinon, on aurait juste à montrer le résultat pour les sous-ensembles A' de A et B' de B pour lesquels $|A'|+|B'|=p+1$ et ensuite on déduirait le résultat en prenant une combinaison linéaire de ces sous-ensembles.

Supposons donc que $|A|+|B|=p+1$. On peut alors choisir un sous-ensemble A' de $\mathbb{Z}_p$ pour lequel $A'\cap B=\{x\}$. Comme ϑ est inversible, on peut trouver une fonction $f\in l^2(A)$ telle que $\hat{f}$ s'annule sur $A'-\{x\}$ et qui est non nulle en x. En utilisant la première partie du théorème, on voit qu'une telle fonction f doit être non nulle sur A

et de même $\hat{f}$ est non nulle sur B. On a donc $supp(f) = A$ et $supp(\hat{f}) = B$, d'où le résultat.□

Remarque. Le dernier théorème nous fournit une preuve très courte du fameux résultat de Davenport, i.e.

$$|A+B| \geq \min\{|A|+|B|-1, p\},$$

si $|A|$ et $|B|$ sont non nuls. En effet, dans ce cas on peut trouver deux ensembles X et Y tels que $|X| = p+1-|A|$, $|Y| = p+1-|B|$ et $|X \cap Y| = \max\{|X|+|Y|-1, 1\}$. Selon le dernier théorème, on peut aussi trouver une fonction f telle que $supp(f) = A$ et $supp(\hat{f}) = X$; de même, on peut trouver une fonction g telle que $supp(g) = B$ et $supp(\hat{g}) = Y$. On voit alors que la fonction $f*g$ est à support dans A+B et sa transformée de Fourier est à support dans $X \cap Y$. On a alors

$$|A+B| + |X \cap Y| \geq p+1.$$

On en déduit le résultat.

Par abus de notation, on utilise $A(\ \ )$ pour désigner la fonction caractéristique de l'ensemble A, i.e.

$$A(n) := \begin{cases} 1 & \text{si } n \in A, \\ 0 & \text{sinon.} \end{cases}$$

On utilise aussi souvent

$$A(x) := |\{a \in A|\ 1 \leq a \leq x\}| = \sum_{n=1}^{[x]} A(n).$$

Dans un contexte fixé, il est facile de ne pas se tromper de définition. La théorie additive des nombres s'inscrit naturellement dans le domaine des transformées de Fourier discrètes. Par exemple, $\hat{A}(n) = \sum_{a \in A} e(\frac{an}{N})$. Ainsi on a la formule de représentation $r_{A-B}(n) = (A*B)(n)$. De plus, on a que

$$\frac{1}{N}\sum_{n=0}^{N-1} |\hat{A}(n)|^2 |\hat{B}(n)|^2 = \sum_{n=0}^{N-1} |(A*B)(n)|^2 = E_N(A,B)$$

d'après (4). Comme

$$\hat{A}(m)\hat{B}(m) = \sum_{n=0}^{N-1} r_{A+B}(n) e\left(\frac{mn}{N}\right),$$

en inversant on obtient

$$r_{A+B}(n) = \frac{1}{N}\sum_{m=0}^{N-1} \hat{A}(m)\hat{B}(m) e\left(\frac{-mn}{N}\right).$$

En particulier,

$$r_{A-A}(n) = \frac{1}{N}\sum_{m=0}^{N-1} |\hat{A}(m)|^2 e\left(\frac{-mn}{N}\right)$$

et on a même en fait que

$$r_{A-A}(n) = \frac{1}{N}\sum_{m=0}^{N-1} |\hat{A}(m)|^2 \cos\left(\frac{2\pi mn}{N}\right),$$

étant donné que $\hat{A}(-m) = \overline{\hat{A}(m)}$.

Traditionnellement, les transformées de Fourier discrètes sont apparues pour comprendre comment un ensemble est réparti modulo N. Soit $(t)_N$ le plus petit résidu non-négatif de t modulo N.

Définition. Un ensemble $A \subseteq \mathbb{Z}_N$ est dit uniformément distribué modulo N si

$$|\{a \in A|\ \alpha N < (na)_N \leq \beta N\}| = (\beta - \alpha)|A|(1 + o(1))$$

pour chaque $n \not\equiv 0$ modulo N et pour chaque $0 \leq \alpha < \beta \leq 1$.

En pratique, le $o(1)$ peut être remplacé par une fonction qui est explicitée dans la preuve du théorème fondamental sur le sujet qui nous vient de Weyl et qui justifie presqu'à lui seul l'usage de sommes exponentielles en mathématiques.

Théorème d'équidistribution de Weyl. *Soit $A \subseteq \mathbb{Z}_N$. Alors A est uniformément distribué modulo N si et seulement si $\hat{A}(n) = o(|A|)$ pour tout $n \not\equiv 0$ modulo N.*

Preuve. Supposons que

$$|\{a \in A|\ \alpha N < (na)_N \leq \beta N\}| = (\beta - \alpha)|A|(1 + o(1))$$

pour $n \not\equiv 0$ modulo N et pour chaque $0 \leq \alpha < \beta \leq 1$. Si $\alpha N < (na)_N \leq \beta N$, alors $e(\frac{an}{N}) = e(\alpha) + O(|\beta - \alpha|)$. En subdivisant $(0, N]$ en intervales $I_j := (\frac{jN}{K}, \frac{(j+1)N}{K}]$ pour un K fixé et choisi de telle sorte que

$$\left|\left\{a \in A|\ \frac{jN}{K} < (na)_N \leq \frac{(j+1)N}{K}\right\}\right| = \frac{1}{K}|A|(1 + o(1)),$$

on trouve

$$\hat{A}(n) = \sum_{j=0}^{K-1} \sum_{\substack{a\in A \\ (na)_N \in I_j}} e\left(\frac{an}{N}\right) = \sum_{j=0}^{K-1}(1 + o(1))\frac{|A|}{K}\left(e\left(\frac{jn}{K}\right) + O\left(\frac{1}{K}\right)\right) \ll |A|/K.$$

En laissant $K = K(N) \to \infty$, on trouve $\hat{A}(n) = o(|A|)$. Dans l'autre sens, on a

$$\begin{aligned}\sum_{\substack{a\in A \\ [\alpha N] \leq (na)_N \leq [\beta N]}} 1 &= \sum_{j=[\alpha N]}^{[\beta N]} \sum_{a\in A} \frac{1}{N}\sum_{i=0}^{N-1} e\left(\frac{i(na-j)}{N}\right) \\ &= (\beta - \alpha)|A| + \frac{1}{N}\sum_{i=1}^{N-1} \hat{A}(in) \sum_{j=[\alpha N]}^{[\beta N]} e\left(\frac{-ij}{N}\right) + O(1).\end{aligned}$$

Pour $i \neq 0$ dans $(-N/2, N/2]$, on a

$$\left| \sum_{j=[\alpha N]}^{[\beta N]} e\left(\frac{-ij}{N}\right) \right| \ll N/|i|.$$

On a donc

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N-1} \hat{A}(in) \sum_{j=[\alpha N]}^{[\beta N]} e\left(\frac{-ij}{N}\right) &\ll \sum_{i=1}^{[N/2]} \frac{|\hat{A}(in)|}{i} \ll \sum_{i=1}^{M} \frac{|\hat{A}(in)|}{i} + \sum_{i=M+1}^{[N/2]} \frac{|\hat{A}(in)|}{i} \\ &\ll (\ln M) \max_{s\neq 0} |\hat{A}(s)| + \left(\sum_{i=1}^{N-1} |\hat{A}(in)|^2\right)^{1/2} \left(\sum_{i=M+1}^{[N/2]} \frac{1}{i^2}\right)^{1/2} \\ &\ll (\ln M) \max_{s\neq 0} |\hat{A}(s)| + (|A|N/M)^{1/2} = o(N)\end{aligned}$$

si on laisse $M \to \infty$ assez lentement.□

Remarque. L'avant-dernière ligne de cette preuve nous donne un terme d'erreur explicite pour la fonction $o(1)$ implicitement définie dans l'énoncé du théorème. Par exemple, si on peut montrer pour un certain ensemble A, où $|A| \gg N^{\delta}$ avec $\delta > 0$, que $\max_{s\neq 0} |\hat{A}(s)| \ll |A|^{1-\epsilon}$ avec $\epsilon > 0$, ce qui est le cas de plusieurs exemples intéressants, alors on peut prendre $M = \frac{N}{|A|^{1-2\epsilon}\ln^2(N/|A|^{1-2\epsilon})}$ et ainsi obtenir par exemple que

$$|\{a \in A|\ \alpha N < (na)_N \leq \beta N\}| = (\beta - \alpha)|A| \left(1 + O\left(\frac{\ln |A|}{\delta|\beta - \alpha||A|^{\epsilon}}\right)\right),$$

ce qui fournit un résultat non trivial pour toute région aussi mince que l'économie qu'on obtient avec l'analyse des sommes exponentielles associées à A.

La notion de distribution uniforme est reliée à la question : Existe-t-il des solutions à l'équation $a + b \equiv c$ avec $a \in A$, $b \in B$ et $c \in C$, trois ensembles de résidus modulo N ?

Proposition 1.3. *Soit $A, B, C \subseteq \mathbb{Z}_N$. Si A est uniformément distribué modulo N avec $|A| \gg N$ et si B, C sont deux ensembles quelconques d'entiers modulo N de cardinalité $\gg N$, alors pour chaque entier x, y et z premier avec N et pour chaque entier n on a que*

$$|\{a \in A, b \in B, c \in C|\ xa + yb + zc \equiv n\ (mod\ N)\}| = \frac{|A||B||C|}{N}(1 + o(1)).$$

Preuve. La cardinalité de cet ensemble est

$$\sum_{a\in A, b\in B, c\in C} \frac{1}{N}\sum_{i=0}^{N-1} e\left(\frac{i(xa + yb + zc - n)}{N}\right) = \frac{1}{N}\sum_{i=0}^{N-1} e\left(\frac{-in}{N}\right) \hat{A}(xi)\hat{B}(yi)\hat{C}(zi).$$

Le terme $i = 0$ donne $|A||B||C|/N$. On regarde le reste comme un terme d'erreur et on le borne par sa valeur absolue qui contribue à au plus

$$\frac{1}{N}\max_{i\neq 0}|\hat{A}(i)|\sum_{i=0}^{N-1}\hat{B}(yi)\hat{C}(zi) \leq \frac{1}{N}\max_{i\neq 0}|\hat{A}(i)|\left(\sum_{i=0}^{N-1}|\hat{B}(yi)|^2\right)^{1/2}\left(\sum_{i=0}^{N-1}|\hat{C}(yi)|^2\right)^{1/2}$$

$$= \frac{1}{N}\max_{i\neq 0}|\hat{A}(i)|(NB)^{1/2}(NC)^{1/2} = (BC)^{1/2}\max_{i\neq 0}|\hat{A}(i)| \leq N\max_{i\neq 0}|\hat{A}(i)|$$

en utilisant Cauchy-Schwarz et l'identité de Parseval. Le résultat découle du fait que $\hat{A}(i) = o(|A|)$ pour tout $i \not\equiv 0$ modulo N, car ça implique que $N\max_{i\neq 0}|\hat{A}(i)| = o(|A||B||C|/N)$.□

Remarque. Cette méthode est bien plus générale que les conditions qui y sont imposées. Par exemple, si on s'intéresse à savoir si un ensemble A possède une progression arithmétique modulo N de longueur trois, il s'agit de vérifier si l'équation $a + b \equiv 2c$ modulo N avec a,b et $c \in A$ tous différents modulo N. En d'autre mots, on veut que

$$|\{a, b, c \in A|\ a + b \equiv 2c\ (mod\ N)\}| > |A|.$$

La preuve nous donne qu'il faut que

$$|A|^3/N - |A|\max_{i\neq 0}|\hat{A}(i)| > |A|,$$

ce qui se réalise si $\max_{i\neq 0}|\hat{A}(i)| < |A|^2/N - 1$.

Les exemples les plus simples de sous-ensembles remarquables de $\mathbb{Z}_N$ sont les sous-groupes multiplicatifs de $\mathbb{Z}_p$. Ces sous-groupes sont exactement les ensembles de puissances k où $k|(p-1)$. C'est l'exemple typique de base. Voici un outil fondamental pour leur étude.

Définition. Pour chaque caractère χ modulo N, on définit la somme de Gauss $\tau(\chi)$ par

$$\tau(\chi) := \sum_{n=0}^{N-1}\chi(n)e\left(\frac{n}{N}\right).$$

Si $(m, N) = 1$, alors on peut écrire

$$\chi(m)\tau(\bar{\chi}) = \sum_{n=0}^{N-1}\bar{\chi}(n)\chi(m)e\left(\frac{n}{N}\right) = \sum_{i=0}^{N-1}\chi(i)e\left(\frac{mi}{N}\right),$$

en posant $n \equiv im$ modulo N. On obtient une représentation de $\chi(m)$ par la formule très importante

$$\chi(m) = \frac{1}{\tau(\bar{\chi})}\sum_{i=0}^{N-1}\chi(i)e\left(\frac{mi}{N}\right),$$

en supposant en plus que $\tau(\bar{\chi}) \neq 0$. Un exemple classique d'application de cette formule est une démonstration de l'importante inégalité de Polya-Vinogradov.

Pour simplifier l'analyse des sommes de Gauss et pour ce qui suit, nous allons supposer à partir de maintenant que N est le nombre premier p. Dans ce cas, on a

$$|\tau(\chi)|^2 = \sum_{n=0}^{p-1}\sum_{m=0}^{p-1} \chi(n)\bar{\chi}(m)e\left(\frac{n-m}{p}\right) = \sum_{n=0}^{p-1}\sum_{k=0}^{p-1} \chi(n)\bar{\chi}(nk)e\left(\frac{n(1-k)}{p}\right),$$

où on a posé $m \equiv kn$ modulo p. Donc

$$|\tau(\chi)|^2 = \sum_{k=1}^{p-1} \bar{\chi}(k) \sum_{n=1}^{p-1} e\left(\frac{n(1-k)}{p}\right) = p - 1 - \sum_{k=2}^{p-1} \bar{\chi}(k) = p,$$

de sorte que $|\tau(\chi)| = p^{1/2}$. Posons maintenant

$$G_k := \{n \in \mathbb{Z}_p^* |\ \exists\ m \text{ tel que } m^k = n\}.$$

Clairement, $G_k = G_{(k,p-1)}$; donc on suppose toujours que k divise $p-1$ et on note $h := (p-1)/k$. De plus,

$$G_k = \{x \in \mathbb{Z}_p |\ x^h = 1\} = \{g^{jk} \in \mathbb{Z}_p |\ j = 1, \ldots, h\}$$

où g est un générateur de $\mathbb{Z}_p^*$. La transformée de Fourier de la fonction caractéristique de cet ensemble est

$$\hat{G}_k(a) = \sum_{i=1}^{h} e\left(\frac{ag^{ik}}{p}\right) = \frac{1}{k}\sum_{n=1}^{p-1} e\left(\frac{an^k}{p}\right).$$

On remarque que pour $a \not\equiv 0$ modulo p, la somme de droite vaut exactement

$$\sum_{\substack{\chi^k = \chi_0 \\ \chi \neq \chi_0}} \bar{\chi}(a)\tau(\chi),$$

car

$$\sum_{m=0}^{p-1} e\left(\frac{m}{p}\right) \sum_{\substack{\chi^k = \chi_0 \\ \chi \neq \chi_0}} \bar{\chi}(a)\chi(m) = \sum_{m=0}^{p-1} e\left(\frac{m}{p}\right) \sum_{b=1}^{k-1} e\left(\frac{bv(m/a)}{k}\right)$$

$$= (k-1) \sum_{\substack{m/a=1 \\ m/a \equiv n^k \bmod p}}^{p-1} e\left(\frac{m}{p}\right) - \sum_{\substack{m/a=1 \\ m/a \not\equiv n^k \bmod p}}^{p-1} e\left(\frac{m}{p}\right).$$

Comme $|\tau(\chi)| = p^{1/2}$, on déduit que

$$|\hat{G}_k(a)| \leq \frac{k-1}{k} p^{1/2} < p^{1/2},$$

ce qui est non trivial dès que $k \leq p^{1/2}$. Une façon encore plus simple de montrer cette inégalité est d'utiliser le fait que $\hat{G}_k(a) = \hat{G}_k(am^k)$ lorsque $m \not\equiv 0$ modulo p, ce qui permet d'écrire

$$(p-1)|\hat{G}_k(a)|^2 = \sum_{m=1}^{p-1} |\hat{G}_k(am^k)|^2 \leq k \sum_{n=1}^{p-1} |\hat{G}_k(n)|^2.$$

D'où

$$h|\hat{G}_k(a)|^2 = \sum_{m_1,m_2=1}^{h} \sum_{n=1}^{p-1} e\left(\frac{(g^{km_1} - g^{km_2})n}{p}\right) \leq h(p-1).$$

On a donc démontré les deux corollaires suivants.

Corollaire 1.1. *Soit G_k le sous-groupe de $\mathbb{Z}_p^*$ d'ordre h. Alors G_k est réparti uniformément modulo p dès que $k = o(p^{1/2}/\ln p)$. De plus, on a l'estimé suivant :*

$$|\{n \in G_k |\ \alpha p < n \leq \beta p\}| = (\beta - \alpha)\left(\frac{p-1}{k}\right)\left(1 + O\left(\frac{k\ln(p)}{|\beta - \alpha|p^{1/2}}\right)\right).$$

Corollaire 1.2. *Il existe au moins une solution non triviale à l'équation de Fermat modulo p lorsque $k < p^{1/4} - 1$, i.e.*

$$x^k + y^k \equiv z^k \ (mod\ p)$$

possède une solution avec $xyz \not\equiv 0$ modulo p.

1.3 Présentation et application de la méthode de Stepanov.

Pour étendre le domaine de validité du corollaire 1.1, il faut introduire une nouvelle famille d'idées qui forment en quelque sorte le sujet central de ce mémoire. Il s'agit d'une façon élémentaire de tantôt compter, tantôt majorer le nombre de solutions d'une équation définie sur un corps fini. Comme on le verra, la même idée sert depuis cent ans dans l'étude des mesures d'irrationalité des nombres algébriques. Cette méthode de Thue fut adaptée par Stepanov, il y a de cela environ 40 ans, à l'étude des corps finis et elle est encore améliorée et généralisée de nos jours. Je me propose d'élaborer cette théorie dans le sens inverse dont l'histoire l'a formée, i.e., après avoir étendu et approfondi le corolaire 1.1, je vais faire un traitement personnalisé de certaines troncations des fonctions "transcendantes" $-\ln(1-x)$ et $\exp(x)$ dans $\mathbb{F}_p$ pour finir avec une preuve de l'hypothèse de Riemann dans $\mathbb{F}_q$. Ensuite, je vais démontrer le fameux théorème de Thue pour mettre en évidence les points communs.

Définition. Soit $K[x]$ l'anneau des polynômes sur K. Soit ∂ l'opérateur de dérivation défini par

$$\partial(\alpha_0 + \alpha_1 x + \ldots + \alpha_t x^t) = \alpha_1 + 2\alpha_2 x + \ldots + t\alpha_t x^{t-1}.$$

On écrit souvent $\partial^k P(x) = P^{(k)}(x)$ pour alléger la notation.

Lemme 1.2. *Soit $\mathbb{F}_q$ un corps de caractéristique p et soit $m \leq p$ un entier positif. Supposons que $P(x) \in \mathbb{F}_q[x]$ avec deg $P(x) \geq m$ et que pour un certain $\beta \in \mathbb{F}_q$ on ait*

$$0 = P(\beta) = P^{(1)}(\beta) = P^{(2)}(\beta) = \ldots = P^{(m-1)}(\beta).$$

Alors $(x-\beta)^m | P(x)$, i.e. $P(x)$ a un zéro d'ordre au moins m en β.

Preuve. Il est facile de voir que pour chaque $\beta \in \mathbb{F}_q$ on peut trouver $c_1, \ldots, c_t$ tels que

$$P(x) := \alpha_0 + \alpha_1 x + \ldots + \alpha_t x^t = c_0 + c_1(x-\beta) + \ldots + c_t(x-\beta)^t.$$

Alors,

$$P^{(k)}(x) = k! \left[c_k + c_{k+1} \binom{k+1}{k} (x-\beta) + \ldots + c_t \binom{t}{k} (x-\beta)^{t-k} \right].$$

En évaluant en β et en utilisant l'hypothèse pour chaque $0 \leq k \leq m-1$, on a que $0 = k!c_k$. Comme $k \leq m-1 < p$, il s'ensuit que $k! \neq 0$ dans $\mathbb{F}_q$. Donc $c_k = 0$ pour $0 \leq k \leq m-1$, d'où $(x-\beta)^m | P(x)$.□

Remarques. La condition $m \leq p$ est essentielle. Par exemple, considérons le polynôme $P(x) = (x(x-1))^p$. Les dérivées s'annulent en 0 pour $0 \leq m \leq 2p$, mais $P(x)$ a un zéro d'ordre seulement p en $x = 0$. Pour contrer ce phénomène qui provient évidement du facteur factoriel qui apparaît à force de dériver, Hasse eut l'idée de définir l'hyperdérivée ϱ^k par

$$\varrho^k(\alpha_0 + \alpha_1 x + \ldots + \alpha_t x^t) := \alpha_k + \binom{k+1}{k} \alpha_{k+1} x + \ldots + \binom{t}{k} \alpha_t x^{t-k}.$$

On écrit souvent $\varrho^k P(x) = P^{[k]}(x)$ pour alléger l'écriture. Formellement, $\varrho^k = \partial^k / k!$ et il possède évidement des propriétés semblables à la dérivé ordinaire. En particulier, on peut écrire le lemme 1.2 sans se limiter à $m \leq p$. Cet opérateur sera beaucoup utilisé aux chapitres 3 et 4.

Lemme 1.3. *Soit $P(x) \in \mathbb{Z}_p[x]$ une somme de $j \geq 1$ termes distincts. Supposons que deg $P(x) < p$. Alors $(x-1)^j$ ne divise pas $P(x)$.*

Preuve. La preuve se fait par induction complète. Le cas $j = 1$ est trivial. Supposons que $j > 1$ et que le résultat est démontré pour tout polynôme de degré $1 \leq i < j$. Soit $P(x) = \sum_{i=1}^{j} \alpha_i x^{l_i}$. Alors

$$xP^{(1)}(x) - l_i P(x) = \sum_{i=1}^{j} l_i(\alpha_i - \alpha_j) x^{l_i}.$$

Le polynôme ainsi formé possède au plus $j-1$ termes. On voit alors que $(x-1)^j$ ne peut pas diviser $P(x)$, car sinon $(x-1)^{j-1}$ diviserait $xP^{(1)}(x) - l_i P(x)$ ce qui contredit l'hypothèse d'induction.□

Remarque. Le principe de Tao nous fournit un résultat d'une tout autre saveur. En effet, on apprend qu'un polynôme à $j+1$ termes ne peut pas posséder plus de j racines dans l'ensemble $\{x \in \mathbb{C} |\ x^p = 1\}$. Un tel polynôme étant essentiellement la transformée de Fourier d'une fonction supportée sur un ensemble à $j+1$ éléments, sa transformée doit avoir un support d'au moins $p-j$ éléments, i.e. au plus j valeurs de $\{x \in \mathbb{C} |\ x^p = 1\}$ peuvent être une racine du polynôme. En utilisant l'argument du lemme 1.3, on déduit alors que le nombre de racines provenant de cet ensemble, comptées avec multiplicité, d'un tel polynôme ne peut pas dépasser j^2.

Une borne sur l'énergie additive de G_k fournit naturellement une majoration de la transformée de Fourier correspondante. Plus précisément, on montre le résultat suivant.

Lemme 1.4. *L'estimé suivant est valide pour tout k lorsque $a \not\equiv 0$ modulo p :*

$$\hat{G}_k(a) \leq (\min\{k, p^{1/2}\}E_p(G_k))^{1/4}.$$

Preuve. Clairement, pour $m \not\equiv 0$ modulo p, on a que

$$\hat{G}_k(a) = \hat{G}_k(am^k).$$

On en déduit donc que

$$(p-1)|\hat{G}_k(a)|^4 = \sum_{m=1}^{p-1} |\hat{G}_k(am^k)|^4 \leq k \sum_{n=1}^{p-1} |\hat{G}_k(n)|^4.$$

Comme chaque valeur revient ou bien 0 ou bien k fois. On a alors

$$h|\hat{G}_k(a)|^4 = \sum_{m_1,m_2,m_3,m_4=1}^{h} \left(\sum_{n=1}^{p-1} e\left(\frac{(g^{km_1} + g^{km_2} - g^{km_3} - g^{km_4})n}{p} \right) \right) \leq (p-1)E_p(G_k),$$

ce qui fournit $|\hat{G}_k(a)| \leq (kE_p(G_k))^{1/4}$. De même, pour l'autre partie, on a

$$h|\hat{G}_k(a)|^2 = \sum_{n=1}^{h} |\hat{G}_k(ag^{nk})|^2 = \sum_{t=0}^{p-1} r_{G_k-G_k}(t)\hat{G}_k(at).$$

En applicant trois fois l'inégalité de Hölder, on obtient

$$h^4|\hat{G}_k(a)|^8 \leq \left(\sum_{t=0}^{p-1} r_{G_k-G_k}(t) \right)^2 \sum_{t=0}^{p-1} r_{G_k-G_k}^2(t) \sum_{t=0}^{p-1} |\hat{G}_k(at)|^4 = ph^4 E_p(G_k)^2,$$

ce qui complète la preuve.□

Lemme 1.5. *Pour chaque $k > p^{1/3}$, on a*

$$E_p(G_k) < 23h^{5/2}.$$

Preuve. Considérons d'abord $R_{G_k-G_k}(n)$. Clairement, $r_{G_k-G_k}(0) = h$ et on peut écrire

$$E_p(G_k) = \sum_{n=0}^{p-1} r_{G_k-G_k}(n)^2 = h^2 + \sum_{n=1}^{p-1} r_{G_k-G_k}(n)^2 = h^2 + h \sum_{n \in \mathbb{Z}_p^*/G_k} r_{G_k-G_k}(n)^2.$$

De même,

$$h^2 = \sum_{n=0}^{p-1} r_{G_k-G_k}(n) = h + \sum_{n=1}^{p-1} r_{G_k-G_k}(n) = h + h \sum_{n \in \mathbb{Z}_p^*/G_k} r_{G_k-G_k}(n),$$

d'où on déduit que

$$\sum_{n\in\mathbb{Z}_p^*/G_k} r_{G_k-G_k}(n) = h-1.$$

Considérons maintenant

$$n^{-1}R^{(1)}_{G_k-G_k}(n) = R^{(1)}_{n^{-1}G_k-n^{-1}G_k}(1)$$

pour tout $n \in \mathbb{Z}_p^*/G_k$. On se fixe alors un sous-ensemble quelconque S de $\mathbb{Z}_p^*/G_k$ et on pose

$$M := \bigcup_{n\in S} n^{-1}R^{(1)}_{G_k-G_k}(n).$$

On commence maintenant à appliquer la méthode de Stepanov dans le but de borner $|M|$. Choisissons un polynôme $\Gamma(x,y,z) \in \mathbb{Z}_p[x,y,z]$ pour lequel

$$deg_x\ \Gamma < A,\ deg_y\ \Gamma < B,\ deg_z\ \Gamma < B.$$

Nous orientons notre choix dans le but que le polynôme $\Theta(x) := \Gamma(x, x^h, (x-1)^h)$ ait un zéro d'ordre au moins D en chaque point $\beta \in M$. On pourra alors conclure que $D|M| \leq deg\ \Theta(x)$, à la condition que $\Theta(x)$ ne soit pas identiquement nul. On remarque que

$$deg\ \Theta \leq deg_x\ \Gamma + h(deg_y\ \Gamma) + h(deg_z\ \Gamma) < A + 2hB.$$

Ainsi

$$|M|D < A + 2hB,$$

si $\Theta(x)$ n'est pas identiquement nul.

Pour faire en sorte que $\Theta(x)$ ait un zéro de multiplicité au moins D en un point β, il faut que

$$\Theta^{(i)}(\beta) = 0 \quad \text{pour} \quad i < D. \tag{1}$$

On observe que $\beta \neq 0, 1$ pour $\beta \in M$. Donc l'expression précédente est équivalente à

$$(\beta(\beta-1))^i\Theta^{(i)}(\beta) = 0 \quad \text{pour} \quad i < D.$$

Clairement, on a

$$\begin{aligned} x^m(x^a)^{(m)} &= (a)_m x^a \quad \text{si} \quad m \leq a, \\ x^m(x^{hb})^{(m)} &= (hb)_m x^{hb} \quad \text{si} \quad m \leq hb \end{aligned}$$

et

$$(x-1)^m((x-1)^{hc})^{(m)} = (hc)_m(x-1)^{hc} \quad \text{si} \quad m \leq hc.$$

Il s'ensuit que

$$(x(x-1))^i(x^a x^{hb}(x-1)^{hc})^{(i)} = P_{a,b,c,i}(x)x^{hb}(x-1)^{hc},$$

où $P_{a,b,c,i}(x)$ est un polynôme qui est ou bien identiquement nul, ou bien de degré $a+i$. On en déduit que

$$(\beta(\beta-1))^i(\beta^a\beta^{hb}(\beta-1)^{hc})^{(i)} = P_{a,b,c,i}(\beta)n^{-hb-hc}$$

pour chaque $\beta \in n^{-1}R^{(1)}_{G_k-G_k}(n)$, étant donné que $\beta^h = (\beta-1)^h = n^{-h}$ pour ces β. On écrit maintenant

$$\Gamma(x,y,z) = \sum_{a,b,c} \lambda_{a,b,c} x^a y^b z^c$$

et

$$P_{n,i}(x) := \sum_{a,b,c} \lambda_{a,b,c} n^{-hb-hc} P_{a,b,c,i}(x),$$

de sorte que $deg\ P_{n,i}(x) < A+i$ et

$$(\beta(\beta-1))^i \Theta^{(i)}(\beta) = P_{n,i}(\beta)$$

pour chaque $\beta \in n^{-1}R^{(1)}_{G_k-G_k}(n)$.

On s'arrange maintenant pour que $P_{n,i}(\beta) = 0$ pour $i < D$ pour chaque $n \in M$ en choisissant les $\lambda_{a,b,c}$ correctement. Ceci nous assure que (1) est satisfait pour $n \in M$. Chacun des $P_{n,i}(x)$ a au plus $A+i$ coefficients, lesquels sont des formes linéaires en $\lambda_{a,b,c}$. Si

$$(DA + D(D-1)/2)|S| < AB^2,$$

alors il y a un ensemble de coefficients $\lambda_{a,b,c}$, non tous nuls, pour lesquels les $P_{n,i}(x)$ sont identiquement nuls pour chaque $i < D$ et $n \in M$. On doit maintenant s'assurer que $\Theta(x)$ n'est pas identiquement nul si $\Gamma(x,y,z)$ ne l'est pas. On écrit

$$\Gamma(x,y,z) = \sum_c \Gamma_c(x,y) z^c,$$

et on prend c_0, la plus petite valeur de c pour laquelle $\Gamma_c(x,y)$ n'est pas identiquement zéro. Il s'ensuit que

$$\Theta(x) = (x-1)^{hc_0} \sum_{c_0 \le c < B} \Gamma_c(x,x^h)(x-1)^{(c-c_0)h},$$

de sorte que si $\Theta(x) \equiv 0$, on doit avoir

$$\Gamma_{c_0}(x,x^h) = 0\ (mod\ (x-1)^h \mathbb{Z}_p[x]).$$

En utilisant le lemme 1.3, on voit que c'est impossible si

$$AB \le h \quad \text{et} \quad A + hB < p.$$

En résumé, si on choisit A, B et D tels que

$$(DA + D(D-1)/2)|S| < AB^2, \quad AB \le h \quad \text{et} \quad A + hB < p,$$

alors on a

$$|M| < \frac{A+2hB}{D}.$$

On pose alors $A := [h^{2/3}/(2.1|S|)^{1/3}]$, $B := [(2.1h|S|)^{1/3}]$ et $D := A$, et on déduit que

$$|M| < 2.1^{5/3}(h|S|)^{2/3} \quad \text{si} \quad h^{2/3}/(2.1|S|)^{1/3} + (2.1h^4|S|)^{1/3} < p.$$

Maintenant, fort de cette borne, on est en mesure de terminer la preuve. Pour ce faire, on ordonne les h valeurs de $r_{G_k-G_k}(n)$ pour $n \in \mathbb{Z}_p^*/G_k$ comme suit :

$$r_{G_k-G_k}(n_1) \geq r_{G_k-G_k}(n_2) \geq \ldots \geq r_{G_k-G_k}(n_h).$$

On prend pour S l'ensemble des n_i avec $i < t$. Ainsi

$$r_{G_k-G_k}(n_t)t \leq |M| < 2.1^{5/3}(ht)^{2/3},$$

D'où on obtient que

$$r_{G_k-G_k}(n_t) < \frac{2.1^{5/3}h^{2/3}}{t^{1/3}}$$

pour tout $t \leq h$. On peut donc écrire

$$\sum_{t=1}^{T} r_{G_k-G_k}(n_t)^2 < 2.1^{10/3}h^{4/3}\sum_{t=1}^{T}\frac{1}{t^{2/3}} < 2.1^{10/3}3h^{4/3}T^{1/3}$$

et de même

$$\sum_{t=T+1}^{h} r_{G_k-G_k}(n_t)^2 < \frac{2.1^{5/3}h^{5/3}}{T^{1/3}}.$$

On pose $T = [h^{1/2}/33.2]$ et on montre ainsi que

$$\sum_{t=1}^{h} r_{G_k-G_k}(n_t)^2 < 22.5h^{3/2},$$

d'où le résultat.□

Théorème 1.1. *Les bornes suivantes sont valides pour $a \neq 0$:*

$$\hat{G}_k(a) \leq \begin{cases} p^{1/2} & \text{si } k \leq 23^{2/3}p^{1/3}, \\ 23^{1/4}p^{5/8}/k^{3/8} & \text{si } 23^{2/3}p^{1/3} < k \leq p^{1/2}, \\ 23^{1/4}p^{3/4}/k^{5/8} & \text{si } p^{1/2} < k \leq p^{2/3}/23^{2/3}, \\ h & \text{sinon.} \end{cases}$$

Il est naturel de se demander de quel ordre de grandeur doit être un sous-groupe de $\mathbb{Z}_p^*$ pour que $\hat{G}_k(a) = o(h)$. Il y a plusieurs auteurs qui ont apportée une réponse partielle à cette question. Il est possible d'améliorer les résultats présentés dans ce chapitre jusqu'à $h = O(p^{1/4+\epsilon})$ en utilisant le même genre de méthode, ce travail est fait dans [25]. Le meilleur résultat connu à ce jour sur le sujet, dans ([13]), est que $\hat{G}_k(a) = o(h)$ pour h aussi petit que $p^{c/\ln\ln(p)}$ lorsque p est suffisamment grand, pour une certaine constante c. C'est un peu pour ces raisons qu'il ne faut pas trop prendre au sérieux les constantes qui apparaissent dans le dernier théorème. Il faut les voir comme la limite de la méthode utilisée.

Définitions. Les éléments de $\mathbb{Z}_p[x]$ qui suivent sont des troncations naturelles des fonctions $-\ln(1-x)$ et $\exp(x)$ conventionnelles :

$$f_p(x) := \sum_{n=0}^{p-1} \frac{x^n}{n},$$

$$g_p(x) := \sum_{n=0}^{p-1} \frac{x^n}{n!}.$$

Pour alléger la notation, on écrit souvent $g(x)$ au lieu de $g_p(x)$ lorsque le contexte est totalement clair. On fait de même pour l'ensemble des fonctions du même genre.

Lemme 1.6. *La fonction $g(x)$ satisfait l'équation différentielle suivante pour $0 \leq n \leq p$:*

$$x^n g^{(n)}(x) = x^n g(x) + x^p h_n(x),$$

où

$$h_n(x) := \sum_{i=0}^{n-1} (-1)^i i! x^{n-1-i}.$$

Preuve. La preuve se fait par induction sur n. Le résultat est trivial pour $n = 0$. Supposons que c'est vrai pour $m < n$. On a alors

$$x\partial(x^{n-1}g^{(n-1)}(x)) = (n-1)x^{n-1}g^{(n-1)}(x) + x^n g^{(n)}(x).$$

En utilisant l'hypothèse d'induction et le fait que

$$g^{(1)}(x) = g(x) - \frac{x^{p-1}}{(p-1)!} = g(x) + x^{p-1}$$

selon le théorème de Wilson, on obtient successivement

$$\begin{aligned} x\partial(x^{n-1}g^{(n-1)}(x)) &= (n-1)x^{n-1}g(x) + x^n g^{(1)}(x) + x^{p+1}h_{n-1}^{(1)}(x) \\ &= (n-1)x^{n-1}g(x) + x^n g(x) + x^{p+n-1} + x^{p+1}h_{n-1}^{(1)}(x). \end{aligned}$$

On a alors que

$$x^n g^{(n)}(x) = x^n g(x) + (n-1)x^{n-1}g(x) - (n-1)x^{n-1}g^{(n-1)}(x) + x^{p+n-1} + x^{p+1}h_{n-1}^{(1)}(x).$$

Donc en utilisant l'hypothèse d'induction une seconde fois, on a

$$x^n g^{(n)}(x) = x^n g(x) - (n-1)x^p h_{n-1}(x) + x^{p+n-1} + x^{p+1}h_{n-1}^{(1)}(x).$$

Il s'agit maintenant de voir que $h_n(x)$ satisfait

$$h_n(x) = xh_{n-1}^{(1)}(x) + x^{n-1} - (n-1)h_{n-1}(x),$$

ce qui se vérifie facilement et ainsi le résultat est démontré.□

Lemme 1.7. *Soit $P_i(x)$ une collection de n polynômes non tous nuls de $\mathbb{Z}_p[x]$ qui satisfont deg $P_i(x) \leq k_i$ pour $0 \leq i \leq n$. Posons $m := \sum_{i=0}^{n} k_i$ et supposons que $m < p - n$. Soit*

$$G_n(x) := \sum_{i=0}^{n} P_i(x) g^i(x).$$

Alors $x^{m+n+1} \nmid G_n(x)$.

Preuve. Définissons la valuation I suivante :

$$I\left(\sum_{i=0}^{n} P_i(x) g^i(x)\right) := \sum_{i=0}^{n} deg\ P_i(x) + n_0,$$

où n_0 est le nombre de polynômes $P_i(x)$ non identiquement nuls, avec $0 \leq i \leq n$, et où par hypothèse, $n_0 \geq 1$. La preuve se fait par induction complète sur $I(G_n(x))$. Le cas $I(G_n(x)) = 1$ provient du fait que x ne divise ni $g(x)$ ni une constante non nulle. Supposons que le résultat a été démontré pour tout $G_n(x)$ avec $I(G_n(x)) < N$. Considérons $G_n(x)$ avec $I(G_n(x)) = N$ et supposons que $P_k(x)$ est le polynôme non identiquement nul qui a le plus petit indice. On peut alors écrire

$$G_n(x) = g^k(x) \sum_{i=k}^{n} P_i(x) g^{i-k}(x)$$

et on pose

$$H_n(x) := \sum_{i=k}^{n} P_i(x) g^{i-k}(x).$$

On a donc

$$H_n^{(1)}(x) = \sum_{i=k}^{n} P_i^{(1)}(x) g^{i-k}(x) + (i-k) P_i(x) g^{i-k-1}(x) g^{(1)}(x),$$

d'où

$$H_n^{(1)}(x) = \sum_{i=k}^{n} P_i^{(1)}(x) g^{i-k}(x) + (i-k) P_i(x) g^{i-k}(x) + (i-k) P_i(x) g^{i-k-1}(x) x^{p-1}$$

$$\equiv \sum_{i=k}^{n} \left(P_i^{(1)}(x) + (i-k) P_i(x)\right) g^{i-k}(x)\ (mod\ x^{p-1}\mathbb{Z}_p[x]),$$

car $g^{(1)}(x) = g(x) + x^{p-1}$. Posons

$$H(x) := \sum_{i=k}^{n} (P_i^{(1)}(x) + (i-k) P_i(x)) g^{i-k}(x).$$

On voit que si $x^{I(G_n(x))} | G_n(x)$ avec $I(G_n(x)) < p$, alors

$$x^{I(G_n(x))-1} | (H(x) - (n-k) H_n(x)).$$

Explicitement, on a que

$$H(x) - (n-k)H_n(x) \equiv$$
$$\sum_{i=k}^{n} \left(P_i^{(1)}(x) + (i-k)P_i(x) - (n-k)P_i(x) \right) g^{i-k}(x) \ (mod\ x^{p-1}\mathbb{Z}_p[x]).$$

Comme

$$P_n^{(1)}(x) + (n-k)P_n(x) - (n-k)P_n(x) = P_n^{(1)}(x),$$

on déduit que

$$I(H(x) - (n-k)H_n(x)) < I(H_n(x)) = I(G_n(x)),$$

une contradiction. Comme $I(G_n(x)) \leq n + m + 1$, la preuve est complète.□

Définition. Soit A, B et C trois sous ensembles d'éléments de $\mathbb{A}$ où ceux de A et C sont non nécessairement distincts contrairement à ceux de B. On dit qu'une loi de composition $\circ : A \times B \to C$ est injective si les fonctions $\iota_a : \{a\} \times B \to C$ avec $\iota_a(b) = a \circ b$ sont injectives pour tout $a \in A$.

Lemme 1.8. *Soit $\circ$ injective. Alors*

$$r_{g(\mathbb{Z}_p) \circ g(\mathbb{Z}_p)}(m) < 29p^{3/2}$$

pour tout m.

Preuve. Nous voulons en particulier borner $E_p(g(\mathbb{Z}_p), \{0\})$. La même preuve nous fournit une borne pour $r_{g(\mathbb{Z}_p) \circ g(\mathbb{Z}_p)}(m)$ et on a

$$p = \sum_{n=0}^{p-1} r_{g(\mathbb{Z}_p) - \{0\}}(n)$$

et

$$r_{g(\mathbb{Z}_p) \circ g(\mathbb{Z}_p)}(m) \leq \sum_{n=0}^{p-1} r_{g(\mathbb{Z}_p) - \{0\}}(n)^2.$$

Soit S un sous-ensemble quelconque de $\mathbb{Z}_p$. On veut choisir un polynôme $\Gamma(x, y, z)$ de $\mathbb{Z}_p[x, y, z]$ pour que $\Theta(x) := \Gamma(x, g(x), x^p)$ ait un zéro d'ordre au moins D en chacun des points t de S. Soit $k \in g^{-1}(S)$. Supposons que

$$deg_x\ \Gamma < A,\ deg_y\ \Gamma < B,\ deg_z\ \Gamma < C,$$

de sorte que $deg\ \Theta(x) < A + p(B + C)$. Posons

$$M := \sum_{t \in S} r_{g(\mathbb{Z}_p) - \{0\}}(t).$$

On aura alors $MD < A + p(B + C)$ si $\Theta(x)$ n'est pas identiquement nul. Pour $k \neq 0$, il est suffisant de faire en sorte que

$$k^i \Theta(k)^{(i)} = 0$$

pour $0 \leq i < D$.

Pour un terme $x^a g^b(x) x^{cp}$ typique, on a

$$x^i (x^a g^b(x) x^{cp})^{(i)} = x^{cp+i} (x^a g(x)^b)^{(i)}$$

$$= x^{cp+i} \sum_{\xi+\eta_1+\ldots+\eta_b=i} \binom{i}{\xi, \eta_1, \ldots, \eta_b} (x^a)^{(\xi)} g^{(\eta_1)}(x) \ldots g^{(\eta_b)}(x)$$

$$= x^{cp} \sum_{\xi+\eta_1+\ldots+\eta_b=i} \binom{i}{\xi, \eta_1, \ldots, \eta_b} (a)_\xi x^a [x^{\eta_1} g(x) + x^p h_{\eta_1}(x)] \ldots [x^{\eta_b} g(x) + x^p h_{\eta_b}(x)].$$

Donc

$$x^i (x^a g^b(x) x^{cp})^{(i)} \equiv x^{a+c} \sum_{j=0}^{b} P_j^{(i)}(x) g^j(x) \ (mod\ (x^p - x)\mathbb{Z}_p[x]),$$

où $deg\ P_j^{(i)}(x) \leq i$ pour $j = 0, \ldots, b$. De plus, on a $deg\ P_b^{(i)}(x) = i$, puisque $P_b^{(i)}(x) = b^i x^i + \ldots$ On en conclut que

$$k^i (k^a g^b(k) k^{cp})^{(i)} = k^{a+c} \sum_{j=0}^{b} P_j^{(i)}(k) g^j(k),$$

car $k^p - k = 0$ pour $k \in \mathbb{Z}_p$.

Pour $g(k) = t$, on définit

$$P_{a,b,c,i,t}(k) := k^{a+c} \sum_{j=0}^{b} P_j^{(i)}(k) t^j.$$

On pose aussi

$$\Gamma(x, y, z) = \sum_{a,b,c} \lambda_{a,b,c} x^a y^b z^c,$$

ce qui permet d'écrire

$$P_i(k) := \Theta^{(i)}(k) = \sum_{a,b,c} \lambda_{a,b,c} P_{a,b,c,i,t}(k),$$

où $deg\ P_i(x) < A + C + i$. Donc, si $D(A + C + (D-1)/2)|S| < ABC$, il existe un choix de $\lambda_{a,b,c}$ non tous nuls tels que $P_i(k) = 0$ pour $n < D$ et $k \in g^{-1}(S)$. Il faut maintenant vérifier que $\Theta(x)$ n'est pas identiquement nul. Pour ça, on écrit

$$\Gamma(x, y, z) = \sum_{c=0}^{C} \Gamma_c(x, y) z^c$$

et on suppose que c_0 est la plus petite valeur de c pour laquelle $\Gamma_c(x, y)$ n'est pas identiquement nul. On a alors

$$\Gamma(x, y, z) = z^{c_0} \sum_{c=c_0}^{C} \Gamma_c(x, y) z^{c-c_0}$$

et il suffit de montrer que $\Gamma_{c_0}(x, g(x)) \neq 0$ modulo x^p. On a que

$$\Gamma_{c_0}(x, g(x)) = \sum_{j=0}^{W} Q_j(x) g(x)^j$$

avec $deg\ Q_j(x) \leq A$ et $W < B$. Selon le lemme 1.7, x^{AB+B+1} ne divise pas $\Gamma_{c_0}(x, g(x))$.

En résumé, il s'agit d'avoir

$$AB + B + 1 < p \quad \text{et} \quad D(A + C + (D-1)/2)|S| < ABC,$$

pour que

$$MD < A + p(B + C).$$

On choisit alors

$$A := [0.27D], \quad B := [2.35(D|S|)^{1/2}], \quad C := [0.85B] \quad \text{et} \quad D := [1.35p^{2/3}/|S|^{1/3}],$$

lorsque $p > 100\ 000$. On peut vérifier les autres nombres directement avec l'ordinateur. Dans tous les cas, on obtient

$$M < 3.745(p|S|)^{2/3}.$$

On est maintenant en mesure de terminer la preuve. On commence par ordonner les valeurs comme suit

$$r_{g(\mathbb{Z}_p)-\{0\}}(n_1) \geq r_{g(\mathbb{Z}_p)-\{0\}}(n_2) \geq \ldots \geq r_{g(\mathbb{Z}_p)-\{0\}}(n_R).$$

On prend ces R valeurs pour former S et on a alors

$$R r_{g(\mathbb{Z}_p)-\{0\}}(n_R) \leq M < 3.745(pR)^{2/3},$$

d'où pour $r \leq [p^{1/2}/3.745^{3/2}] =: R$, nous avons

$$r_{g(\mathbb{Z}_p)-\{0\}}(n_r) < \frac{3.745p^{2/3}}{r^{1/3}}.$$

Sinon, on prend $r_{g(\mathbb{Z}_p)-\{0\}}(n_R) < p/R$. On a donc

$$\sum_{r=1}^{R} r_{g(\mathbb{Z}_p)-\{0\}}(n_r)^2 < \sum_{r=1}^{R} \frac{3.745^2 p^{4/3}}{R^{2/3}} = 3(3.745)^2 p^{4/3} R^{1/3}$$

et

$$\sum_{r=R+1}^{p} r_{g(\mathbb{Z}_p)-\{0\}}(n_r)^2 < p^2/R.$$

On en conclut que

$$r_{g(\mathbb{Z}_p) \circ g(\mathbb{Z}_p)}(m) < 29p^{3/2}$$

pour tout m. □

Remarques. On déduit de ce dernier résultat que la fonction exponentielle, telle que définie, prend au moins $(1/14)p^{1/2}$ valeurs distinctes de $\mathbb{Z}_p$. Comme cette fonction ne semble pas satisfaire une identité fonctionnelle avec seulement elle-même, un argument heuristique simple nous laisse penser qu'elle prend $(1 - 1/e + o(1))p$ valeurs de $\mathbb{Z}_p$. En effet, si on considère $g(x)$ comme p tirages dans $\mathbb{Z}_p$, indépendants entre eux, avec remise et à probabilité uniforme de $1/p$, alors la probabilité que la valeur n ne soit pas prise est de $(1 - 1/p)^p = 1/e + o(1)$, d'où le résultat. Comme on le verra, c'est un peu différent pour la fonction $f(x)$ qui ressemble à $-ln(1 - x)$. Contrairement à ce à quoi on pourrait s'attendre, il ne semble pas y avoir une relation fonctionnelle entre $f(x)$ et $g(x)$, mais il y a des relations remarquables pour la dérivée de la composée des deux. Il y a aussi d'autres relations de $g(x)$ avec elle-même qui nous rappellent la vraie fonction exponentielle. On peut facilement démontrer qu'il n'y a pas de morphisme autre que la fonction identiquement 1 ou 0 qui transforme l'addition en produit dans un corps fini, i.e. si $h(x + y) = h(x)h(y)$ pour tout x et y dans $\mathbb{F}_q$ avec $q = p^a$, alors $h \equiv 1$ ou $h \equiv 0$. Pour le voir, on remarque que $h(0) = 0$ ou 1 et dans ce dernier cas, il s'agit d'appliquer a fois le Frobenius σ à $h(x)$; on a alors

$$\sigma^a(h(x)) = h^{p^a}(x) = h(p^a x) = h(0)$$

et aussi $h^{p^a}(x) = h(x)$, d'où $h(x) = h(0)$. Il est possible, en utilisant exactement les mêmes arguments, de démontrer le même genre de résultats que le lemme 1.8 pour une large classe de fonctions de type "exponentielles tronquées". Comme

$$(p - i)!(i - 1)! \equiv (-1)^i \ (mod\ p),$$

pour $1 \leq i \leq p$, on déduit que la fonction

$$\kappa_p(x) := \sum_{n=0}^{p-1} n! x^n$$

satisfait l'identité

$$g(x) = \begin{cases} -\kappa(-x^{-1}) & \text{si } x \not\equiv 0 \ (mod\ p), \\ 1 & \text{sinon.} \end{cases}$$

Ce polynôme, qui est directement relié à la conjecture originale de Kurepa, jouit donc des mêmes propriétés que celles dont $g(x)$ hérite grâce au lemme 1.8 et aux remarques précédentes. Il en sera plus longuement question dans le prochain chapitre.

L'étude de la fonction $f(x)$ fournit un autre type d'avantages puisqu'elle est reliée de très près à une somme exponentielle célèbre, i.e. la somme de Heilbronn. Elle est définie comme suit :

$$H(a) := \sum_{n=1}^{p-1} e\left(\frac{an^p}{p^2}\right).$$

En fait, nous allons présenter ici un résultat sur une somme plus générale. Pour $l|p - 1$, on pose

$$H_l(a) := \sum_{n=1}^{p-1} e\left(\frac{an^{lp}}{p^2}\right).$$

Lemme 1.9. *L'estimé suivant est valide pour $a \not\equiv 0$ modulo p :*

$$\sum_{i=0}^{p-1} |H_l(a+ip)|^4 \leq l^3(p^2-1)((p-1)+r_{f(\mathbb{Z}_p)-f(\mathbb{Z}_p)}(0)).$$

Preuve. On remarque d'abord que pour $am \not\equiv 0$ modulo p,

$$H_l(a) = H_l(am^{lp}).$$

Ainsi on peut écrire

$$(p-1)\sum_{i=0}^{p-1} |H_l(a+ip)|^4 = \sum_{m=1}^{p-1}\sum_{i=0}^{p-1} |H_l((a+ip)m^{lp})|^4 \leq l \sum_{\substack{n=1 \\ n\not\equiv 0 \bmod p}}^{p^2-1} |H_l(n)|^4.$$

On a donc

$$\frac{p-1}{l}\sum_{i=0}^{p-1} |H_l(a+ip)|^4 \leq \sum_{m_1,m_2,m_3,m_4=1}^{p-1} \left(\sum_{n=1}^{p^2-1} e\left(\frac{(m_1^{lp}+m_2^{lp}-m_3^{lp}-m_4^{lp})n}{p^2} \right) \right)$$

$$\leq (p^2-1)E_{p^2}((\mathbb{Z}_p^*)^{lp}).$$

Un élément de $E_{p^2}((\mathbb{Z}_p^*)^{lp})$ satisfait $m_1^l+m_2^l \equiv m_3^l+m_4^l$ modulo p avec $m_1, m_2, m_3, m_4 \not\equiv 0$ modulo p. On peut alors écrire $m_1^l - m_3^l \equiv m_4^l - m_2^l \equiv t$ modulo p. Le cas $t \equiv 0$ modulo p est satisfait par $l^2(p-1)^2$ couples ; sinon on pose $m_1 \equiv v_1 t$ et $m_4 \equiv v_2 t$ modulo p. On a donc

$$m_1^{lp} - m_3^{lp} \equiv (v_1^{lp}t^{(l-1)p} - (v_1^l t^{l-1}-1)^p)t^p \equiv m_4^{lp} - m_2^{lp} \equiv (v_2^{lp}t^{(l-1)p} - (v_2^l t^{l-1}-1)^p)t^p \ (mod\ p^2).$$

Pour une variable x, nous avons la relation

$$x^p - \sum_{k=0}^{p} \binom{p}{k} (-1)^{p-k}x^k \equiv 1 - pf(x) \ (mod\ p^2),$$

qui provient du fait que $\binom{p}{k} \equiv (-1)^{k-1}p/k$ modulo p^2 pour $1 \leq k \leq p-1$. On déduit ainsi que

$$f(v_1^l t^{l-1}) \equiv f(v_2^l t^{l-1}) \ (mod\ p).$$

Il y a alors $p-1$ choix possibles pour t et pour chacun d'eux il y a au plus $l^2 r_{f(\mathbb{Z}_p)-f(\mathbb{Z}_p)}(0)$ choix possibles pour (v_1, v_2). En résumé, on a

$$\frac{p-1}{l}\sum_{i=0}^{p} |H(a+ip)|^4 \leq (p^2-1)(l^2(p-1)^2 + (p-1)l^2 r_{f(\mathbb{Z}_p)-f(\mathbb{Z}_p)}(0)),$$

d'où le résultat.□

Encore une fois, exactement le même genre d'analyse qui nous à permis de prouver le lemme 1.8 nous fournit le résultat suivant.

Corollaire 1.3. *L'estimé suivant est valide pour tout $a \not\equiv 0$ modulo p :*

$$\sum_{i=0}^{p-1} |H_l(a+ip)|^4 < 29l^3p^{7/2} + l^3p^3;$$

en particulier on a

$$|H_l(a)| < 2.33l^{3/4}p^{7/8}.$$

Remarques. Le dernier résultat nous dit en particulier que la suite $(n^p - n)/p$ est répartie uniformément modulo p. Plus précisément, dans un intervalle $(\alpha p, \beta p]$, il y a

$$(\beta - \alpha)p\left(1 + O\left(\frac{\ln(p)}{|\beta - \alpha|p^{1/8}}\right)\right)$$

telles valeurs d'après les remarques qui suivent le théorème de Weyl. Le corollaire 1.3 généralise le résultat de [22] et améliore celui de [35], l'inégalité étant non trivial pour $l = o(p^{1/6})$. De plus, encore une fois, les mêmes arguments nous fournissent une preuve que la fonction $f(x)$ prend au moins $(1/14)p^{1/2}$ valeurs distinctes de $\mathbb{Z}_p$. Cependant, on vérifie facilement avec l'identité $pf(x) = 1 + (x-1)^p - x^p$ que $f(x) = f(1-x)$, d'où on déduit que le nombre de valeurs distinctes ne peut pas dépasser $(p+1)/2$. En fait, cette fois, il faut s'attendre à ce qu'il y ait $(1 - 1/\sqrt{e} + o(1))p$ valeurs distinctes prises par cette fonction dans $\mathbb{Z}_p$. L'argument est exactement le même, sauf qu'on fait $(p+1)/2$ tirages.

Il existe une méthode standard pour borner une somme exponentielle incomplète à l'aide de la somme complète correspondante. En effet, on a par exemple que

$$\sum_{n=M}^{M+N} e\left(\frac{an^p}{p^2}\right) = \frac{1}{p}\sum_{r,s=0}^{p-1} e\left(\frac{an^p}{p^2}\right)\sum_{n=M}^{M+N} e\left(\frac{r(s-n)}{p}\right),$$

ce qui entraîne facilement que

$$\sum_{n=M}^{M+N} e\left(\frac{an^p}{p^2}\right) \ll p^{5/8}N^{1/4}.$$

Le quotient de Fermat $q(x)$ est défini comme suit :

$$q(x) := \frac{x^{p-1} - 1}{p} \quad \text{pour} \quad x \not\equiv 0 \ (mod\ p).$$

On remarque que la fonction $f(x)$ nous fournit une représentation du quotient de Fermat $q(2)$ de plusieurs façons différentes si on utilise les nombres harmoniques H_{p-1} et $H_{(p-1)/2}$ en plus. Ce genre de phénomènes se généralise probablement aux fonctions polylogues

d'ordre plus élevé que trois, mais les identités deviennent vite compliquées (voir [19] et [6]). Rappelons que la k-ième fonction polylogue modulo p est définie par

$$f_{p,k}(x) := \sum_{n=1}^{p-1} \frac{x^n}{n^k}.$$

Par exemple, en dérivant $f^2(x)$, il nous vient l'identité

$$f^2(x) = 2f_2(1-x)(x^p-1) - 2x^p f_2(x) \ (mod\ p).$$

La fonction dilogue satisfait aussi l'importante identité

$$f_2(x) = f_2(1-x) + x^p f_2(1-1/x) \ (mod\ p),$$

qui se démontre facilement en dérivant le membre de droite. Comme $x^p f_2(x^{-1}) = f_2(x)$, on obtient la relation $q(2)^2 = -f_2(2)$.

2 Calcul symbolique et théorème de Kurepa.

Après avoir remarqué que certaines suites de polynômes classiques satisfont des identités familières, nos ancêtres ont eu l'idée de faire du *calcul symbolique* pour découvrir de nouvelles formules. Étant incapables de démontrer formellement leurs résulats par cette méthode, ils utilisaient d'autres façons pour arriver à leurs fins. C'est vers les année 1970 que cette théorie fut complétée par Gian-Carlo Rota. En gros, il s'agissait simplement de bien écrire les choses et de définir les bons opérateurs. Ensuite, le reste n'est que de la simple algèbre linéaire. Souvent, on démontre des choses presque triviales pour une base qui réagit bien à un opérateur, après on écrit l'expression "par linéarité nous avons" et on obtient une identité comme par magie. Telle est la force de cette théorie. Ce qui ajoute encore une dimension de richesse, c'est que la linéarité des opérateurs laisse cohérente l'analyse dans les anneaux $\mathbb{Z}_N$ dans le sens direct. Pour les opérations inverses on ne peut pas toujours conclure.

Cette théorie porte aussi le nom de *calcul ombral*. Elle généralise une foule de résultats très élémentaires que nous voyons normalement avant même d'entrer à l'université. Après avoir bien défini le sujet, nous allons donner quelques exemples pour aboutir à la preuve de la conjecture originale de Kurepa.

Théorème (Kurepa, Barsky, Benzaghou). *Pour tout n plus grand que* 2,

$$n \not| \kappa_n(1) =: \kappa_n,$$

i.e.

$$0! + 1! + \ldots + (n-1)! \not\equiv 0 \ (mod\ n).$$

Comme on le verra, la preuve de ce théorème est loin d'être triviale et elle fait intervenir tout un arsenal d'idées qui vont, sans aucun doute, intéresser un lecteur profane. Il n'est pas très difficile de voir qu'il est suffisant de prouver le théorème dans le cas où $n = p$, un nombre premier. C'est un petit miracle que κ_n ne s'annule jamais modulo n pour $n > 2$. Une analyse naïve de ce problème nous dirait que cette fonction doit s'annuler environ $\ln\ln(N)$ fois dans l'ensemble des $p \leq N$, mais en fait il existe un raison pour que la réponse soit 1 pour $N \geq 2$ et c'est le but ultime de ce chapitre que de le prouver. Dans la littérature, la notation $!n$ est utilisée pour représenter les valeurs de κ_n et elle est prononcée "left factorial". Il est donc très facile de constater que le dernier théorème est en fait équivalent au fait que $(n!, !n) = 2$ pour tout $n \geq 2$. Rappelons que nous avons démontré, dans la remarque qui suit le lemme 1.8, que le polynôme $\kappa_p(x)$ prend toujours au moins $(1/14)p^{1/2}$ des valeurs de $\mathbb{Z}_p$.

2.1 Quelques définitions préliminaires.

Soit $\mathbb{A}$ un anneau unitaire commutatif. En pratique, $\mathbb{A}$ est souvent $\mathbb{Z}$, $\mathbb{Z}[z]$ ou l'ensemble des nombres p-adiques. On appele *delta-opérateur* un opérateur linéaire δ :

$\mathbb{A}[x] \to \mathbb{A}[x]$ qui satisfait $\delta(x) \neq 0$ et qui commute avec les opérateurs de translation

$$\tau_a : \mathbb{A}[x] \to \mathbb{A}[x],\ P(x) \mapsto P(x+a)\ (a \in \mathbb{A}),$$

où $P(x) \in \mathbb{A}[x]$ ici et par la suite. On vérifie que ces conditions entraînent que $\delta(a) = 0$ pour tout $a \in \mathbb{A}$ et de même, par induction, que $deg\ \delta P(x) = deg\ P(x) - 1$. On dit qu'une famille $(P_n(x))_{n \in N}$ de polynômes est *associée* à un delta-opérateur si

- $P_0(x) = 1$ et $P_n(0) = 0\ (n \geq 1)$: Propriété de normalisation,
- $\delta P_n(x) = nP_{n-1}(x)\ (n \in N)$: Propriété symbolique (dite de Scheffer).

On en déduit que $(P_n(x))_{n \in \mathbb{Z}}$ est en fait une famille de polynômes unitaires qui forment une base de $\mathbb{A}[x]$.

L'exemple le plus familier est sans aucun doute celui de la base canonique x^n, avec $n \geq 0$, qui est naturellement associée à l'opérateur de dérivation ∂. Un premier exemple moins évident est celui de la base dite de Pochhammer, à savoir

$$(x)_n := \prod_{i=0}^{n-1}(x-i) \quad \text{pour} \quad n \geq 0,$$

qui est associée avec l'opérateur de différence finie défini par

$$\Delta : \mathbb{A}[x] \to \mathbb{A}[x],\ P(x) \mapsto P(x+1) - P(x).$$

En effet, on vérifie facilement que $\Delta(x)_n = n(x)_{n-1}$.

Il est alors très pertinent de définir l'opérateur de changement de base naturel entre la base canonique et la base de Pochhammer, i.e.

$$\phi : \mathbb{A}[x] \to \mathbb{A}[x],\ (x)_n \mapsto x^n.$$

Cet opérateur est très riche en propriétés. Premièrement, c'est évidement un opérateur linéraire, ce qui implique en particulier qu'il rend cohérent l'analyse dans les anneaux $\mathbb{Z}_N$. Ceci signifie que si

$$f(x) \equiv g(x)\ (mod\ N),$$

alors

$$\phi(f(x)) \equiv \phi(g(x))\ (mod\ N),$$

où $f(x)$ et $g(x)$ sont dans $\mathbb{Z}[x]$. Une propriété moins évidente, de genre fonctionnelle, nous vient de la structure multiplicative des bases impliquées. En effet, on a successivement pour tout m et n dans $\mathbb{N}$,

$$x^{n+m} = x^n\phi((x)_m) = \phi((x)_{n+m}) = \phi((x)_n(x-n)_m),$$

d'où on déduit par linéarité que

$$x^n\phi(P(x)) = \phi((x)_nP(x-n))$$

pour tout polynôme $P(x)$ de $\mathbb{A}[x]$.

Il est utile de remarquer de quelle façon l'opérateur ϕ relie l'action des delta-opérateurs associés aux bases sur lesquelles ϕ est naturellement défini. On remarque simplement que $\partial\phi((x)_n) = nx^{n-1} = \phi\Delta((x)_n)$, d'où on déduit par linéarité que $\partial\phi = \phi\Delta$ et par induction que $\partial^n\phi = \phi\Delta^n$ pour $n \geq 0$.

Il existe aussi un lien direct entre les opérateur ∂ et Δ. En effet, on a

$$e^{\partial}x^n = \sum_{i=0}^{\infty}\frac{\partial^i x^n}{i!} = \sum_{i=0}^{\infty}\frac{(n)_i x^{n-i}}{i!} = (x+1)^n = (1+\Delta)x^n,$$

ce qui montre, par linéarité, que $e^{\partial} = 1 + \Delta$. En fait, il n'est pas étonnant qu'un tel lien existe, car on peut facilement montrer que tous les opérateurs $\mathcal{B}$ qui commutent avec les translations peuvent se développer en fonction d'un delta-opérateur δ qui a $P_n(x)$ comme base associée. En effet, l'expression

$$\mathcal{B} := \sum_{n=0}^{\infty}\frac{\alpha_n\delta^n}{n!} \quad \text{où} \quad \alpha_n := \mathcal{B}P_n(0) \in \mathbb{A}$$

définit un tel développement. De plus, il est facile de voir que si $\mathcal{B}$ est un delta-opérateur, alors $\alpha_0 = 0$ et $\alpha_1 \neq 0$.

Il est possible de définir une sorte de dérivation, dite de Pincherle, sur l'ensemble des opérateurs. En effet, en définisant l'opérateur χ de multiplication par x

$$\chi : \mathbb{A}[x] \to \mathbb{A}[x],\ P(x) \mapsto xP(x),$$

on peut considérer le commutateur

$$\mathcal{T}' := [\mathcal{T}, \chi] := \mathcal{T}\chi - \chi\mathcal{T}.$$

Alors on montre facilement que $(\mathcal{T}\mathcal{V})' = \mathcal{T}'\mathcal{V} + \mathcal{T}\mathcal{V}'$, où $\mathcal{T}$ et $\mathcal{V}$ sont des opérateurs. Par exemple, pour $F(x) \in \mathbb{A}[x]$, on montre aussi que

$$[F(\partial), \chi] = F'(\partial) \text{ et que } [\partial, F(\chi)] = F'(\chi).$$

Un cas particulièrement intéressant est $[\partial, \chi] = 1$ qui définit formellement ce qu'on appelle l'algèbre de Weyl $W := \mathbb{A}[[\partial, \chi]]$.

Comme les $P_n(x)$ avec $n \in \mathbb{N}$ forment une base de $\mathbb{A}[x]$, il s'ensuit qu'il existe un développement pour toutes les fonctions $f(x) \in \mathbb{A}[x]$ de la forme

$$f(x+a) = \sum_{n=0}^{\infty}\frac{\delta^n f(a)}{n!}P_n(x),$$

qui est un généralisation du développement de Taylor.

Définition. Les polynômes de Bell sont définis implicitement par

$$B_n(x) := \phi(x^n),$$

pour $n \geq 0$.

L'équation fonctionnelle de l'opérateur ϕ nous permet alors d'écrire

$$x\phi((x+1)^n) = \phi(x^{n+1}),$$

d'où la relation de récurence

$$B_{n+1}(x) = x \sum_{i=0}^{n} \binom{n}{i} B_i(x)$$

avec $B_0(x) = 1$. Les nombres de Bell sont alors définis par $B_n := B_n(1)$ et ainsi on a

$$B_{n+1} = \sum_{i=0}^{n} \binom{n}{i} B_i.$$

On remarque aussi que

$$B_n^{(1)}(x) = \phi((x+1)^n - x^n) = \sum_{i=0}^{n-1} \binom{n}{i} B_i(x),$$

d'où on déduit que

$$B_{n+1}(x) = x(B_n(x) + B_n^{(1)}(x)),$$

ce qui fournit une façon plus économique pour le calcul explicite de ces polynômes. On peut alors écrire

$$B_n(x) = (\chi(I + \partial))^n(1),$$

où I est l'opérateur identité. Ces considérations nous permettent d'obtenir une caractérisation combinatoire des nombres de Bell. Pour ce faire, on écrit implicitement

$$B_n(x) =: \sum_{i=0}^{n} \begin{bmatrix} n \\ i \end{bmatrix} x^i,$$

d'où on déduit que

$$\begin{bmatrix} n+1 \\ i \end{bmatrix} = \begin{bmatrix} n \\ i-1 \end{bmatrix} + i \begin{bmatrix} n \\ i \end{bmatrix},$$

avec

$$\begin{bmatrix} n+1 \\ n+1 \end{bmatrix} = \begin{bmatrix} n+1 \\ 1 \end{bmatrix} = 1.$$

On a ainsi montré que les nombres de Stirling de deuxième espèce sont les coefficients des polynômes de Bell. Il n'est pas trop difficile de voir que $\begin{bmatrix} n \\ i \end{bmatrix}$ est le nombre de façons de partitionner un ensemble de n éléments considérés comme distincts en i sous-ensembles non vides. En effet, il y a deux façons de fabriquer une telle partition. La première est d'ajouter à une partition de $n-1$ éléments en $i-1$ sous-ensembles un élément dans un nouvel ensemble et la deuxième est d'ajouter cet élément à une partition de $n-1$ éléments en i sous-ensembles, et ce de i façons possibles. Ce qui se traduit exactement

par la relation de récurence pour les nombres de Stirling. Comme le n-ième nombre de Bell est défini comme étant la somme des nombres de Stirling d'ordre n, il s'ensuit que les nombres de Bell représentent le nombre de façons de partitionner un ensemble de n éléments considérés comme distincts en sous-ensembles non vides. Historiquement, c'est cette définition qui fut utilisée pour les définir.

La suite des polynômes de Bell est associée à l'opérateur

$$\nabla : \mathbb{A}[x] \to \mathbb{A}[x],\ P(x) \mapsto \ln(1+\partial)P(x).$$

En effet, on a successivement

$$\nabla B_n(x) = \ln(1+\partial)\phi(x^n) = \phi\ln(1+\Delta)(x^n) = \phi\partial x^n = nB_{n-1}(x).$$

On peut donc former le graphique commutatif

$$\begin{array}{ccccc}
(x)_{n+1} & \stackrel{\phi}{\longrightarrow} & x^{n+1} & \stackrel{\phi}{\longrightarrow} & B_{n+1}(x) \\
\uparrow \Gamma & & \uparrow \chi & & \uparrow \Omega \\
(x)_n & \stackrel{\phi}{\longrightarrow} & x^n & \stackrel{\phi}{\longrightarrow} & B_n(x) \\
\downarrow \Delta & & \downarrow \partial & & \downarrow \nabla \\
n(x)_{n-1} & \stackrel{\phi}{\longrightarrow} & nx^{n-1} & \stackrel{\phi}{\longrightarrow} & nB_{n-1}(x)
\end{array}$$

pour clarifier la structure de ce qui a été rencontré jusqu'ici. On a posé $\Gamma := \chi\tau_{-1}$ et $\Omega = \chi(\partial + I)$ pour alléger le tableau. Ce graphique peut être prolongé indéfiniment dans toutes les directions de l'espace $\mathbb{Z}^3$, cependant vers le bas il finit par aboutir à 0 partout. Le seul ingrédient qui nous manque encore est l'opérateur de convolution exponentielle qui nous permet de voyager dans le sens de la profondeur. Si on fixe $a \in \mathbb{A}$ et une suite de polynômes $P_n(x)$ avec $n \in \mathbb{N}$ et $deg\ P_n(x) = n$, on peut alors définir

$$T_a : \mathbb{A}[x] \to \mathbb{A}[x],\ P_n(x) \mapsto \sum_{i=0}^{n} \binom{n}{i} a^{n-i} P_i(x).$$

Il est clair que $T_0 = I$ et on vérifie facilement que $T_a T_b = T_{a+b}$, ce qui implique en particulier que $T_a T_{-a} = I$.

Une autre propriété remarquable de l'opérateur ϕ est

$$\phi((x)_n) = x^n = e^{-x}\sum_{i=0}^{\infty}\frac{x^{n+i}}{i!} = e^{-x}\sum_{i=n}^{\infty}\frac{x^i}{(i-n)!} = e^{-x}\sum_{i=0}^{\infty}\frac{(i)_n x^i}{i!},$$

d'où on déduit par linéarité que

$$\phi(f(x)) = e^{-x}\sum_{i=0}^{\infty}\frac{f(i)x^i}{i!}.$$

Un cas particulier remarquable est obtenu en posant $f(x) = x^n$, ce qui donne

$$B_n(x) = e^{-x}\sum_{i=0}^{\infty}\frac{i^n x^i}{i!},$$

d'où en particulier l'identité

$$B_n = e^{-1} \sum_{i=0}^{\infty} \frac{i^n}{i!}$$

qui est souvent attribuée à Dobinski dans la littérature. On peut utiliser cette dernière identité pour obtenir l'estimé de de Bruijn

$$\ln(B_n) = n \left(\ln(n) - \ln\ln(n) - 1 + \frac{\ln\ln(n)}{\ln(n)} + \frac{1}{\ln(n)} + O\left(\frac{(\ln\ln(n))^2}{(\ln(n))^2} \right) \right),$$

de même que pour voir que la série génératrice exponentielle des polynômes de Bell est donnée par

$$e^{x(e^z-1)} = \sum_{n=0}^{\infty} \frac{B_n(x) z^n}{n!},$$

qui se réduit à

$$e^{e^z-1} = \sum_{n=0}^{\infty} \frac{B_n z^n}{n!}$$

dans le cas des nombres de Bell.

Une autre propriété remarquable de l'opérateur ϕ est la façon dont il agit sur le produit de deux polynômes $f(x)$ et $g(x)$ de $\mathbb{A}[x]$. En utilisant la généralisation du théorème de Taylor, on peut écrire

$$\begin{aligned} \phi((x)_n(x)_m) = \phi((x)_n)\phi((x+n)_m) &= \sum_{i \geq 0} \frac{1}{i!} (n)_i \phi((x)_n)(m)_i \phi((x)_{m-i}) \\ &= \sum_{i \geq 0} \frac{x^i}{i!} \partial^i \phi((x)_n) \partial^i \phi((x)_m) \end{aligned}$$

et on en conclut par bilinéarité que

$$\phi(f(x)g(x)) = \sum_{i \geq 0} \frac{x^i}{i!} \partial^i \phi(f(x)) \partial^i \phi(g(x)).$$

Un cas particulier intéressant est que lorsqu'on pose $f(x) = x^n$ et $g(x) = x^m$, on obtient

$$B_{n+m}(x) = \sum_{i \geq 0} \frac{x^i}{i!} B_n^{(i)}(x) B_m^{(i)}(x).$$

Cette dernière identité nous permet de factoriser la matrice de Hankel d'ordre N (i.e. $(P_{i+j}(x))_{0 \leq i,j \leq N}$) associée aux polynômes de Bell en deux matrices triangulaires et ainsi de montrer que son déterminant vaut exactement

$$x^{\frac{N(N+1)}{2}} \prod_{n=0}^{N} n!;$$

cette formule pour le déterminant, en plus d'être très jolie, nous permet de montrer assez facilement que la série génératrice ordinaire des nombres de Bell (voir le lemme 2.5) est irrationnelle.

2.2 Lemmes utiles.

Lemme 2.1. *Supposons que* $f(x) \equiv g(x)$ *modulo* $p^v\mathbb{Z}[x]$ *pour un certain entier* $v \geq 1$. *Alors*

$$f^p(x) \equiv g^p(x) \ (mod\ p^{v+1}\mathbb{Z}[x]).$$

Preuve. Par hypothèse, on a $f(x) = g(x) + p^v h(x)$ avec $h(x) \in \mathbb{Z}[x]$. Ainsi

$$f^p(x) = (g(x) + p^v h(x))^p = g^p(x) + p^{v+1} r(x),$$

d'où le résultat.□

Lemme 2.2. *Supposons que* p *est un nombre premier impair et que* $v \geq 1$ *est un entier. Alors la congruence suivante est valide :*

$$\prod_{n=0}^{p-1} f(x - np^v) \equiv f^p(x) \ (mod\ p^{v+1}\mathbb{Z}[x]).$$

Dans le cas où $v = 0$, *on a*

$$\prod_{n=0}^{p-1} f(x-n) \equiv \prod_{f(\beta)=0} ((x^p - x) - (\beta^p - \beta)) \ (mod\ p\mathbb{Z}[x]).$$

Preuve. En utilisant le développement de Taylor, on a

$$f(x - np^v) \equiv f(x) - np^v f^{(1)}(x) \ (mod\ p^{2v}),$$

d'où on obtient

$$\begin{aligned} \prod_{n=0}^{p-1} f(x - np^v) &\equiv f^p(x) - p^v f^{(1)}(x) \sum_{n=1}^{p-1} n \ (mod\ p^{2v}\mathbb{Z}[x]) \\ &\equiv f^p(x) - \frac{(p-1)}{2} p^{v+1} f^{(1)}(x) f^{p-1}(x) \ (mod\ p^{v+1}\mathbb{Z}[x]), \end{aligned}$$

ce qui termine la première partie. Pour la deuxième partie il est suffisant de montrer que $(x)_p \equiv x^p - x$ modulo $p\mathbb{A}[x]$, i.e. que $n^p \equiv n$ modulo p pour tout $n \in \mathbb{Z}_p$. La façon la plus simple est de le faire par inducion sur n en utilisant le fait évident que les coefficients binomiaux satisfont

$$\binom{p}{k} \equiv 0 \ (mod\ p) \quad \text{pour} \quad 1 \leq k \leq p-1.$$

□

Remarque. Le cas $p = 2$ nous fournit

$$f(x) f(x - 2^v) \equiv f^2(x) \ (mod\ 2^v).$$

Lemme 2.3. *Pour $v \geq 1$, la congruence*

$$(x)_{p^v} \equiv (x^p - x)^{p^{v-1}} \ (mod\ p^v\mathbb{Z}[x])$$

est vérifiée pour $p > 2$. Pour $p = 2$, nous avons

$$(x)_{2^v} \equiv (x^2 - x)^{2^{v-1}} \ (mod\ \max\{2, 2^{v-1}\}\mathbb{Z}[x]).$$

Preuve. La preuve de la première partie se fait par induction sur v. Pour $v = 1$, on a

$$(x)_p \equiv x^p - x \ (mod\ p\mathbb{Z}[x]);$$

ceci provient du fait que les deux polynômes ont les mêmes racines comme on l'a vu au lemme précédent. Supposons que

$$(x)_{p^v} \equiv (x^p - x)^{p^{v-1}} \ (mod\ p^v\mathbb{Z}[x])$$

pour un certain $v \geq 1$. Le lemme 2.2 entraîne alors

$$(x)_{p^{v+1}} = \prod_{n=0}^{p-1}(x - np^v)_{p^v} \equiv (x)_{p^v}^p \ (mod\ p^{v+1}\mathbb{Z}[x]) \equiv (x^p - x)^{p^v} \ (mod\ p^{v+1}\mathbb{Z}[x]),$$

ce qui démontre le résultat pour p impair ; on fait de même pour $p = 2$.□

Lemme 2.4. *Pour $v \geq 1$, nous avons la congruence*

$$\phi((x^p - x)^{p^{v-1}} f(x)) \equiv x^{p^v}\phi(f(x)) \ (mod\ p^v\mathbb{Z}[x])$$

pour p impair et

$$\phi((x^2 - x)^{2^{v-1}} f(x)) \equiv x^{2^v}\phi(f(x)) \ (mod\ \max\{2, 2^{v-1}\}\mathbb{Z}[x])$$

pour $p = 2$.

Preuve. Clairement, on a

$$x^{p^v}\phi(f(x)) \equiv \phi((x)_{p^v} f(x)) \equiv \phi((x^p - x)^{p^{v-1}} f(x)) \ (mod\ p^v\mathbb{Z}[x])$$

en utilisant le lemme 2.3. Le cas $p = 2$ se fait de la même façon.□

Proposition 2.1. *Pour $v \geq 1$ et $m \geq 0$, la congruence*

$$\phi((x^{p^v} - x)^m f(x)) \equiv \left(\sum_{i=1}^{v} x^{p^i}\right)^m \phi(f(x)) \ (mod\ p\mathbb{Z}[x])$$

est valide.

Preuve. La preuve se fait par induction sur v. Pour $v = 1$ il s'agit tout simplement du fait que

$$x^{mp}\phi(f(x)) = \phi((x)_{mp} f(x - mp)) \equiv \phi((x^p - x)^m f(x)) \ (mod\ p\mathbb{Z}[x]),$$

où on a utilisé le lemme 2.3. Supposons que le résultat est vérifié pour tout $n \leq v-1$. On considère alors

$$\begin{aligned}
\left(\sum_{i=1}^{v} x^{p^i}\right)^m \phi(f(x)) &\equiv \sum_{j=0}^{m} \binom{m}{j} \left(\sum_{i=1}^{v-1} x^{p^i}\right)^j x^{(m-j)p^v} \phi(f(x)) \ (mod\ p\mathbb{Z}[x]) \\
&\equiv \sum_{j=0}^{m} \binom{m}{j} x^{(m-j)p^v} \phi((x^{p^{v-1}} - x)^j f(x)) \ (mod\ p\mathbb{Z}[x]) \\
&\equiv \phi\left(\sum_{j=0}^{m} \binom{m}{j} (x)_{(m-j)p^v} (x^{p^{v-1}} - x)^j f(x)\right) \ (mod\ p\mathbb{Z}[x]) \\
&\equiv \phi\left(\sum_{j=0}^{m} \binom{m}{j} (x^p - x)^{(m-j)p^{v-1}} (x^{p^{v-1}} - x)^j f(x)\right) \ (mod\ p\mathbb{Z}[x]) \\
&\equiv \phi(((x^p - x)^{p^{v-1}} + (x^{p^{v-1}} - x))^m f(x)) \ (mod\ p\mathbb{Z}[x]) \\
&\equiv \phi((x^{p^v} - x)^m f(x)) \ (mod\ p\mathbb{Z}[x]),
\end{aligned}$$

où on a utilisé l'hypothèse d'induction et le lemme 2.3.□

Corolaire 2.1. *La conguence*

$$\phi(x^{mp^v} f(x)) \equiv \sum_{j=0}^{m} \binom{m}{j} \left(\sum_{i=1}^{v} x^{p^i}\right)^{m-j} \phi(x^j f(x)) \ (mod\ p\mathbb{Z}[x])$$

est valide pour $v \geq 1$, $m \geq 0$ *et pour tout* $f(x) \in \mathbb{Z}[x]$.

Preuve. Il s'agit tout simplement d'appliquer la dernière proposition au fait trivial

$$x^{mp^v} = ((x^{p^v} - x) + x)^m.$$

□

Remarques. En prenant le polynôme $f(x) = x^n$ avec $n \in \mathbb{N}$, on obtient l'importante congruence

$$B_{n+mp^v}(x) \equiv \sum_{j=0}^{m} \binom{m}{j} \left(\sum_{i=1}^{v} x^{p^i}\right)^{m-j} B_{n+j}(x) \ (mod\ p\mathbb{Z}[x]),$$

qui implique à elle seule plusieurs cas remarquables dont

$$B_{n+p}(x) \equiv x^p B_n(x) + B_{n+1}(x) \ (mod\ p\mathbb{Z}[x]),$$

$$B_{n+p^v}(x) \equiv \left(\sum_{i=1}^{v} x^{p^i}\right) B_n(x) + B_{n+1}(x) \ (mod\ p\mathbb{Z}[x])$$

et

$$B_{mp}(x) \equiv \sum_{j=0}^{m} \binom{m}{j} x^{p(m-j)} B_j(x) \ (mod\ p\mathbb{Z}[x]),$$

ce dernier se réduisant à

$$B_{mp} \equiv B_{m+1} \ (mod\ p)$$

dans le cas des nombres de Bell. La première de ces congruences est de Touchard, la deuxième de Radoux, la troisième de Junod [24] et la dernière de Comtet.

2.3 Vers la preuve de la conjecture de Kurepa.

Lemme 2.5. *La fonction génératrice ordinaire des polynômes de Bell $F(x,z)$ satisfait*

$$F(x,z) := \sum_{n\geq 0} B_n(x)z^n = \sum_{n\geq 0} \frac{(xz)^n}{\prod_{j=1}^{n}(1-jz)}.$$

De plus, on a la congruence

$$F(x,z) \equiv F_p(x,z) \ (mod\ p\mathbb{Z}[x][[z]]),$$

où

$$F_p(x,z) := \frac{\sum_{n=0}^{p-1}(xz)^n \prod_{i=n+1}^{p-1}(1-iz)}{1-z^{p-1}-(xz)^p}$$

se réduit à

$$F_p(x,z) \equiv \frac{\sum_{n=0}^{p-2} B_n(x)z^n + (B_{p-1}(x)-1)z^{p-1}}{1-z^{p-1}-(xz)^p} \ (mod\ p\mathbb{Z}[x][[z]]).$$

Preuve. Pour la première partie, on pose

$$F(z,x) := \sum_{n\geq 0} B_n(x)z^n.$$

Clairement, on a successivement

$$F(x,z) = 1+\phi\left(xz\sum_{n\geq 0}(xz)^n\right) = 1+xz\phi\left(\sum_{n\geq 0}((x+1)z)^n\right) = 1+xz\phi\left(\frac{1}{1-z-xz}\right),$$

où on a utilisé la linéarité et la propriété fonctionelle de l'opérateur ϕ qui agit seulement sur x. On poursuit en écrivant

$$\begin{aligned} F(x,z) &= 1+\frac{xz}{1-z}\phi\left(\frac{1}{1-xz/(1-z)}\right) \\ &= 1+\frac{xz}{1-z}\phi\left(\sum_{n\geq 0}\left(\frac{xz}{1-z}\right)^n\right) \\ &= 1+xh(z)\phi(F(x,h(z))), \end{aligned}$$

où $h(z) := \dfrac{z}{1-z}$ est une homographie. Les composées de $h(z)$ sont $h^k(z) = \dfrac{z}{1-kz}$, donc par induction on a le développement

$$F(x,z) = \sum_{n=0}^{m-1} x^n \prod_{k=1}^{n} h^k(z) + x^m F(x,h^m(z)) \prod_{k=1}^{m} h^k(z),$$

d'où on obtient

$$F(x,z) = \sum_{n\geq 0} \frac{(xz)^n}{\prod_{k=1}^{n}(1-kz)}.$$

Pour la deuxième partie, on écrit d'abord

$$\begin{aligned} F(x,z) &= \sum_{n=0}^{p-1}\sum_{i\geq 0} \frac{(xz)^{ip+n}}{\prod_{j=1}^{ip+n}(1-jz)} \\ &= \sum_{n=0}^{p-1}\sum_{i\geq 0} \frac{(xz)^{n}}{\prod_{j=ip+1}^{ip+n}(1-jz)} \frac{(xz)^{ip}}{\prod_{k=1}^{ip}(1-kz)}. \end{aligned}$$

Comme

$$\frac{(xz)^n}{\prod_{j=ip+1}^{ip+n}(1-jz)} \equiv \frac{(xz)^n}{\prod_{j=1}^{n}(1-jz)} \ (mod\ p\mathbb{Z}[x][[z]])$$

et

$$\frac{(xz)^{ip}}{\prod_{k=1}^{ip}(1-kz)} \equiv \left(\frac{(xz)^p}{\prod_{k=1}^{p}(1-kz)}\right)^i \ (mod\ p\mathbb{Z}[x][[z]]),$$

on déduit que

$$F(x,z) \equiv \sum_{n=0}^{p-1} \frac{(xz)^n}{\prod_{j=1}^{n}(1-jz)} \sum_{i\geq 0} \left(\frac{(xz)^p}{\prod_{k=1}^{p}(1-kz)}\right)^i \ (mod\ p\mathbb{Z}[x][[z]])$$

et le résultat s'ensuit. Pour la dernière partie, on multiplie $F_p(x,z)$ par $1-z^{p-1}-(xz)^p$ et on développe en utilisant la congruence de Touchard. On obtient alors

$$\begin{aligned} (1-z^{p-1}-(xz)^p)F_p(x,z) &\equiv \sum_{n\geq 0} B_n(x)z^n - B_n(x)z^{n+p-1} - x^pB_n(x)z^{n+p} \\ &\equiv \sum_{n=0}^{p-2} B_n(x)z^n - (B_{p-1}(x)-1)z^{p-1} \ (mod\ p\mathbb{Z}[x][[z]]), \end{aligned}$$

d'où le résultat.□

La démonstration du théorème de Kurepa est possible grâce à une étude approfondie des zéros dans $\overline{\mathbb{F}}_p$ du dénominateur $\varepsilon_p(x,z)$ de $F_p(x,z)$.

Lemme 2.6. *Soit $\varepsilon_p(x,z) := 1-z^{p-1}-(xz)^p$ avec $x \in \mathbb{F}_p^*$. Notons $\theta_i^{-1}(x)$, $1 \leq i \leq p$, les racines de $\varepsilon_p(x,z)$ dans $\overline{\mathbb{F}}_p$. Elles vérifient alors $\theta_i^p(x) = \theta_i(x) + x$; de plus, on a $\theta_i(x) \neq \theta_j(x)$ dans $\overline{\mathbb{F}}_p$ si $i \neq j$. L'extension $K_p(x) := \mathbb{F}_p(\theta(x))$, dite d'Artin-Schreier, est de degré p et son groupe de Galois, isomorphe à $\mathbb{Z}_p$ additif, est engendré par le Frobenius $\sigma_p : a \mapsto a^p$.*

Preuve. Essentiellement, il faut juste montrer que $\varepsilon_p(x,z)$ est irréductible sur $\mathbb{F}_p$ et possède que des racines simples. Pour la deuxième partie, il suffit de remarquer que 0 n'est pas une racine et que $\varepsilon_p^{(1)}(x,z) = z^{p-2}$ sur $\mathbb{F}_p$. Pour la première partie, on remarque

que si $\theta(x)$ est une racine de $z^p - z - x = z^p\varepsilon_p(x, z^{-1})$, alors il en est de même pour $\theta(x)+i$, car $(\theta(x)+i)^p-(\theta(x)+i)-x=0$. On en déduit alors que les racines de $\varepsilon_p(x,z)$ sont de la forme

$$\frac{1}{\theta(x)+i} \quad \text{pour} \quad 0 \leq i \leq p-1.$$

On voit alors que si $\varepsilon_p(x,z)$ était réductible, alors le degré du plus petit facteur diviserait p. Comme $x \not\equiv 0$ modulo p, il s'ensuit que $\varepsilon_p(x,z)$ n'a pas de racine dans $\mathbb{F}_p$ et donc le lemme est démontré.□

Notation. Les deux entiers suivants vont jouer un rôle particulier par la suite :

$$t_p := \sum_{n=0}^{p-1} p^n = \frac{p^p-1}{p-1}$$

et

$$c_p := \sum_{n=1}^{p-1} np^{n-1} = \frac{p^p - t_p}{p-1}.$$

Dans tout ce qui suit, on note

$$Tr(\alpha) := Tr_{K_p(x)/\mathbb{F}_p}(\alpha) = \sum_{i=0}^{p-1} \sigma^i(\alpha)$$

pour tout $\alpha \in K_p(x)$ qui est bien sûr isomorphe à $\mathbb{F}_{p^p}$ pour tout $x \neq 0$ de $\mathbb{F}_p$.

Lemme 2.7. *Soit $\theta(x)$ une racine de $z^p - z - x$ dans $K_p(x)$, où $x \in \mathbb{F}_p^*$. On a alors*

(1) $\theta^{p^s}(x) = \sigma_p^s(\theta(x)) = \theta(x) + sx,$

(2) $Tr(\theta^i(x)) = \begin{cases} 0 & si\ 0 \leq i \leq p-2, \\ -1 & si\ i = p-1, \end{cases}$

(3) $Tr(\theta^{-1}(x)) = -x^{-1},$

(4) $\theta^{t_p}(x) = x,$

(5) $(n\theta^{c_p}(x))^{p-1} = x^{-1}\theta(x) \quad pour \quad 1 \leq n \leq p-1,$

(6) $\sigma_p^s(\theta^{c_p}(x)) = x^{-s}\theta^{c_p}(x)\prod_{i=1}^{s}(\theta(x)+(i-1)x).$

Preuve. Le point (1) se fait par induction sur s en utilisant le polynôme minimal. Le point (2) se fait en substituant l'identité du point (1) dans la trace de $\theta^i(x)$ et en observant que

$$\sum_{j=0}^{p-1} j^a = \begin{cases} -1 & \text{si } a \equiv 0\ (mod\ p-1), \\ 0 & \text{sinon.} \end{cases}$$

Le point (3) découle du fait que $\theta^{-1}(x)$ satisfait le polynôme $\varepsilon_p(x,z)$. De même, le point (4) vient du fait que $\theta^{t_p}(x)$ n'est rien d'autre que la norme de $\theta(x)$. Pour ce qui est du

point (5), on utilise (1) et (4). On démontre ainsi que $n\theta^{c_p}(x)$, avec $1 \leq n \leq p-1$, sont les racines $(p-1)$-ièmes de $x^{-1}\theta(x)$ dans $K_p(x)$. Pour finir, le point (6) se fait par induction sur s en utilisant (1) et (4).□

Lemme 2.8. *Les zéros de $\varepsilon_p(x,z)$, avec $x \in \mathbb{F}_p^*$, forment une $\mathbb{F}_p$-base normale de $K_p(x)$. De plus, pour un élément $\omega \in K_p(x)$ de la forme*

$$\omega := \sum_{n=0}^{p-1} \frac{v_n}{\theta(x)+n},$$

où v_n dans $\mathbb{F}_p$, avec $0 \leq n \leq p-1$, nous avons

$$v_n = x^2 Tr\left(\frac{\omega}{\theta(x)+n}\right) \quad \text{pour} \quad 0 \leq n \leq p-1$$

et

$$Tr(\omega) = -x^{-1}\sum_{n=0}^{p-1} v_n.$$

Preuve. Commençons pas montrer que les $\dfrac{1}{\theta(x)+n}$, avec $0 \leq n \leq p-1$, forment une base de $K_p(x)$. En fait, il suffit de remarquer que

$$\frac{x}{\theta(x)+n} = (\theta(x)+n)^{p-1} - 1 = \prod_{i=1}^{p-1}(\theta(x)+n+i)$$

ce qui permet de voir facilement qu'il n'y a pas de représentation non triviale de 0 dans $K_p(x)$. Comme les $\dfrac{1}{\theta(x)+n}$, avec $0 \leq n \leq p-1$ sont les zéros de $\varepsilon_p(x,z)$, il s'ensuit qu'ils forment une $\mathbb{F}_p$-base normale de $K_p(x)$. On remarque alors que

$$Tr\left(\frac{1}{\theta(x)+n}\right) = -x^{-1} \quad \text{pour} \quad 0 \leq n \leq p-1,$$

de sorte que

$$\left(Tr\left(\frac{1}{\theta(x)+n}\right)\right)^2 = Tr\left(\frac{1}{(\theta(x)+n)^2}\right) + 2\sum_{0\leq i<j\leq p-1}\frac{1}{(\theta(x)+i)(\theta(x)+j)} = x^{-2}.$$

Pour $p \geq 3$, comme le coefficient de z^{p-2} dans $\varepsilon_p(x,z)$ est 0, il s'ensuit que

$$Tr\left(\frac{1}{(\theta(x)+n)^2}\right) = x^{-2} \quad \text{pour} \quad 0 \leq n \leq p-1.$$

Le même résultat est trivialement valide pour $p = 2$. De plus, on a

$$Tr\left(\frac{1}{(\theta(x)+i)(\theta(x)+j)}\right) =$$

$$\begin{cases} \frac{1}{j-i}\left(Tr\left(\frac{1}{(\theta(x)+i)}\right) - Tr\left(\frac{1}{(\theta(x)+j)}\right)\right) & \text{si } 0 \leq i < j \leq p-1, \\ Tr\left(\frac{1}{(\theta(x)+i)^2}\right) & \text{si } 0 \leq i = j \leq p-1. \end{cases}$$

On a donc montré que

$$Tr\left(\frac{1}{(\theta(x)+i)(\theta(x)+j)}\right) = \begin{cases} x^{-2} & \text{si } 0 \leq i = j \leq p-1, \\ 0 & \text{si } 0 \leq i < j \leq p-1, \end{cases}$$

ce qui termine la démonstration du lemme.□

Pour ce qui suit, on utilise la notation $\theta := \theta(1)$, de même que $F_p(z) := F_p(1, z)$.

Lemme 2.9. *Soit $\theta(x)$ une racine de $z^p - z - x$ dans $K_p(x)$,où $x \in \mathbb{F}_p^*$. Alors, il existe un développement pour la fonction $F_p(x, z)$ de la forme*

$$F_p(x, z) = \sum_{\theta^p(x)=\theta(x)+x} \frac{\mu_{\theta(x)}}{1 - \theta(x)z}.$$

On a ainsi

$$\mu_{\theta(x)} = -\sum_{n=0}^{p-1} x^{-n} \prod_{i=1}^{n} (\theta(x) + i) = -\left(\sum_{n=0}^{p-2} B_n(x)\theta^{p-1-n}(x) + B_{p-1}(x) - 1\right)$$

et de plus

$$\mu_\theta = -\theta^{-c_p-1} Tr(\theta^{c_p}).$$

Preuve. Pour ce qui est de l'existence, on fait comme on aurait pu faire au lemme précédent, i.e. on suppose qu'on a une représentation de zéro avec un nombre minimal de coefficients non nuls. On multiplie par le produit des dénominateurs et on obtient une contradiction avec la minimalité de notre choix en mettant un des facteurs en évidence. Maintenant, pour calculer la valeur des coefficients $\mu_{\theta(x)}$, il suffit de multiplier le tout par le produit des dénominateurs et d'évaluer en $\theta^{-1}(x)$. Après avoir utilisé la deuxième partie du lemme 2.5 et le théorème de Wilson, on obtient

$$\begin{aligned} \mu_{\theta(x)} &= -\theta^{p-1}(x) \sum_{n=0}^{p-1} x^n \theta^{-n}(x) \prod_{i=n+1}^{p-1} (1 - i\theta^{-1}(x)) \\ &= -\sum_{n=0}^{p-1} x^n \prod_{i=n+1}^{p-1} (\theta(x) - i) \\ &= -\sum_{n=0}^{p-1} x^{-n} \prod_{i=1}^{n} (\theta(x) + i), \end{aligned}$$

ce qui prouve la première partie de l'énoncé.

Pour la deuxième partie, on utilise la troisième partie du lemme 2.5. Pour finir, on utilise le point (6) du lemme 2.7, et on découvre que

$$\begin{aligned}\mu_\theta &= -\theta^{-c_p-1}\left(\sum_{n=0}^{p-1}\theta^{c_p+1}\prod_{i=1}^{n+1}(\theta+i)\right)\\ &= -\theta^{-c_p-1}\left(\theta^{c_p}\left(\sum_{n=1}^{p-1}\prod_{i=1}^{n}(\theta+i-1)+1\right)\right)\\ &= -\theta^{-c_p-1}Tr(\theta^{c_p}),\end{aligned}$$

ce qui termine la démonstration.□

La fonction rationnelle $F_p(x,z)$ possède un développement formel de Laurent de la forme

$$F_p(x,z) = -\sum_{n\geq 1}\frac{B_{-n}(x)}{z^n};$$

en effet

$$\begin{aligned}F_p(x,z) &= \frac{\sum_{n=0}^{p-1}(xz)^n\prod_{i=n+1}^{p-1}(1-iz)}{1-z^{p-1}-(xz)^p}\\ &= \frac{1}{z}\frac{\sum_{n=0}^{p-1}x^{n-1}\prod_{i=n+1}^{p-1}(z^{-1}-i)}{(xz)^{-p}-(xz)^{-1}-1}\end{aligned}$$

est développable, où on a supposé que $x\in\mathbb{F}_p^*$. En identifiant avec $n=1$, on trouve

$$\begin{aligned}B_{-1}(x) &\equiv \sum_{i=0}^{p-1}x^{i-1}\prod_{j=i+1}^{p-1}(p-j)\equiv\sum_{i=1}^{p}x^{p-i-1}\prod_{j=p-i+1}^{p-1}(p-j)\ (mod\ p)\\ &= \sum_{i=1}^{p}x^{-i}(i-1)!=\sum_{i=0}^{p-1}x^{-i-1}i!\ (mod\ p),\end{aligned}$$

d'où on déduit que

$$xB_{-1}(x)\equiv\kappa(x^{-1})\ (mod\ p),$$

lorsque $x\in\mathbb{F}_p^*$.

Proposition 2.2. *La congruence*

$$B_n\equiv -Tr(\theta^{c_p})Tr(\theta^{c_p-1+n})\ (mod\ p)$$

est valide pour $n\in\mathbb{Z}$.

Preuve. En utilisant le lemme 2.9 et les dernières remarques, nous avons

$$B_n(x)\equiv\sum_{\theta^p(x)=\theta(x)+x}\mu_{\theta(x)}\theta^n(x)\ (mod\ p)$$

pour $n\in\mathbb{Z}$. En effet, il suffit de comparer les coefficients dans les développements. On en déduit alors que

$$\begin{aligned}B_n &\equiv -\sum_{\theta^p=\theta+1}\theta^{-c_p-1+n}Tr(\theta^{c_p})\ (mod\ p)\\ &\equiv -Tr(\theta^{c_p})Tr(\theta^{-c_p-1+n})\ (mod\ p),\end{aligned}$$

ce qui démontre le résultat.□

Remarque. On déduit de ce dernier résultat que les nombres de Bell ont une période modulo p qui divise t_p. Plus précisément, la période des nombres de Bell modulo p est la même que celle de θ. Comme $B_0(x) = 1$, le dernier résultat nous apprend aussi que le nombre de Kurepa est nul modulo p si et seulement si $Tr(\theta^{-c_p-2})$ l'est. Ce dernier élément est égal à

$$Tr\left(\frac{1}{\theta^{c_p}(\theta-1)}\right),$$

étant donné que le point 6 du lemme 2.7 nous dit que $\theta^{-c_p-2} = \theta^{-pc_p}\theta^{-1} = \sigma(\theta^{-c_p}(\theta - 1)^{-1})$ et que le groupe de Galois est cyclique. Cette forme algébrique pour cet élément est mieux adaptée à notre argumentation.

Pour la suite, on définit $\lambda_i \in \mathbb{F}_p$, $0 \leq i \leq p-1$, par

$$\theta^{-c_p} := \sum_{i=0}^{p-1} \frac{\lambda_i}{\theta + i}.$$

Le lemme 2.8 nous assure que de tels λ_i existent et de plus que

$$\lambda_i = Tr\left(\frac{1}{\theta^{c_p}(\theta+i)}\right) \quad \text{pour} \quad 0 \leq i \leq p-1.$$

La proposition précédente et les derniers commentaires nous apprennent ainsi que

$$\kappa_p \equiv 0 \ (mod\ p) \iff \lambda_{p-1} \equiv 0 \ (mod\ p).$$

Lemme 2.10. *Pour $p > 2$, les coefficients $\lambda_i \in \mathbb{F}_p$, $0 \leq i \leq p-1$, satisfont*

$$\sum_{i=1}^{p-1} \frac{\lambda_i}{i} - \lambda_0 = \lambda_{p-1}$$

ainsi que

$$\frac{\lambda_0 - \lambda_i}{i} = \lambda_{i-1} \quad \text{pour} \quad 1 \leq i \leq p-1.$$

En particulier, on a toujours $\lambda_1 = 0$. De plus, pour le cas $p = 2$, on a simplement $c_2 = 1$ et $\theta^{-c_2} = \theta^{-1}$.

Preuve. Clairement, on a

$$\theta^{-pc_p} = \sum_{i=0}^{p-1} \frac{\lambda_i}{\theta + i + 1}$$

et aussi

$$\theta^{-c_p-1} = \sum_{i=0}^{p-1} \frac{\lambda_i}{\theta(\theta+i)} = \frac{\lambda_0}{\theta^2} + \sum_{i=1}^{p-1} \frac{\lambda_i}{\theta(\theta+i)}.$$

Le point 6 du lemme 2.7 nous dit que $\theta^{-pc_p} = \theta^{-c_p-1}$. De plus, le point 3 du même lemme nous permet d'écrire

$$-\frac{1}{\theta} = \sum_{i=0}^{p-1} \frac{1}{\theta(\theta+i)} = \frac{1}{\theta^2} + \sum_{i=1}^{p-1} \left(\frac{i^{-1}}{\theta} - \frac{i^{-1}}{\theta+i} \right)$$

et comme $p > 2$, on a que

$$\sum_{i=1}^{p-1} \frac{1}{i} \equiv 0 \ (mod\ p).$$

On en déduit que

$$\theta^{-c_p-1} = \frac{1}{\theta} \left(\sum_{i=1}^{p-1} \frac{\lambda_i}{i} - \lambda_0 \right) + \sum_{i=1}^{p-1} \left(\frac{\lambda_0 - \lambda_i}{i} \right) \frac{1}{\theta+i}$$

et le résultat s'ensuit en utilisant le lemme 2.8 et en comparant cette expression avec celle pour θ^{-pc_p}.□

2.4 Preuve du théorème de Kurepa, Barsky, Benzaghou.

La preuve se fait par contradiction. On va supposer que $\kappa_p \equiv 0$ modulo p, avec $p > 2$ et on va en déduire que $\lambda_i = 0$ pour $0 \leq i \leq p-1$, ce qui est absurde étant donné que θ est non nul en tant que solution de $x^p \equiv x+1$ modulo p. Supposons donc que $\kappa_p \equiv 0$ modulo p. Nous avons alors

$$\sum_{i=1}^{p-1} \frac{\lambda_i}{i} - \lambda_0 = 0$$

et on fait le changement de variables évident

$$X_0 := \lambda_0 \text{ et } X_i := \frac{\lambda_0 - \lambda_i}{i} \quad \text{pour} \quad 1 \leq i \leq p-1.$$

On obtient ainsi le nouveau système

$$\sum_{i=1}^{p-1} X_i = -X_0$$

ainsi que

$$X_i = -(i-1)X_{i-1} + X_0 \quad \text{pour} \quad 1 \leq i \leq p-1.$$

Il est alors très facile de voir par induction que ces dernières relations sont équivalentes à

$$X_{i+1} = X_0 \sum_{j=0}^{i} (-1)^j \frac{i!}{(i-j)!} \quad \text{pour} \quad 0 \leq i \leq p-2.$$

En sommant, on obtient

$$\begin{aligned}
\sum_{k=1}^{p-1} X_k &= X_0 \sum_{i=0}^{p-2} \sum_{j=0}^{i} (-1)^j \frac{i!}{(i-j)!} \\
&= X_0 \sum_{j=0}^{\infty} (-1)^j \sum_{i=0}^{p-2} \frac{i!}{(i-j)!} \\
&= X_0 \sum_{j=0}^{\infty} (-1)^j j! \binom{p-1}{j+1} \\
&\equiv -X_0 \sum_{j=0}^{p-1} j! \ (mod\ p) \\
&\equiv -X_0 \kappa_p \ (mod\ p),
\end{aligned}$$

où on a utilisé le fait que

$$\binom{p-1}{j+1} = \prod_{k=1}^{j+1} \left(\frac{p-k}{k} \right) \equiv (-1)^{j+1} \ (mod\ p).$$

On a donc montré que

$$\kappa_p \equiv 0 \ (mod\ p) \Rightarrow \sum_{i=1}^{p-1} X_i \equiv 0 \ (mod\ p) \Rightarrow X_0 \equiv 0 \ (mod\ p)$$

et il est facile de voir que ceci entraîne que $\lambda_i = 0$ pour $0 \leq i \leq p-1$. La démonstration est donc complète.□

Remarques. Comme on a vu, le fait que $\kappa_p \not\equiv 0$ modulo p est une sorte de petit miracle comme il est très difficile d'en trouver en mathématiques. En effet, en modifiant très légèrement le problème, on obtient des résultats bien différents. Par exemple, les nombres de Kurepa alternés, i.e. les entiers $n > 1$ pour lesquels $n | \kappa_n(-1)$, sont plus fréquents. Il est facile de voir qu'un nombre de Kurepa alterné se factorise en nombres qui sont des puissances de nombres premiers et de Kurepa alterné. Les seuls générateurs connus sont 2, 4, 5, 13, 37 et 463 ; ce qui donne lieu à 47 nombres de Kurepa alternés. Il est possible de montrer qu'il n'y a pas d'autres générateurs inférieurs à 3 millions. En conclusion, il semble que rien n'empêche l'existence d'une infinité de nombres de Kurepa alternés ce qui contraste avec les nombres de Kurepa. Bien sûr, on a remarqué tout au long du chapitre les endroits où les chemins divergent significativement dans les structures pour donner naissance à ce phénomène.

3 Hypothèse de Riemann.

Jusqu'ici, nous avons considérées des équations sur $\mathbb{F}_p$ que nous pouvons, en quelque sorte, voir comme transcendantes. Dans ce chapitre, nous allons démontrer un résultat très général pour les fonctions algébriques de deux variables sur $\mathbb{F}_q$, ce résultat étant équivalent à l'hypothèse de Riemann pour les courbes sur les corps finis. Nous avons déjà vu que dans le cas d'un polynôme $f(x)$ à une variable de $\mathbb{F}_q[x]$, le nombre de zéros de $f(x)$ est compris dans l'intervalle $[0, deg\ f(x)]$ et que chacune des valeurs peuvent être prises si on n'a pas d'autre information sur $f(x)$. Dans le cas des polynômes de $\mathbb{F}_q[x, y]$ il y a certaines similarités, mais en général la situation est complètement différente et beaucoup plus complexe. Premièrement, il faut restreindre l'ensemble des polynômes considérés pour faire un énoncé clair, vrai et général. En effet, il faut gérer la façon dont la courbe se factorise sur $\overline{\mathbb{F}}_q$ et à partir de ça on peut ensuite tirer des conclusions sur son nombre de zéros dans un cas donné (sans en faire le calcul explicite).

Définitions. On dit qu'une courbe $C : f(x, y) = 0$ où $f(x, y) \in \mathbb{F}_q[x, y]$ est absolument irréductible si elle est irréductible dans $\overline{\mathbb{F}}_q[x, y]$. On note le nombre de zéros de $f(x, y)$ dans $\mathbb{F}_{q^n} \times \mathbb{F}_{q^n}$ par N_n. Dans la cas où $f(x, y)$ est absolument irréductible, on définit la fonction zêta $\mathcal{Z}(t, C)$ associée à C par

$$\mathcal{Z}(t, C) := exp\left(\sum_{n=1}^{\infty} \frac{N_n t^n}{n}\right).$$

Le point clef, c'est qu'il est possible de montrer que $\mathcal{Z}(t, C)$ est en fait une fonction rationnelle en t de la forme

$$\mathcal{Z}(t, C) = \frac{\prod_{i=1}^{2g}(1 - \alpha_i t)}{(1 - t)(1 - qt)}$$

où g est le genre de la courbe C (voir [28]). On peut donc déduire, en prenant le logarithme de chaque coté, que

$$N_n = q^n + 1 - \sum_{i=1}^{2g} \alpha_i^n.$$

L'hypothèse de Riemann est la supposition que les $(\alpha_i)_{1 \leq i \leq 2g}$ valent exactement $q^{1/2}$ en valeur absolue. On obtient ainsi la fantastique borne

$$|N_n - q^n - 1| \leq 2gq^{n/2} \leq (d - 1)(d - 2)q^{n/2}$$

étant donné que $g \leq (d - 1)(d - 2)/2$ (voir [28]).

La richesse de cette structure nous apprend, au prix d'un léger raisonnement qui utilise le principe des nids de pigeons, qu'une borne de la forme

$$|N_n - q^n - 1| \leq A_C q^{n/2},$$

où A_C est une constante quelconque qui ne dépend que de C, est suffisante pour montrer l'hypothèse de Riemann. En effet, il faut juste voir essentiellement qu'il existe une infinité de valeurs de n qui alignent les racines plutôt bien et on prend la limite lorsque n tend vers l'infini.

Le reste de ce chapitre est voué à une preuve de cette dernière borne. Nous suivons plus on moins le procédé et la notation tels que présentés dans le livre de Wolfgang M. Schmidt [31], qui sont basés sur la méthode de Stepanov.

3.1 Résultats préliminaires.

Définitions. Un polynôme à d variables est dit symétrique s'il est invariant aux permutations de S_d. Les polynômes symétriques élémentaires sont

$$s_1(x_1, \ldots, x_d) := \sum_{i=1}^{d} x_i,$$

$$s_2(x_1, \ldots, x_d) := \sum_{1 \leq i < j \leq d} x_i x_j,$$

$$\vdots$$

$$s_d(x_1, \ldots, x_d) := \prod_{i=1}^{d} x_i.$$

Lemme 3.1. *Soit $a(x_1, \ldots, x_d)$ un polynôme symétrique à coefficients dans le corps K. Alors, il existe un polynôme $b(z_1, \ldots, z_d)$ à coefficients dans K pour lequel*

$$a(x_1, \ldots, x_d) = b(s_1(x_1, \ldots, x_d), \ldots, s_d(x_1, \ldots, x_d)).$$

De plus, si $a(x_1, \ldots, x_d)$ est de degré δ pour chaque variable x_i avec $1 \leq i \leq d$, alors $b(z_1, \ldots, z_d)$ est de degré total δ. De même, si tous les monômes de $a(x_1, \ldots, x_d)$ sont de degré e, alors chaque monôme de $b(z_1, \ldots, z_d)$ de la forme $cz_1^{i_1} \ldots z_d^{i_d}$ avec $c \in K$ satisfait la relation

$$\sum_{j=1}^{d} j i_j = e.$$

Preuve. La preuve se fait par induction complète sur le degré total et on se restreint facilement au cas où tous les monômes sont de même degré, disons e. Le cas $e = 1$ est trivial ; donc supposons que $e > 1$. On impose alors un ordre lexicographique sur les variables, par exemple

$$x_1 > \ldots > x_d.$$

Dans ce cas, on peut comparer deux monômes quelconques et il est très facile de voir qu'on peut fabriquer de façon unique le terme dominant avec les polynômes symétriques élémentaires et un élément de K. Il suffit alors de soustraire les deux expressions et nous

obtenons un polynôme "plus petit". Après un nombre fini d'étapes, on obtient ainsi un polynôme de degré total strictement inférieur à e et le résultat s'ensuit.□

Il est intéressant de remarquer qu'un polynôme $f(x,y)$ absolument irréductible sur K reste absolument irréductible sur $K(x)$. En effet, supposons le contraire ; alors

$$f(x,y) = f_1(x,y)f_2(x,y)$$

avec $f_i(x,y)$ des polynômes en y à coefficients dans $K(x)$ pour $i = 1,2$. Pour un polynôme quelconque $h(x,y) \in K(x)[y]$, on peut toujours écrire de façon unique

$$h(x,y) = \frac{v(x)}{w(x)}\hat{h}(x,y) =: g(h)\hat{h}(x,y),$$

où $v(x)$ et $w(x)$ sont des polynômes premiers entre eux et

$$\hat{h}(x,y) = \sum_j c_j(x)y^j$$

avec $c_j(x)$ pour $0 \leq j \leq deg_y\ h(x,y)$ des polynômes aussi premiers entre eux. Comme $K[x]$ est un domaine à factorisation unique, il s'ensuit que $g(f_1f_2) = g(f_1)g(f_2)$. Étant donné que $f(x,y)$ est irréductible, on a que $g(f) = 1$. On en déduit donc que

$$f(x,y) = g(f_1)\hat{f}_1(x,y)g(f_2)\hat{f}_2(x,y) = \hat{f}_1(x,y)\hat{f}_2(x,y),$$

ce qui contredit l'irréductibilité de $f(x,y)$. On en déduit que si $f(x,\beta) = 0$ avec $deg_y\ f(x,y)$ et $\beta \in \overline{K(x)}$, alors $[K(x,\beta),K(x)] = d$.

Lemme 3.2. *Soit $f(x,y)$ et $g(z,t)$ deux polynômes absoluments irréductibles à coefficients dans K avec $deg_y\ f(x,y) = d$ et $deg_y\ g(z,t) = e$, où $ed > 0$. Supposons que $\beta \in \overline{K(x)}$ et $\gamma \in \overline{K(z)}$ sont des valeurs pour lesquelles $f(x,\beta) = g(z,\gamma) = 0$. Alors*

$$[K(x,z,\beta,\gamma),K(x,z)] = ed.$$

Preuve. Il faut montrer que

$$[K(x,z,\beta,\gamma),K(x,z,\beta)] = e$$

et que

$$[K(x,z,\beta),K(x,z)] = d.$$

Pour ce qui est de la première relation, il suffit de montrer que $g(z,t)$ reste irréductible sur $K(x,\beta)$. Supposons le contraire ; alors

$$g(z,t) = g_1(z,t)g_2(z,t),$$

où $g_i(z,t)$ est à coefficients dans $K(x,\beta)$ et est de degré strictement inférieur à e pour $i = 1,2$. Écrivons

$$g_i(z,t) = \sum_{j,k} c_{ijk}(x,\beta)z^jt^k, \quad i = 1,2,$$

où

$$c_{ijk}(x,\beta) := \sum_{l=0}^{d-1} r_{ijkl}(x)\beta^l$$

pour certaines fonctions rationnelles $r_{ijkl}(x)$ avec $0 \leq l \leq d-1$. Choisissons alors $a \in \overline{K}$ pour lequel les dénominateurs des $r_{ijkl}(a)$ avec $0 \leq l \leq d-1$ sont non nuls et tel que si

$$f(x,y) = \sum_{m=0}^{d} \alpha_m(x) y^m,$$

alors $\alpha_d(a)$ est non nul. Choisissons ensuite $b \in \overline{K}$ pour lequel $f(a,b) = 0$. Ainsi, la paire (a,b) satisfait toutes les équations qui sont satisfaites par (x,β). On pose alors

$$\bar{g}_i(z,t) = \sum_{j,k} c_{ijk}(a,b) z^j t^k, \quad i = 1,2,$$

et on observe que les valeurs $c_{ijk}(a,b)$ sont dans $\overline{K}$ pour tout i,j et k. On obtient la factorisation

$$g(z,t) = \bar{g}_1(z,t)\bar{g}_2(z,t),$$

ce qui contredit le fait que $g(z,t)$ est absolument irréductible. L'autre relation se fait de la même façon, mais c'est un peu plus court.□

Remarque importante. Si $f(x,y)$ est un polynôme de degré total d, alors sans perte de généralité, on peut toujours supposer que $f(x,y)$ est séparable en y, i.e. n'est pas un polynôme en y^p et qu'il est de la forme

$$f(x,y) = y^d + \sum_{i=1}^{d} \alpha_i(x) y^{d-i},$$

où $deg\ \alpha_i(x) \leq i$ pour $1 \leq i \leq d$.

Preuve. Supposons que $f(x,y) = h(x,y^p)$, où $q = p^n$. Comme $y \mapsto y^p$ est un automorphisme de $\mathbb{F}_q$, il s'ensuit que cette opération permute les éléments de $\mathbb{F}_q$. On voit alors que le nombre de zéros de $h(x,y)$ est le même que celui de $f(x,y)$, sauf que $deg_y\ h(x,y) \leq deg_y\ f(x,y)$. On peut répéter le même argument et après un nombre fini d'étapes on obtient un polynôme séparable en y.

Si

$$f(x,y) = \sum_{i,j=0}^{d} \alpha_{i,j} x^i y^j$$

est séparable en y, alors il existe $\alpha_{i_0,j_0} \neq 0$ avec $p \not| j_0$. Posons alors

$$g(x,y) := f(x+by,y) = \sum_{i,j=0}^{d} \alpha_{i,j}(b) x^i y^j.$$

Ainsi, les coefficients $\alpha_{i,j}(b)$ sont des polynômes en b de degré au plus d. Le polynôme $\alpha_{i_0,j_0}(b)$ n'est pas identiquement 0 et le coefficient de y^d est exactement $\alpha_{0,d}(b) = f_d(b,1)$,

où $f_d(x,y)$ est la partie de $f(x,y)$ qui est de degré total d qui est non identiquement nulle. Si $q > 2d$, ce qui ne pose pas de problème pour le résultat final, alors on peut choisir une valeur de $b \in \mathbb{F}_q$ pour laquelle $\alpha_{0,d}(b) \neq 0$ et $\alpha_{i_0,j_0}(b) \neq 0$. On obtient alors un polynôme $g(x,y)$ pour lequel le coefficient de y^d est non nul. Le résultat s'ensuit en divisant par la bonne constante.□

Lemme 3.3. *Soit K un corps de caractéristique p et $f(x,y) \in K[x,y]$ un polynôme de degré d absolument irréductible de la forme*

$$f(x,y) = y^d + \sum_{i=1}^{d} \alpha_i(x) y^{d-i},$$

où $deg\ \alpha_i(x) \leq i$ pour $1 \leq i \leq d$. Supposons que $f(x,\beta) = 0$ avec $\beta \in \overline{K(x)}$. Si $a(x,y,z,t)$ est un polynôme non identiquement nul pour lequel

$$\begin{cases} deg_x\ a & \leq q/d - d, \\ deg_y\ a & \leq d-1, \\ deg_t\ a & \leq d-1, \end{cases}$$

alors $a(x,\beta,x^q,\beta^q) \neq 0$.

Preuve. Posons

$$\hat{a}(x,y,z;t_1,\ldots,t_d) := \prod_{i=1}^{d} a(x,y,z,t_i).$$

C'est un polynôme à $d+3$ variables, symétrique en $t_1,\ldots,t_d$. D'après le lemme 3.1, on peut écrire

$$\hat{a}(x,y,z,t_1,\ldots,t_d) = b(x,y,z;s_1(t_1,\ldots,t_d),\ldots,s_d(t_1,\ldots,t_d)),$$

où le degré de $b(x,y,z;v_1,\ldots,v_d)$ en $v_1,\ldots,v_d$ est d'au plus $d-1$.

Comme

$$\beta^d = -\sum_{i=1}^{d} \alpha_i(x)\beta^{d-i},$$

nous pouvons écrire

$$\beta^{d+t-1} = \sum_{i=1}^{d} \alpha_{i,t}(x)\beta^{d-i}$$

pour chaque entier positif t. On vérifie facilement par induction que $deg\ \alpha_{i,t}(x) \leq i+t-1$ pour $1 \leq i \leq d$. Puisque $deg_y\ b \leq d(d-1) = (d-1) + (d-1)^2$, on peut utiliser les dernières relations avec $t \leq (d-1)^2$ pour écrire

$$b(x,\beta,z;v_1,\ldots,v_d) = c(x,\beta,z;v_1,\ldots,v_d),$$

avec $deg_y\ c \leq d-1$. De plus,

$$\begin{cases} deg_x\ c \leq & deg_x\ b + (d + (d-1)^2 - 1), \\ deg_x\ c \leq & d(deg_x\ a) + d(d-1), \\ deg_x\ c < & q. \end{cases}$$

Supposons que $a(x, \beta, x^q, \beta^q) = 0$. Posons $\beta_1 = \beta$ et

$$f(x, y) = \prod_{i=1}^{d} (y - \beta_i).$$

On a donc que $\hat{a}(x, \beta, x^q; \beta_1^q, \ldots, \beta_d^q) = 0$ et comme $s_i(\beta_1, \ldots, \beta_d) = (-1)^i \alpha_i(x)$ pour $1 \leq i \leq d$, il s'ensuit que $s_i(\beta_1^q, \ldots, \beta_d^q) = (-1)^i (\alpha_i(x))^q$ pour $1 \leq i \leq d$. D'où

$$b(x, \beta, x^q; -(\alpha_1(x))^q, \ldots, (-1)^d (\alpha_d(x))^q) = 0.$$

Ainsi

$$c(x, \beta, x^q; -(\alpha_1(x))^q, \ldots, (-1)^d (\alpha_d(x))^q) = 0.$$

Étant donné que $deg_y\ c \leq d - 1$ et que β est un nombre algébrique de degré d, il s'ensuit que

$$c(x, y, x^q; -(\alpha_1(x))^q, \ldots, (-1)^d (\alpha_d(x))^q) = 0.$$

Si on pose $x = x_1 + x_2$, on obtient que pour un certain polynôme $r(x_1, x_2, y)$ l'expression

$$c(x_1 + x_2, y, x_1^q; -(\alpha_1(x_1))^q, \ldots, (-1)^d (\alpha_d(x_1))^q) + x_2^q r(x_1, x_2, y) = 0$$

est vérifiée. Maintenant, puisque $deg_x\ c < q$, il s'ensuit que

$$c(x_1 + x_2, y, x_1^q; -(\alpha_1(x_1))^q, \ldots, (-1)^d (\alpha_d(x_1))^q) = 0$$

et comme les variables $x_1 + x_2$, y et x_1^q sont algébriquement indépendantes, on en conclut qu'on peut écrire

$$c(x, y, z; -\bar{\alpha}_1(z), \ldots, (-1)^d \bar{\alpha}_d(z)) = 0,$$

où on a utilisé la notation $\bar{\alpha}_i(z)$ avec $1 \leq i \leq d$ pour signifier que les coefficients des $\alpha_i(x)$ sont élevés à la puissance q. En posant $y = \beta$, on a

$$b(x, \beta, z; -\bar{\alpha}_1(z), \ldots, (-1)^d \bar{\alpha}_d(z)) = 0.$$

Supposons que $\bar{f}(z, \gamma) = 0$ avec $\gamma = \gamma_1$ et $\gamma \in \overline{K(x)}$. On peut alors écrire

$$\bar{f}(z, t) = \prod_{i=1}^{d} (t - \gamma_i)$$

et donc $s_i(\gamma_1, \ldots, \gamma_d) = (-1)^i \bar{\alpha}_i(z)$ pour $1 \leq i \leq d$. On a donc

$$\hat{a}(x, \beta, z; \gamma_1, \ldots, \gamma_d) = 0,$$

d'où on déduit qu'au moins un des facteurs s'annule, par exemple en $i = 1$ et $a(x, \beta, z, \gamma) = 0$. En utilisant le lemme 3.2 et le fait que $f(x, \beta) = \bar{f}(z, \gamma) = 0$, on a que les d^2 valeurs $\beta^i \gamma^j$ avec $0 \leq i, j \leq d - 1$ sont linéairement indépendantes et donc que $a(x, y, z, t) = 0$, une contradiction.□

Lemme 3.4. *Soit*

$$f(x,y) = y^d + \sum_{i=1}^{d} \alpha_i(x) y^{d-i},$$

où deg $\alpha_i(x) \leq i$ *pour* $1 \leq i \leq d$. *Supposons que*

$$f(x,y) = \prod_{i=1}^{d} (y - \beta_i),$$

avec $\beta_i \in \overline{K(x)}$ *pour* $1 \leq i \leq d$. *Si* $a(x; y_1, \ldots, y_d)$ *est un polynôme symétrique en* $y_1, \ldots, y_d$, *alors* $a(x; \beta_1, \ldots, \beta_d) = b(x)$ *avec deg* $b \leq$ *deg total* $a(x; y_1, \ldots, y_d)$.

Preuve. Soit δ le degré total de $a(x; y_1, \ldots, y_d)$. On peut écrire

$$a(x; y_1, \ldots, y_d) = \sum_{k=0}^{\delta} x^k c_k(y_1, \ldots, y_d),$$

où $c_k(y_1, \ldots, y_d)$ est un polynôme de degré au plus $\delta - k$, symétrique en $y_1, \ldots, y_d$. D'après le lemme 3.1,

$$c_k(y_1, \ldots, y_d) = h_k(s_1(y_1, \ldots, y_d), \ldots, s_d(y_1, \ldots, y_d)).$$

De plus, pour chaque terme $s_1^{i_1} \ldots s_d^{i_d}$, on a

$$\sum_{j=1}^{d} j i_j \leq \delta - k.$$

On en déduit que

$$c_k(\beta_1, \ldots, \beta_d) = h_k(-\alpha_1(x), \ldots, (-1)^d \alpha_d(x))$$

et que pour chaque terme $\alpha_1^{i_1}(x) \ldots \alpha_d^{i_d}(x)$ on a

$$\sum_{j=1}^{d} j i_j \leq \delta - k.$$

Le terme $x^k c_k(\beta_1, \ldots, \beta_d)$ de $a(x; \beta_1, \ldots, \beta_d)$ est donc un polynôme de degré au plus δ et le résultat s'ensuit.□

Lemme 3.5. *Soit* $f(x,y) \in \mathbb{F}_q[x,y]$ *un polynôme absolument irréductible et* $\beta \in \overline{K(x)}$ *tel que* $f(x,\beta) = 0$. *Soit* M *un entier positif et* $a(x,y)$ *un polynôme. Alors pour* $0 \leq l \leq M$,

$$\varrho^l(f_y^{2M}(x,\beta)a(x,\beta)) = f_y^{2M-2l}(x,\beta)a_l(x,\beta),$$

où $a_l(x,y)$ *est un polynôme avec* $a_l(x,y) \leq$ *deg* $a(x,y) + (2d-3)l$.

Preuve. La preuve se fait par induction sur l. Pour $l = 0$, il n'y a rien à prouver. Supposons que le lemme est vrai pour $0 \leq l < M$. On a donc

$$\begin{aligned}\varrho^{l+1}(f_y^{2M}(x,\beta)a(x,\beta)) &= \varrho(f_y^{2M-2l}(x,\beta)a_l(x,\beta)) \\ &= \tfrac{2M-2l}{l+1} f_y^{2M-2l-1}(x,\beta)(f_{yx}(x,\beta) + f_{yy}(x,\beta)\partial\beta)a_l(x,\beta) \\ &\quad + \tfrac{1}{l+1} f_y^{2M-2l}(x,\beta)(a_{ly}(x,\beta) + a_{lx}(x,\beta)\partial\beta).\end{aligned}$$

Premièrement, on a que

$$\partial\beta = -\frac{f_x(x,y)}{f_y(x,y)}.$$

Deuxièmement, il est très important de remarquer que, malgrés le changement de variables inductif qui ne fait que déguiser la situation, à la $(l+1)$-ième étape tous les coefficients sont divisibles par $l+1$. Ce facteur provient de l'ensemble des $l+1$ dérivations déjà faites et donc il n'y a aucun danger de division par 0 dans le cas où $l+1$ est un multiple de p. Il faut regarder les coefficients de ces polynômes d'une façon symbolique comme si on était en caractéristique 0. En somme, on obtient

$$\begin{aligned}&f_y^{2M-2(l+1)}(x,\beta)(\tfrac{1}{l+1}((2M-2l)(f_y(x,\beta)f_{yx}(x,\beta) - f_{yy}(x,\beta)f_x(x,\beta))a_l(x,\beta) \\ &+(f_y^2(x,\beta)a_{ly}(x,\beta) - a_{lx}(x,\beta)f_x(x,\beta)f_y(x,\beta)))) = f_y^{2M-2(l+1)}(x,\beta)a_{l+1}(x,\beta).\end{aligned}$$

Il est donc clair que

$$deg\ a_{l+1}(x,y) \leq deg\ a_l(x,y) + 2d - 3 \leq deg\ a(x,y) + (2d-3)(l+1),$$

et le résultat s'ensuit.□

Remarques. Supposons que $0 < m < M \leq q$. Alors il est facile de voir que $\varrho^m(x^{qj}) = 0$ et que de même $\varrho^m(\beta^{qk}) = 0$.

3.2 Construction de certains polynômes.

Supposons que $d > 1$, sinon les résultats de ce chapitre sont triviaux. On suppose aussi que $f(x,y) \in \mathbb{F}_q[x,y]$ est un polynôme absolument irréductible de degré d sous sa forme réduite, tel que $f(x,\beta) = 0$, où $\beta \in \overline{\mathbb{F}_q(x)}$. Afin de conserver une certaine élégance avec nos estimés, nous allons supposer que $q > 250d^5$ ce qui ne change rien au résultat final.

Soit

$$f(y) = y^d + \sum_{n=1}^{d} a_n y^{d-n} = \prod_{n=0}^{d}(y - \gamma_n).$$

Le discriminant $\Delta(f)$, qui ne devrait pas être confondu avec l'opérateur de différence finie, est défini par

$$\Delta(f) := \prod_{1 \leq i < j \leq d} (\gamma_j - \gamma_i)^2.$$

On déduit du lemme 3.1 que $\Delta(f)$ est un polynôme de degré $2d-2$ en $(\gamma_i)_{1\leq i\leq d}$ et que pour chaque monôme de la forme $a\gamma_1^{i_1}\dots\gamma_d^{i_d}$ la relation

$$\sum_{j=1}^{d} ji_j = d(d-1)$$

est valide.

Soit maintenant

$$f(x,y) = y^d + \sum_{i=1}^{d} \alpha_i(x)y^{d-i},$$

où $deg\ \alpha_i(x) \leq i$ pour $1 \leq i \leq d$. Notons par $\Delta(x)$ le discriminant de $f(x,y)$ vu comme polynôme en y. Clairement, $\Delta(x)$ est un polynôme en x et $deg\ \Delta(x) \leq d(d-1)$.

Soit $\mathcal{G}$ l'ensemble des valeurs de $v \in \mathbb{F}_q$ avec $\Delta(v) \neq 0$. Alors

$$q - d(d-1) \leq |\mathcal{G}| \leq q.$$

Si $v \in \mathcal{G}$, le polynôme $f(v,y)$ a d racines distinctes $\xi_1, \dots, \xi_d \in \overline{\mathbb{F}}_q$. On s'intéresse surtout aux racines qui sont dans $\mathbb{F}_q$. Posons alors

$$\mathcal{I}_1(v) := \{\xi \in \mathbb{F}_q |\ f(v,\xi) = 0\}$$

et

$$\mathcal{I}_2(v) := \{\xi \in \overline{\mathbb{F}}_q |\ f(v,\xi) = 0,\ \xi \notin \mathbb{F}_q\}.$$

On a donc pour chaque $v \in \mathcal{G}$ que

$$|\mathcal{I}_1(v)| + |\mathcal{I}_2(v)| = d.$$

Posons $\alpha_0(x) = 1$, puis définissons

$$\mathcal{E}_1(x,y,\bar{y}) := y - \bar{y}$$

et

$$\mathcal{E}_2(x,y,\bar{y}) := \sum_{j=1}^{d} \alpha_{d-j}(x) \sum_{i=0}^{j-1} y^{j-1-i}\bar{y}^i.$$

On a donc

$$f(x,y) - f(x,\bar{y}) = \mathcal{E}_1(x,y,\bar{y})\mathcal{E}_2(x,y,\bar{y}).$$

Si $v \in \mathcal{G}$ et $\xi \in \mathcal{I}_1(v) \cup \mathcal{I}_2(v)$, alors

$$0 = f(v,\xi) = (f(v,\xi))^q = f(v,\xi^q)$$

et donc

$$0 = f(v,\xi) - f(v,\xi^q) = (\xi - \xi^q)\mathcal{E}_2(v,\xi,\xi^q).$$

Si $\xi \in \mathcal{I}_1(v)$, alors $\xi \in \mathbb{F}_q$ et $\xi - \xi^q = 0$; et comme y est une racine simple de $f(v,y)$, il s'ensuit que $\mathcal{E}_2(v,\xi,\xi^q) \neq 0$. Si $\xi \in \mathcal{I}_2(v)$, alors $\xi \neq \xi^q$ et $\mathcal{E}_2(v,\xi,\xi^q) = 0$. Ainsi pour $\lambda = 1$ ou 2, $\mathcal{I}_\lambda(v)$ est l'ensemble des ξ avec

$$f(v,\xi) = 0 \quad \text{et} \quad \mathcal{E}_\lambda(v,\xi,\xi^q) = 0.$$

Pour simplifier la notation, on pose

$$\epsilon_1 := deg\ \mathcal{E}_1(v,\xi,\xi^q) = 1 \quad \text{et} \quad \epsilon_2 := deg\ \mathcal{E}_2(v,\xi,\xi^q) = d-1.$$

Lemme 3.6. *Supposons que $\lambda = 1$ ou 2. Soit M un entier positif tel que*

$$d|M,\ M \geq d^2 \quad \text{et} \quad 2(d-1)(M+8)^2 \leq q.$$

Alors il existe un polynôme $a(x,y)$ pour lequel $a(x,\beta) \neq 0$ et tel que si $a_l(x,y)$ est le polynôme défini au lemme 3.5, alors

$$a_l(v,\xi) = 0 \quad \text{pour} \quad 0 \leq l < M,$$

où $v \in \mathcal{G}$ et $\xi \in \mathcal{I}_\lambda(v)$. De plus, la borne

$$deg\ a(x,y) \leq (\epsilon_\lambda/d)qM + q(d-3/2)$$

est satisfaite.

Preuve. Posons

$$h(x,y,z,t) := \sum_{\substack{(j,k)=(0,0)\\ j+k\leq N}}^{(N,d-1)} b_{jk}(x,y) z^j t^k$$

avec

$$b_{jk}(x,y) = \sum_{i=0}^{d-1} a_{ijk}(x) y^i,$$

où

$$deg\ a_{ijk}(x) \leq q/d - d - k - j - i$$

et

$$N := (\epsilon_\lambda/d)M + d - 2.$$

Posons alors $a(x,y) := h(x,y,x^q,y^q)$.

D'après le lemme 3.3, si les polynômes $a_{ijk}(x)$ ne sont pas tous identiquement nuls, alors $a(x,\beta) \neq 0$. Comme $\varrho^m(x^{qj}) = \varrho^m(\beta^{qk}) = 0$ lorsque $1 \leq m \leq q$, il s'ensuit que

$$a_l(x,y) = \sum_{\substack{(j,k)=(0,0)\\ j+k\leq N}}^{(N,d-1)} b_{jkl}(x,y) x^{qj} y^{qk}$$

et de plus selon le lemme 3.5 on a

$$deg\ b_{jkl}(x,y) \leq deg\ b_{jk}(x,y) + (2d-3)l \leq q/d - d - k - j + (2d-3)l.$$

On veut donc que $a_l(v,\xi) = 0$ pour $0 \leq l < M$, où $v \in \mathcal{G}$ et $\xi \in \mathcal{I}_\lambda(v)$.

Cas 1 : $\lambda = 1$. On a $v, \xi \in \mathbb{F}_q$, donc $v^q = v$ et $\xi^q = \xi$. Alors le polynôme

$$c_l(x,y) = \sum_{\substack{(j,k)=(0,0) \\ j+k \leq N}}^{(N,d-1)} b_{jkl}(x,y) x^j y^k$$

s'annule pour chaque paire (v,ξ) considérée. Notons que

$$deg\ c_l(x,y) \leq q/d + (2d-3)l - 2.$$

Cas 2 : $\lambda = 2$. On a $v \in \mathbb{F}_q$, $f(x,\xi) = 0$ et $\mathcal{E}_2(v,\xi,\xi^q) = 0$. Donc $v^q = v$ et

$$0 = \mathcal{E}_2(v,\xi,\xi^q) = \sum_{j=1}^{d} \alpha_{d-j}(v) \sum_{i=0}^{j-1} \xi^{j-1-i} \xi^{qi}.$$

Ainsi, on peut exprimer $\xi^{q(d-1)}$ en termes de $(\xi^{qj})_{0 \leq j \leq d-2}$ avec des coefficients qui sont des polynômes en v et ξ de degré au plus $d-1$. On a donc besoin qu'un certain polynôme $c_l(x,y,\bar{y})$ s'annule en (v,ξ,ξ^q), avec $deg_{\bar{y}}\ c_l(x,y,\bar{y}) \leq q-2$ et un degré total d'au plus $q/d + (2d-3)l - 2$ en x, y.

Dans les deux cas, on a besoin qu'un certain polynôme $c_l(x,y,\bar{y})$ s'annule en (v,ξ,ξ^q), où

$$deg\ total_{x,y}\ c_l(x,y,\bar{y}) \leq q/d + (2d-3)l - 2$$

et

$$deg_{\bar{y}}\ c_l(x,y,\bar{y}) \leq \epsilon_\lambda - 1.$$

Nous savons que pour une paire (v,ξ) avec $f(v,\xi) = 0$,

$$\xi^d = -\sum_{n=1}^{d} \alpha_n(v) \xi^{d-n}$$

et que pour un entier positif t,

$$\xi^{d+t-1} = \sum_{n=1}^{d} \alpha_{nt}(v) \xi^{d-n}$$

avec $deg\ \alpha_{nt}(x) \leq t - i + n$. Donc, $c_l(v,\xi,\xi^q) = 0$ précisément si un certain polynôme $d_l(v,\xi,\xi^q) = 0$, où

$$\begin{cases} deg_x\ d_l(x,y,\bar{y}) \leq q/d + (2d-3)l - 2, \\ deg_y\ d_l(x,y,\bar{y}) \leq d-1, \\ deg_{\bar{y}}\ d_l(x,y,\bar{y}) \leq \epsilon_\lambda - 1. \end{cases}$$

Il suffit donc de construire $d_l(x,y,\bar{y})$ pour qu'il soit identiquement nul pour $0 \leq l < M$.

Le nombre de coefficients de $d_l(x, y, \bar{y})$ est inférieur à $\epsilon_\lambda q + (2d^2 - 3d)\epsilon_\lambda l$. Le nombre total B de l'ensemble des coefficients de $d_l(x, y, \bar{y})$ avec $0 \leq l < M$ satisfait

$$B < \epsilon_\lambda qM + (d^2 - 3d/2)\epsilon_\lambda M^2.$$

Ces coefficients sont des combinaisons linéaires des coefficients de $a_{ijk}(x)$. On obtient alors un système d'équations homogènes linéaires en les coefficients jusqu'ici indéterminés des polynômes $a_{ijk}(x)$. Le nombre de coefficients disponibles pour $a_{ijk}(x)$ est d'au moins

$$q/d - d - k - j - i > q/d - j - 3d.$$

Le nombre de coefficients disponibles en sommant sur j, avec $0 \leq j \leq N - k$, est d'au moins

$$(q/d - 3d)(N + 1) - (N - k)(N - k + 1)/2 - kq/d.$$

En sommant sur k, avec $0 \leq j \leq d - 1$, on obtient

$$(q - 3d^2)(N + 1) - N^2 d/2 - q(d - 1)/2$$

et on obtient le nombre total A de coefficients disponibles en sommant sur i, avec $0 \leq j \leq d - 1$. On a alors

$$\begin{aligned} A &> (q - 3d^2)(Nd + d) - N^2 d^2/2 - qd(d - 1)/2 \\ &> (q - 3d^2)(\epsilon_\lambda M + d^2 - d) - (\epsilon_\lambda M + d^2)^2/2 - qd(d - 1)/2 \end{aligned}$$

et comme par hypothèse $M \geq d^2$, il s'ensuit que

$$A > \epsilon_\lambda Mq + (d^2 - d)q/2 - \epsilon_\lambda^2 M^2/2 - 8\epsilon_\lambda M d^2.$$

Donc, pour que les polynômes $d_l(x, y, \bar{y})$ s'annulent, il faut résoudre un système homogène à B équations en A variables. Pour être certain d'avoir une solution, il suffit que $B < A$. On a donc besoin que

$$\epsilon_\lambda M^2(2d^2 - 3d + \epsilon_\lambda) + 16\epsilon_\lambda M d^2 < d(d - 1)q$$

et comme $\epsilon_\lambda = 1$ ou $d - 1$, il s'ensuit que

$$M^2(d - 1)(2d^2 - 2d - 1) + 16Md^2(d - 1) < d(d - 1)q$$

est suffisant. Cette inégalité est satisfaite si

$$2M(d - 1) + 16Md < q,$$

ce qui est vrai selon notre hypothèse de départ : $2(d - 1)(M + 8)^2 \leq q$. Finalement,

$$deg\ a(x, y) \leq Nq + q/d \leq (\epsilon_\lambda/d)qM + q(d - 3/2)$$

et le lemme est démontré.□

Si on pose

$$c(x,y) = f_y^{2M}(x,y)a(x,y),$$

alors $c(x,\beta) \neq 0$ et de plus pour $0 \leq l < M$ on a

$$\varrho^l c(x,\beta) = f_y^{2M-2l}(x,\beta)a_l(x,\beta).$$

Ainsi, pour $v \in \mathcal{G}$ et $\xi \in \mathcal{I}_\lambda(v)$ on a $\varrho^l c(v,\xi) = 0$. De plus

$$deg\ c(x,y) \leq (\epsilon_\lambda/d)qM + q(d-3/2) + 2Md,$$

mais si $q > 250d^5$, alors $2Md \leq 2dq^{1/2} = \dfrac{2d}{q^{1/2}}q < \dfrac{q}{2}$ et donc

$$deg\ c(x,y) \leq (\epsilon_\lambda/d)qM + q(d-1).$$

Lemme 3.7. *Supposons que M satisfait les conditions du lemme 3.6. Soit $\lambda = 1$ ou 2 fixé. Alors il existe un polynôme $r(x) \neq 0$ avec*

$$\varrho^l r(v) = 0 \quad \text{pour} \quad v \in \mathcal{G} \quad \text{et} \quad 0 \leq l < M|\mathcal{I}_\lambda(v)|$$

et

$$deg\ r(x) \leq \epsilon_\lambda qM + qd(d-1).$$

Preuve. Soit $\mathcal{R}$ la norme de $\mathbb{F}_q(x,\beta)$ sur $\mathbb{F}_q(x)$ et posons $r(x) := \mathcal{R}(c(x,\beta))$, i.e. si

$$f(x,y) = \prod_{i=1}^{d}(y-\beta_i),$$

alors

$$r(x) = \prod_{i=1}^{d} c(x,\beta_i)$$

et

$$\varrho^l r(x) = \sum_{a_1+\ldots+a_d=l} (\varrho^{a_1}c(x,\beta_1))\ldots(\varrho^{a_d}c(x,\beta_d)).$$

Le coté droit de cette égalité est un polynôme symétrique en $(\beta_i)_{1\leq i\leq d}$, donc un polynôme en les fonctions symétriques élémentaires en $(\beta_i)_{1\leq i\leq d}$ et on peut écrire

$$\varrho^l r(x) = k_l(x; -\alpha_1(x), \ldots, (-1)^d\alpha_d(x)).$$

Pour $v \in \mathbb{F}_q$, on a

$$\varrho^l r(v) = k_l(v; -\alpha_1(v), \ldots, (-1)^d\alpha_d(v)).$$

Si $v \in \mathcal{G}$, alors $f(v,y)$ a d racines distinctes $(\xi_i)_{1\leq i\leq d} \in \overline{\mathbb{F}}_q$ et $s_i(\xi_1,\ldots,\xi_d) = (-1)^i\alpha_i(v)$. Ainsi

$$\varrho^l r(v) = \sum_{a_1+\ldots+a_d=l} (\varrho^{a_1}c(v,\xi_1))\ldots(\varrho^{a_d}c(v,\xi_d)).$$

On a

$$\{\xi_1, \ldots, \xi_d\} = \mathcal{I}_1(v) \cup \mathcal{I}_2(v).$$

Supposons sans perte de généralité que

$$\{\xi_1, \ldots, \xi_t\} = \mathcal{I}_\lambda(v);$$

ainsi $|\mathcal{I}_\lambda(v)| = t$. Chaque terme du développement de $\varrho^l r(v)$ satisfait

$$\sum_{i=1}^{t} a_i \leq l.$$

Il y a donc un entier s avec $1 \leq s \leq t$ pour lequel

$$a_s \leq \frac{l}{t} = \frac{l}{|\mathcal{I}_\lambda(v)|} < \frac{M|\mathcal{I}_\lambda(v)|}{|\mathcal{I}_\lambda(v)|} = M.$$

En utilisant les commentaires qui précèdent ce lemme, on a que $\varrho^{a_s} c(v, \xi_s) = 0$ et donc pour chaque $x \in \mathcal{G}$,

$$\varrho^l r(v) = 0 \quad \text{pour} \quad 0 \leq l < M|\mathcal{I}_\lambda(v)|.$$

Maintenant, on a que

$$r(x) = \prod_{i=1}^{d} c(x, \beta_i)$$

est un polynôme en x et $(\beta_i)_{1 \leq i \leq d}$ qui est symétrique en $(\beta_i)_{1 \leq i \leq d}$ et de degré total au plus

$$\epsilon_\lambda q M + qd(d-1)$$

en utilisant encore une fois les commentaires qui précèdent ce lemme. Ce qui complète la preuve de ce résultat.□

3.3 Preuve de l'hypothèse de Riemann pour les corps finis.

En utilisant le lemme 1.2, on conclut que le polynôme $r_\lambda(x)$ possède un zéro d'une multiplicité d'au moins M en chaque $v \in \mathcal{G}$. Posons

$$\mathcal{N}_\lambda = \sum_{v \in \mathcal{G}} |\mathcal{I}_\lambda(v)| \quad \lambda = 1, 2.$$

On a donc

$$d(q - d(d-1)) \leq \mathcal{N}_1 + \mathcal{N}_2 = d|\mathcal{G}| \leq dq.$$

Comme le nombre de zéros de $r_\lambda(x)$, comptés avec multiplicités, ne peut pas dépasser son degré, il s'ensuit que

$$M\mathcal{N}_\lambda \leq deg\ r_\lambda(x)$$

et que

$$\mathcal{N}_\lambda \leq \epsilon_\lambda q + \frac{qd(d-1)}{M}.$$

Maintenant, $\mathcal{N}_1$ est le nombre de zéros $(v, \xi) \in \mathbb{F}_q^2$ avec $\Delta(v) \neq 0$ de $f(x, y)$. Comme $deg\ \Delta(x) \leq d(d-1)$, il s'ensuit que si on pose $N := N_1$ alors on a

$$N \leq \mathcal{N}_1 + d^2(d-1) \leq q + \frac{qd(d-1)}{M} + d^2(d-1).$$

De même

$$N \geq \mathcal{N}_1 - d^2(d-1) - \mathcal{N}_2 \geq q - \frac{qd(d-1)}{M} - d^2(d-1).$$

On a donc

$$|N - q - 1| < \frac{qd(d-1)}{M} + d^3,$$

où M peut être n'importe quel entier qui satisfait les conditions du lemme 3.6. On choisit alors le multiple de d qui satisfait

$$\left(\frac{q}{2d}\right)^{1/2} - 5d < M \leq \left(\frac{q}{2d}\right)^{1/2} - 4d$$

et sous l'hypothèse $q > 250d^5$, on obtient

$$|N - q - 1| < (2d^5 q)^{1/2}.$$

En utilisant les commentaires du début du chapitre, on conclut que

$$|N - q - 1| \leq (d-1)(d-2)q^{1/2}$$

sans restriction. Le résultat est démontré.□

Le nombre d'applications de cette borne est astronomique. En effet, c'est un outil incontournable et fondamental pour étudier autant les sommes de caractères, le théorème de Burgess (voir [36]) et plusieurs autres questions importantes en théorie des nombres.

4 Finitude du nombre de solutions de l'équation de Thue.

4.1 Lemmes techniques.

Lemme de Siegel (Principe des nids de pigeons). *Soit α_{ij} une suite d'entiers non tous nuls avec $1 \leq i \leq m < n$ et $1 \leq j \leq n$. Supposons que $M \geq \max_{i,j} |a_{i,j}|$ et que $A \geq 0$ est un entier. Alors le système d'équations linéaires*

$$\sum_{j=1}^{n} \alpha_{ij} x_j = 0, \ 1 \leq i \leq m$$

possède au moins A solutions non triviales en entiers x_j avec $1 \leq j \leq n$ tels que

$$\max_j |x_j| \leq A^{1/(n-m)} (2nM)^{m/(n-m)}.$$

Preuve. Considérons l'application $\rho : [0, N]^n \to \mathbb{Z}^m$ définie par

$$\rho(v_1, \ldots, v_n) = \left(\sum_{j=1}^{n} \alpha_{1j} v_j, \ldots, \sum_{j=1}^{n} \alpha_{mj} v_j \right).$$

Clairement, l'image de ρ est contenue dans $[-nMN, nMN]^m$. On déduit que si

$$(N+1)^n > A(2nMN+1)^m, \tag{2}$$

alors il existe une valeur spéciale S de $[-nMN, nMN]^m$ qui apparaît au moins $A+1$ fois dans $\rho([0, N]^n)$. On note alors les vecteurs solutions de S par $V_1, \ldots, V_{A+1}$ et on obtient les A solutions recherchées avec $V_1 - V_{A+1}, \ldots, V_A - V_{A+1}$. On vérifie qu'il est suffisant d'avoir

$$N > A^{1/(n-m)} (2nM)^{m/(n-m)} - 1$$

pour satisfaire (2). Le résultat s'ensuit.□

Remarque. Si on pose $V := \{V_1, \ldots, V_{A+1}\}$, alors il est facile de voir que la preuve du dernier résultat nous fournit exactement $|V - V|$ solutions différentes avec cette borne. Il serait très intéressant de montrer que $|V - V|$ devient grand plus vite que linéairement en A au moins quand n est "grand".

Proposition 4.1. *Soit β un entier algébrique de degré $n \geq 2$ et α un nombre complexe quelconque. Supposons que $P(x)$ et $Q(x)$ sont des polynômes à coefficients entiers de degré inférieur à M tels que l'expression $P(x) - \beta Q(x)$ a un zéro d'ordre au moins D en $x = \beta$ avec $D \geq 2$. Si $M \leq Dn$, alors pour tout complexe ξ qui n'est pas un conjugué de β, l'ordre d'annulation G en $x = \xi$ de $P(x) - \alpha Q(x)$ est 0 ou au plus $2M - 2 - (D-1)n$.*

Preuve. Clairement, on peut écrire

$$P(x) - \beta Q(x) = (x - \beta)^D R_1(x)$$

ainsi que

$$P^{(1)}(x) - \beta Q^{(1)}(x) = (x - \beta)^{D-1} S_1(x)$$

pour certains polynômes $R_1(x)$ et $S_1(x)$. Comme $M \leq Dn$, il s'ensuit que $Q(x)$ est de degré plus grand que 0, car sinon on aurait que $x - \beta$ diviserait $P^{(1)}(x)$, d'où le fait que $Q(x)$ serait en fait 0 et donc on aurait que le polynôme minimal de β diviserait $P(x)$ avec une multiplicité d'au moins D, une contradiction. En éliminant β, on trouve

$$P(x)Q^{(1)}(x) - P^{(1)}(x)Q(x) = (x - \beta)^{D-1} T_1(x)$$

pour un certain polynôme $T_1(x)$. De la même façon on trouve

$$P(x) - \alpha Q(x) = (x - \xi)^G R_2(x),$$

$$P^{(1)}(x) - \alpha Q^{(1)}(x) = (x - \xi)^{G-1} S_2(x)$$

et

$$P(x)Q^{(1)}(x) - P^{(1)}(x)Q(x) = (x - \xi)^{G-1} T_2(x)$$

pour certains polynômes $R_2(x)$, $S_2(x)$ et $T_2(x)$. On déduit qu'il existe un polynôme $T(x)$ pour lequel on a

$$P(x)Q^{(1)}(x) - P^{(1)}(x)Q(x) = (x - \beta)^{D-1}(x - \xi)^{G-1} T(x).$$

Remarquons que $T(x)$ n'est pas identiquement nul, car sinon on aurait $P(x) = cQ(x)$ pour une certaine constante rationnelle c, cependant la première identité entraînerait que le polynôme minimal de β diviserait $Q(x)$ avec une multiplicité d'au moins D ce qui contredirait l'hypothèse selon laquelle $M \leq Dn$. Finalement, le membre de gauche est un polynôme à coefficients entiers de degré au plus $2M - 3$, il est donc divisible par le polynôme minimal de β. On déduit alors que $G - 1 \leq 2M - 3 - (D - 1)n$, d'où le résultat.□

Définitions. Soit γ un nombre irrationnel et $\frac{p}{q}$ une approximation de γ avec $(p, q) = 1$. La *qualité* de l'approximation de γ est le nombre $u > 0$ pour lequel

$$\left| \gamma - \frac{p}{q} \right| = \frac{1}{q^u};$$

si ce nombre n'existe pas, on pose $u = 0$. On définit alors la *mesure d'irrationalité* comme étant la limite supérieure prise sur toutes les approximations rationnelles de γ.

4.2 Preuve du théorème de Thue.

Théorème (Axel Thue). *Soit β un entier algébrique de degré $n > 2$. Alors la mesure d'irrationnalité de β est d'au plus $n/2 + 1$.*

Preuve. Sans perte de généralité, on peut toujours supposer que $\beta > 1$. Supposons, pour une contradiction, qu'il existe une infinité de nombres rationnels p_i/q_i pour lesquels nous avons

$$\left|\beta - \frac{p_i}{q_i}\right| < \frac{1}{q_i^{n/2+1+\epsilon}}$$

pour un certain $\epsilon > 0$. Nous allons alors choisir deux polynômes $P(x)$ et $Q(x)$ de degré inférieur à M, à coefficients entiers et pour lesquels l'expression $P(x) - \beta Q(x)$ a un zéro d'ordre au moins D en $x = \beta$. Pour ce faire, il suffit d'avoir

$$P^{[j]}(\beta) - \beta Q^{[j]}(\beta) = 0 \text{ pour } 0 \leq j < D.$$

Un tel système se réécrit en Dn équations à $2M$ inconnues. Supposons que le polynôme minimal de β est

$$\alpha_n x^n + \ldots + \alpha_1 x + \alpha_0$$

et que le maximum des coefficients est majoré par W. On observe alors que $\alpha_n^{m-n+1}\beta^m$ avec $m \geq n$ peut s'écrire en termes des nombres β^i avec $i < n$ en n'utilisant que des coefficients inférieurs à $2^{m-n}W^{m-n+1}$ en valeurs absolues. On déduit alors que le système à résoudre est à coefficients entiers majorés en valeurs absolues par

$$2^{M-n}(\alpha_n W)^{M-n+1} \max_{d<D} \binom{M}{d} < 2^{2M-n}(\alpha_n W)^{M-n+1}.$$

En utilisant le principe des nids de pigeons, on déduit qu'il existe une solution non triviale avec des coefficients qui satisfont

$$\lambda_{max} := \max_{0\leq i\leq 2M} |\lambda_i| < (2^{2M-n+2}(\alpha_n W)^{M-n+1} M)^{\frac{Dn}{2M-Dn}} = (2^n M(4\alpha_n W)^{M-n+1})^{\frac{Dn}{2M-Dn}}.$$

Soit $q_i \ll q_j$ deux grands nombres qui sont les dénominateurs d'une bonne approximation de β. Si on pose

$$\Lambda := \left| P\left(\frac{p_i}{q_i}\right) - \frac{p_j}{q_j} Q\left(\frac{p_i}{q_i}\right)\right|,$$

on remarque alors que

$$\begin{aligned}
\Lambda &\leq \left|P\left(\frac{p_i}{q_i}\right) - \beta Q\left(\frac{p_i}{q_i}\right)\right| + \left|\left(\beta - \frac{p_j}{q_j}\right) Q\left(\frac{p_i}{q_i}\right)\right| \\
&\leq \left|\beta - \frac{p_i}{q_i}\right|^D \lambda_{max} M \left(\frac{2p_i}{q_i}\right)^M + \left|\beta - \frac{p_j}{q_j}\right| \lambda_{max} \left(\frac{p_j}{q_j}\right)^M \\
&\leq \lambda_{max}(3\beta)^M \left(\left|\beta - \frac{p_i}{q_i}\right|^D + \left|\beta - \frac{p_j}{q_j}\right|\right),
\end{aligned}$$

où on a supposé que $|\beta - p_i/q_i| < 1/5$, que $|\beta - p_j/q_j| < 1$ et que $M \leq (5/4)^M$. Avec ces paramètres, on a

$$\lambda_{max} \leq (5\alpha_n W)^{\frac{MDn}{2M-Dn}}.$$

En supposant que Λ est non nul, il est facile de voir qu'il est supérieur à $1/q_i^M q_j$. On a donc

$$\frac{1}{q_i^M q_j} \leq (3\beta)^M (5\alpha_n W)^{\frac{MDn}{2M-Dn}} \left(\left| \beta - \frac{p_i}{q_i} \right|^D + \left| \beta - \frac{p_j}{q_j} \right| \right)$$

et dans le cas où Λ est nul, on applique la proposition précédente pour déduire qu'il existe une dérivée d'ordre G qui vaut au plus $2M - 2 - (D-1)n$ de $P(x) - p_j/q_j Q(x)$ qui ne s'annule pas en $x = p_i/q_i$, et ce, dans le cas où $M < Dn$. On pose donc

$$\Lambda' := \left| P^{[G]}\left(\frac{p_i}{q_i}\right) - \frac{p_j}{q_j} Q^{[G]}\left(\frac{p_i}{q_i}\right) \right|.$$

D'où

$$\begin{aligned}
\Lambda' &\leq \left| P^{[G]}\left(\frac{p_i}{q_i}\right) - \beta Q^{[G]}\left(\frac{p_i}{q_i}\right) \right| + \left| \left(\beta - \frac{p_j}{q_j}\right) Q^{[G]}\left(\frac{p_i}{q_i}\right) \right| \\
&\leq \left| \beta - \frac{p_i}{q_i} \right|^{D-G} \lambda_{max} M 2^M \left(\frac{p_i}{q_i}\right)^{M-G} + \left| \beta - \frac{p_j}{q_j} \right| \lambda_{max} 2^M \left(\frac{p_j}{q_j}\right)^{M-G} \\
&\leq \lambda_{max} 3^M \beta^{M-G} \left(\left| \beta - \frac{p_i}{q_i} \right|^{D-G} + \left| \beta - \frac{p_j}{q_j} \right| \right),
\end{aligned}$$

où on a supposé que $|\beta - p_i/q_i| < 1/5$, que $|\beta - p_j/q_j| < 1/4$ et que $M \leq (5/4)^M$. Cette fois, on a bien

$$\frac{1}{q_i^{M-G} q_j} \leq 3^M \beta^{M-G} (5\alpha_n W)^{\frac{MDn}{2M-Dn}} \left(\left| \beta - \frac{p_i}{q_i} \right|^{D-G} + \left| \beta - \frac{p_j}{q_j} \right| \right).$$

Nous avons le contrôle sur M et sur D, on peut donc s'assurer d'avoir $M \leq Dn$. En fait, on peut se rendre jusqu'à $D = (2/n - \delta)M$; le pire cas possible est lorsque G est très près de $2M - 2 - (D-1)n$ qui vaut au plus $\delta n M + n - 2 = \delta_1 n M$. Si on pose μ_i, μ_j pour les qualités d'approximation de β par q_i, q_j et que $q_j = q_i^a$, on peut écrire

$$\frac{1}{q_i^{M(1-\delta_1 n)+a}} < (15\alpha_n \beta W)^{D/\delta} \left(\frac{1}{q_i^{\mu_i((2/n-\delta-n\delta_1)M)}} + \frac{1}{q_i^{a\mu_j}} \right),$$

où encore

$$(3) \quad 1 < 2(15\alpha_n \beta W)^{D/\delta} \max \left(\frac{1}{q_i^{\mu_i((2/n-\delta-n\delta_1)M) - M(1-n\delta_1) - a}}, \frac{1}{q_i^{a(\mu_j - 1) - M(1-\delta_1 n)}} \right).$$

Pour bien voir ce qui se passe lorsque δ est fixé petit, M est fixé suffisament grand et $n \geq 3$. On suppose que q_i est très grand et que (p_i, q_i), $(p_j . q_j)$ sont des solutions à l'équation

$$|\alpha_n x^n + \alpha_{n-1} x^{n-1} y \ldots + \alpha_1 x y^{n-1} + \alpha_0 y^n| < |t|$$

pour un $|t|$ très petit par rapport à q_i. Dans ce cas, il est clair que μ_i et μ_j sont très près de n. En fait, si on pose

$$B := \min_{1 \le i < j \le n} |\beta_i - \beta_j|,$$

où les β_i sont les conjugués de β, alors on montre que

$$\left| \beta - \frac{p_i}{q_i} \right| < \frac{2^{n-1}|t|}{\alpha_n B^{n-1} q_i^n}$$

dès que $y > \max(2, [2/B] + 1)$. L'inégalité (3) nous fournit alors

$$0 > \mu_i((2/n - \delta - n\delta_1)M) - M(1 - n\delta_1) - a > (1 - \epsilon_1)M - a$$

et/ou

$$0 > a(\mu_j - 1) - M(1 - \delta_1 n) > a(n - 1 - \epsilon_2) - M(1 - \delta_1 n)$$

pour certains ϵ_1 et ϵ_2 aussi petit qu'on veut. Il n'est pas trop difficile de voir que ces deux énoncés sont en contradiction l'un avec l'autre, donc le mot "ou" doit être pris en considération. On en déduit que a est borné ou bien il est plus grand que $(1 - \epsilon_1)M$. Si a est borné, nous avons terminé. On suppose donc que a est plus grand que $(1 - \epsilon_1)M$. L'idée est maintenant de choisir une valeur de M un peu plus grande, disons $M' := 6/5M$ et de choisir D' de telle sorte que le δ' associé soit environs pareil à un petit facteur près de l'ordre de $1 + \epsilon_3$. Si q_i est suffisament grand, on a alors un nouveau minorant pour a qui est de l'ordre de $(1 - \epsilon_4)6/5M$. Le résultat s'ensuit en poursuivant ce processus.

Pour le cas général, on fait exactement le même genre de raisonnements. On remarque que l'argument fonctionne pour μ_i et μ_j de l'ordre de $n/2 + 1 + \epsilon_5 > n/2 + 1$; en effet dans ce cas, on se retrouve dans la même situation, i.e. nous avons un intervalle dans lequel a ne peut pas se trouver et on définit une suite de valeurs de (M_i, D_i) avec δ_i presque fixe pour recouvrir l'intervalle $[q, \infty)$. Un point important, c'est que D_i grandit moins vite que M_i et comme δ_i est presque fixe, le facteur exponentiel ne grossit pas assez vite pour causer des problèmes lorsque q est suffisament grand.

Le tout est parfaitement explicite. Le problème c'est que l'existence ou non de (p_i, q_i) n'est pas claire et qu'en général on ne peut donc pas conclure. On peut seulement montrer qu'une équation donnée ne peut pas avoir plus d'une solution avec $|t| < t_0$ pour $q > q_0$ explicitement calculable. On peut aussi utiliser ces méthodes pour borner le nombre total de solutions d'une équation fixée. En effet, pour les petites valeurs de q nous avons un "gap principle" universel, i.e. qui ne dépend pas de la nature algébrique de la quantité β considérée. Il suffit de remarquer que si

$$\left| \beta - \frac{p_1}{q_1} \right| < \frac{1}{q_1^\mu} \quad \text{et} \quad \left| \beta - \frac{p_2}{q_2} \right| < \frac{1}{q_2^\mu}$$

avec $q_1 < q_2$ et $p_1 q_2 \neq p_2 q_1$, alors on a

$$\frac{1}{q_1 q_2} \le \left| \frac{p_1}{q_1} - \frac{p_2}{q_2} \right| \le \left| \beta - \frac{p_1}{q_1} \right| + \left| \beta - \frac{p_2}{q_2} \right| < \frac{2}{q_1^\mu},$$

d'où on déduit que $q_2 > \dfrac{q_1^{\mu-1}}{2}$ et ensuite, pour les grandes valeurs de q, on sait déjà qu'on a au plus une solution.□

Références

[1] D. Barsky, B. Benzaghou, *Nombres de Bell et somme de factorielles*, J. de Théorie des Nombres de Bordeaux, 2004, 1-17.

[2] B. J. Birch, H. P. F. Swinnerton-Dyer, *Note on a problem of Chowla*, Acta Arithmetica, 1959, 417-423.

[3] R. Crandall, K. Dilcher, C. Pomerance, *A search for Wieferich and Wilson primes*, Math. of Computation, Volume 66, Number 217, January 1997, 433-449.

[4] E. Croot, *An outline of the Thue-Siegel theorem*, expository notes, 2007.

[5] E. Croot, *Stepanov's method for elliptic curves*, expository notes, 2007.

[6] K. Dilcher, L. Skula, *The cube of the Fermat quotient*, Electronic Journal of Combinatorial Number Theory **6**, 2006.

[7] H. Davenport, *Multiplicative number theory, Third edition*, Springer, 2000.

[8] D. S. Dummit, R. M. Foote, *Abstract Algebra*, Third Edition, John Wiley and Sons, 2004.

[9] P. Erdös, C. Pomerance, A. Sárközy, *On locally repeated values of certain arithmetic functions, IV*, The Ramanujan J., 1997, 227-241.

[10] K. Ford, *The distribution of integers with a divisor in a given interval*, Annals of Math., **168**, 2008, 367-433.

[11] K. Ford, *Integers with a divisor in* $(y, 2y]$, Anatomy of Integers, CRM Proceedings and Lecture notes **46**, AMS, 2008, 65-80.

[12] K. Ford, M. R. Khan, I. E. Shparlinski, C. L. Yankov, *On the maximal difference between an element and its inverse in residue rings*, Proceedings AMS **133**, 2005, 3463-3468.

[13] M. Z. Garaev, *Sums and products of sets and estimates of rational trigonometric sums in fields of prime order*, Preprint May 2009.

[14] M. Z. Garaev, *On uniform distribution modulo one*, Preprint, 2005.

[15] M. Z. Garaev, F. Luca, I. E. Shparlinski, *Exponential sums and congruences with factorials*, J. reine Angew Math., Band 584, 2005, 29-44.

[16] M. Z. Garaev, F. Luca, I. E. Shparlinski, *Character sums and congruences with* $n!$, Transactions AMS, Volume 356, Number 12, 2004, 5089-5102.

[17] A. Gertsch, A. M. Robert, *Some congruences concerning the Bell numbers*, Bull. Belgian Math. Soc. **3**, 1996, 467-475.

[18] A. Granville, *Additive combinatorics*, Lecture notes of winter, 2005.

[19] A. Granville, *The square of the Fermat quotient*, Electronic J. Combinatorial Number Theory **4**, 2004.

[20] A. Granville, Z. Rudnick, *Uniform distribution*, Springer Netherlands, Volume 237, 2007, 1-13.

[21] R. Heath-Brown, *An estimate for Heilbronn's exponential sum*, Analytic number theory. Vol. 2. Proc. of the International Conference in honor of Heini Halberstam held in Allerton Park, Illinois, May 16-20, 1995, 451-463.

[22] R. Heath-Brown, S. Konyagin, *New bounds for Gauss sums derived from k-th powers, and for Heilbronn's exponential sum*, Oxford J., Mathematics and Physical Sciences, Quaterly J. of Mathematics, Volume 51, Number 2, 2000, 221-235.

[23] A. Ivić, Ž. Mijajlović, *On Kurepa's problems in number theory*, Publications de l'Institut Mathématique; Nouvelle série, tome **57** (71), 1995, 19-28.

[24] A. Junod, *Congruences par l'analyse p-adique et le calcul symbolique*, Thèse de doctorat, Université de Neuchâtel, 2003.

[25] S. Konyagin, *Estimates for trigonometric sums over subgroups and for Gauss sums*, IV International Conference "Modern Problems of Number Theory and its Applications" : Current Problems, Part III (Russian) (Tula, 2001), Mosk. Gos. Univ. im. Lomonosova, Mekh.-Mat. Fak., Moscow, 2002, 86114.

[26] F. Luca, A. Mukhopadhyay, K. Srinivas, *On the Oppenheim's "factorisatio numerorum" function*, 2008.

[27] R. Meshulam, *An uncertainty inequality for finite abelian groups*, European J. of Combinatorics, Volume 27, Issue 1, January 2006, 63-67.

[28] C. J. Moreno, *Algebraic curves over finite fields*, Cambridge U. Press, Cambridge, 1991.

[29] P. D. Proinov, *On the Erdös-Turán inequality on uniform distribution. I*, Proc. Japan Acad., **64**, Ser. **A**, 1988.

[30] P. D. Proinov, *On the Erdös-Turán inequality on uniform distribution. II*, Proc. Japan Acad., **64**, Ser. **A**, 1988.

[31] W. M. Schmidt, *Equations over finite fields : an elementary approach*, Springer-Verlag, Berlin, 536, 1976.

[32] S. A. Stepanov, *Elementary method in the theory of congruences for a prime modulus*, Acta Arithmetica **17**, 1970, 231-247.

[33] K. B. Stolarsky, *Algebraic numbers and diophantine approximation*, Pure and applied Math., New York, 1974, 85-97.

[34] T. Tao, *An uncertainty principle for cyclic groups of prime order*, Mathematical Research Letters **12**, 2005, 121-127.

[35] H. Bing Yu, *Note on Heath-Brown's estimate for Heilbronn's exponential sum*, Proc. AMS, Volume 127, Number 7, 1999, 1995-1998.

[36] L. Zhao, *Burgess bound for character sums*, expository notes, 2007.