Construction des suites binaires pseudo-aléatoires

Shea Ming Oon

UFR S.T.M.I.A.
École Doctorale IAE + M
Université Henri Poincaré - Nancy I
D.F.D. Mathématiques

Thèse

présentée pour l'obtention du titre de

Docteur de l'Université Henri Poincaré, Nancy-I
en Mathématiques

par

OON Shea Ming

Construction des suites binaires pseudo-aléatoires

soutenue publiquement le 12 Juillet 2005

Membres du jury :

Christian MAUDUIT	Professeur à l'université Aix-Marseille 2 (rapporteur)
András SARKOZY	Professeur à l'université ELTE de Budapest (rapporteur)
Cécile DARTYGE	Maître de Conférence à l'université Henri Poincaré Nancy 1
Gérald TENENBAUM	Professeur à l'université Henri Poincaré Nancy 1 (président)
Jie WU	Chargé de Recherche CNRS à Nancy
Joël RIVAT	Professeur à l'université Aix-Marseille 2 (directeur de thèse)

Institut Élie Cartan Nancy, Laboratoire de Mathématiques, B. P. 239, 54506
Vandoeuvre-lès-Nancy

Remerciements

À l'issue de ce travail, je tiens à remercier en premier lieu mon directeur de thèse Joël Rivat. Grâce à ses connaissances en théorie de nombres, j'ai pu être bien initié dans ce domaine de recherche, plus particulièrement dans l'étude des suites pseudo-aléatoires. Outre ses compétences mathématiques, je voudrais le remercier également pour ses conseils en informatique qui m'ont permis d'améliorer ce travail.

Je suis sensible à l'honneur que me fait Gérald Tenenbaum en acceptant de présider ce jury. Merci à lui d'avoir bien voulu endosser cette responsabilité. Je lui suis reconnaissant aussi pour ses idées dans l'exploration de nouvelles problématiques.

De même, un grand merci à Christian Mauduit et András Sárközy pour avoir accepté d'être rapporteurs et de prendre soin de lire cette thèse. Leurs suggestions ont aussi été précieuses dans la progression de ma thèse.

Je tiens également à remercier Cécile Dartyge et Jie Wu pour leur disponibilité et de faire partie des membres du jury. Leur conseils durant quelques discussions occasionnelles ont été pareillement judicieux pour approfondir mes connaissances.

Je suis également reconnaissant à d'autres professeurs ou doctorants du laboratoire pour des discussions intéressantes. J'ai apprécié plus particulièrement celles avec Sylvain Col et Tom Krantz qui m'inspirèrent sur d'autres types de problèmes.

Je ne pourrais jamais assez remercier l'institut Élie Cartan de Nancy dans son ensemble. La direction m'a soutenu et fourni des conditions de travail excellentes. Les secrétaires m'ont simplifié beaucoup de tâches administratives. Les ingénieurs et techniciens m'ont aidé à résoudre de nombreux problèmes informatiques. Et les bibliothécaires sont toujours disponibles, efficaces et souriantes.

Les courts séjours à Bordeaux et à Marseille ont aussi été enrichissants pour mon apprentissage. Je remercie pour leur accueil ces institutions universitaires.

Enfin, je remercie le soutien de ma famille car sans eux la réussite de mes études n'aurait pas été possible.

Introduction

Le « hasard » est un sujet essentiel en cryptographie. Il n'est pas considéré comme un phénomène indésirable mais au contraire indispensable pour le développement de cette discipline. Son caractère non déterministe rend toute tentative de définition quelque peu arbitraire.

Le chapitre 1 est consacré à explorer en profondeur ce concept. Imaginons que le « hasard » est représenté par la sortie d'une suite depuis une « boîte noire ». Nous aborderons trois aspects importants. Premièrement, la répartition des éléments de cette suite : c'est le point de vue statistique qui est privilégié. La théorie des probabilités permet de construire des modèles afin de vérifier la fréquence d'apparition des motifs ou l'indépendance de ces éléments. Deuxièmement, la compressibilité de cette suite : c'est le point de vue algorithmique. La théorie de l'information permet de nous donner une idée sur la complexité requise pour une suite aléatoire. Troisièmement, l'imprévisibilité de la sortie de cette suite. Cela relève de l'essence même du « hasard », et est bien sûr crucial en cryptographie.

Après avoir exposé ces notions, nous explorons dans la suite du chapitre 1 quelques méthodes classiques pour obtenir le hasard en pratique. On peut soit se procurer des valeurs aléatoires par la voie naturelle qui consiste à observer des phénomènes physiques supposés aléatoires de manière intrinsèque, soit par des générateurs artificiels dits pseudo-aléatoires. Dans tous les cas, pour se convaincre qu'on obtient effectivement une bonne suite aléatoire, il est nécessaire de lui faire passer des tests statistiques. Cependant, il serait encore mieux de pouvoir prédire sa qualité avant sa génération. C'est la notion de test *a priori*, introduite par Knuth. C. Mauduit et A. Sárközy ont proposé deux mesures dans [**43**] qui seront notre sujet d'étude essentiel dans la suite. Elles sont définies pour les suites binaires à valeurs dans $\{-1,1\}$. Soit $E_N = (e_j)_{j\in[1,N]}$ avec $e_j \in \{-1,1\}$. Alors, la mesure de « *bonne distribution* » est

$$W(E_N) = \max_{a,b,t} \left| \sum_{j=0}^{t-1} e_{a+jb} \right|$$

où le maximum est pris pour tous $a, b, t \in \mathbb{N}$ tels que $1 \leqslant a \leqslant a + (t-1)b \leqslant N$. La mesure de « *corrélation d'ordre ℓ* » est

$$C_\ell(E_N) = \max_{M,D} \left| \sum_{n=1}^{M} e_{n+d_1} e_{n+d_2} \cdots e_{n+d_\ell} \right|$$

où le maximum est pris pour tous $M \in \mathbb{N}$ et $D = (d_1, \ldots, d_\ell) \in \mathbb{N}^\ell$ tels que $0 \leqslant d_1 < \cdots < d_\ell \leqslant N - M$.

J. Rivat et A. Sárközy [**54**] ont établi quelques liens sur ces mesures avec d'autres tests statistiques. Les tests statistiques usuels comme le test d'équidistribution ou test de série sont également présentés afin de mieux comprendre l'intérêt de ces mesures.

À la fin du chapitre 1, nous donnons quatre exemples de générateurs pseudo-aléatoires afin d'illustrer les différents aspect du hasard discutés précédemment. Ils sont le générateur à congruence linéaire, le générateur feed-back linéaire à registres décalés, le générateur Blum-Blum-Shub et le générateur RSA. Nous comparons également leurs avantages et inconvénients.

Le chapitre 2 présente quelques résultats connus sur les mesures W et C_ℓ. En premier lieu, nous pouvons nous demander quels sont les ordres de grandeurs pour ces deux mesures si l'on a affaire à une vraie suite aléatoire, dans l'optique de la théorie des probabilités. Nous rappelons les résultats des articles [**9, 2**], qui donnent les bons ordres de grandeur de $W(E_N)$ et $C_\ell(E_N)$, conformes aux valeurs attendues en vertu du théorème de la limite centrale.

Dans la suite du chapitre 2, nous passons en revue diverses constructions dont nous discutons l'intérêt et les techniques mises en œuvre pour les étudier, en apportant quelques améliorations ponctuelles. Les constructions connues de suites évaluées dans l'optique de Mauduit et Sárközy sont inspirées par des problèmes de théorie des nombres. La fonction de Liouville définie par $\lambda(n) = (-1)^{\Omega(n)}$ en est un exemple intéressant, en raison de son lien avec l'hypothèse de Riemann. On peut étendre cette construction à d'autres fonctions multiplicatives. La partie fractionnaire d'une suite réelle équirépartie modulo 1 peut aussi fournir des exemples intéressants. En revanche, les suites usuelles de type automatique ne fournissent jamais des exemples satisfaisants. La suite de Champernowne en base 2 qui consiste à concaténer les entiers naturels écrits en base 2 (en changeant 0 en -1) conduit à des valeurs trop grandes aussi bien pour la mesure de bonne distribution que pour la mesure de corrélation à l'ordre de 2.

Une classe de « bonnes » suites pseudo-aléatoires a été obtenue grâce aux caractères définis sur les corps finis. Les résultats importants utilisés dans cette étude sont des conséquences liées aux travaux d'André Weil sur les courbes algébriques et les variétés qui s'en déduisent. Nous présenterons les résultats principaux, avec ponctuellement des améliorations de certains lemmes utilisés dans les articles originaux sur les suites pseudoaléatoires. Ainsi nous pouvons obtenir une meilleure constante (2 au lieu de 9) dans la majoration suivante, fondamentale pour l'étude des sommes de caractères de Dirichlet (voir la Proposition 7) :

PROPOSITION 1. *Soient $p \geqslant 5$ un nombre premier, $\chi \neq \chi_0$ un caractère de Dirichlet d'ordre d sur $\mathbb{F}_p$ et $f \in \mathbb{F}_q[X]$ un polynôme tel que sa factorisation dans $\overline{\mathbb{F}}_p$ s'écrit $f(X) = a(X - x_1)^{d_1} \cdots (X - x_s)^{d_s}$ avec*

$$(d, d_1, d_2, \ldots, d_s) = 1.$$

Alors pour tout $(X, Y) \in \mathbb{R} \times \mathbb{R}^+$ avec $0 < Y \leqslant p$, on a

$$\left| \sum_{X < n \leqslant X+Y} \chi\left(f(n)\right) \right| \leqslant 2sp^{1/2} \log p.$$

La constante 2 ainsi obtenue n'est pas la meilleure possible, comme le suggère le test sur machine réalisé dans le chapitre 6. Néanmoins, cela permet d'obtenir des applications numériques intéressantes.

La première suite de ce type qui a été étudiée est la suite définie par le symbole de Legendre (cf. [**43**]) :

$$e_n = \left(\frac{n}{p}\right) \qquad (\text{pour } 0 < n < p)$$

où p est un nombre premier impair.

Il serait intéressant pour les applications dans la cryptographie d'arriver à généraliser cette construction. Dans l'article [**23**], L. Goubin, C. Mauduit et A. Sárközy ont notamment réussi à remplacer n par un polynôme $f(n)$ de degré k à racines simples dans la clôture algébrique de $\mathbb{Z}/p\mathbb{Z}$, avec d'autres hypothèses supplémentaires sur p, k et ℓ. Ceci donne une grande famille des suites :

$$e_n = \begin{cases} \left(\frac{f(n)}{p}\right) & \text{si} \quad p \nmid f(n), \\ 1 & \text{si} \quad p \mid f(n). \end{cases} \tag{1}$$

Dans le chapitre 3 nous abordons deux questions probabilistes posées par A. Sárközy, inspirées par l'étude de ces deux mesures W et C_ℓ : si $(\varepsilon_n)_{n\in\mathbb{N}}$ désigne une suite de variables aléatoires indépendantes suivant la loi de Bernoulli symétrique et si on pose $S_n = \sum_{n=1}^{m} \varepsilon_n$,

- est-il vrai que la quantité

$$\lim_{N\to\infty} \left(\max_{N<m<2N} \frac{|S_m|}{\sqrt{m}} \right)$$

est presque sûrement infinie ?

- Est-il vrai que la quantité

$$\limsup_{N\to\infty} \frac{\left|\sum_{n=1}^{N} \varepsilon_n \varepsilon_{n+1}\right|}{\sqrt{N}}$$

est presque sûrement infinie ?

Pour cette dernière question, nous avons été en mesure d'apporter une réponse complète sous une forme plus générale :

THÉORÈME 1. *Pour k fixé, soient $0 < d_1 < \cdots < d_k$ des entiers fixés et $(\varepsilon_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes suivant la loi de Bernoulli symétrique. Alors presque sûrement*

$$\limsup_{N\to\infty} \frac{\sum_{n=1}^{N} \varepsilon_{n+d_1} \cdots \varepsilon_{n+d_k}}{\sqrt{2N \log_2 N}} = 1,$$

et

$$\liminf_{N\to\infty} \frac{\sum_{n=1}^{N} \varepsilon_{n+d_1} \cdots \varepsilon_{n+d_k}}{\sqrt{2N \log_2 N}} = -1.$$

La première question est plus délicate. Il n'est pas évident que la limite existe donc on s'intéresse à la limite inférieure. Après quelques réflexions, nous avons été convaincu que même si on remplace le 2 par toute autre quantité finie, on n'aura pas une quantité presque sûrement infinie, sans pour autant formuler une démonstration valide. G. Tenenbaum a reformulé cette question après avoir pris connaissance de ce problème et a démontré que pour $f : \mathbb{N}^* \to \mathbb{R}^+$ une fonction croissante, la quantité

$$\liminf_{N\to\infty} \max_{N\leqslant n\leqslant f(N)N} \frac{|S_n|}{\sqrt{n}}$$

est presque sûrement finie si

$$\limsup_{N\to\infty} \frac{f(N)(\log_2 N)}{\log N} < \infty,$$

alors qu'elle est p.s. infinie si

$$\lim_{N\to\infty} \frac{\log f(N)}{\log_2 N} = \infty.$$

G. Tenenbaum nous a fait l'honneur d'autoriser la reproduction de sa démonstration dans le chapitre 3. La description complète de la limite inférieure en fonction de f reste encore une question ouverte, semble-t-il difficile, puisqu'il existe une plage de croissance de f pour laquelle on ne la connaît pas.

Le chapitre 4 traite de constructions inspirées par les sommes de caractères de Dirichlet. Une question naturelle consiste à se demander quand est-ce que deux suites définies par (1) coïncident. Nous apportons une réponse partielle dans le théorème suivant (cf. le Théorème 11) :

THÉORÈME 2. *Soit $p \geqslant 5$ un nombre premier, f, g deux polynômes à coefficients dans $(\mathbb{Z}/p\mathbb{Z})$ de degré respectif k et ℓ avec $k, \ell \leqslant \sqrt{p}$. Si les deux suites $\left(\frac{f(n)}{p}\right)$ et $\left(\frac{g(n)}{p}\right)$ ont $> 3(k+\ell)p^{1/2}\log p$ valeurs coïncidentes consécutives, alors ces deux suites sont identiques. De plus, si f et g sont unitaires et ne contiennent pas de facteur carré dans leur factorisation dans $(\mathbb{Z}/p\mathbb{Z})[X]$, alors $f = g$.*

Une autre possibilité de généralisation est de remplacer le symbole de Legendre par des caractères de Dirichlet. C'est une approche suggérée par E. Fouvry qui a proposé de poser

$$e_n = \begin{cases} +1 & \text{si } \mathfrak{Re}(\chi(n)) \geqslant 0, \\ -1 & \text{sinon.} \end{cases} \tag{2}$$

Nous avons étudié en détail cette construction, et nous présentons nos résultats dans le chapitre 4, dont certains sont en cours de publication dans le *Ramanujan Journal*.

Le résultat important est l'obtention d'une nouvelle grande famille de « bonnes » suites pseudo-aléatoires. L'essentiel est résumé dans le théorème suivant (voir le Théorème 12) :

THÉORÈME 3. *Soient $p \geqslant 2999$ un nombre premier, $\ell \geqslant 2$ un entier et $\chi \bmod p$ un caractère non principal d'ordre $d \geqslant 2$. Soit $E_p = (e_i)_{i\in[1,p]}$ la suite définie par (2). Alors, on a*

$$W(E_p) \leqslant 2p^{1/2}(\log p)^2 + 4p/d$$

et

$$C_\ell(E_p) \leqslant 3\ell p^{1/2}(\log 8p)^{\ell+1} + 4\ell p/d.$$

Lorsque $d > \sqrt{p}$, on peut supprimer les termes contenant p/d dans les deux estimations.

La restriction sur l'ordre d du caractère modulo p n'est pas contraignante, car presque tous les caractères multiplicatifs modulo p sont d'ordre $> \sqrt{p}$.

L'un des outils employés, outre les estimations sur les sommes des caractères, est une approximation par des polynômes trigonométriques due à Vaaler. Une généralisation de cette construction consiste à procéder comme dans le cas du symbole de Legendre, remplacer n par un polynôme $g(n)$. Nous donnons le théorème suivant pour clore ce chapitre 4 (voir le Théorème 13) :

THÉORÈME 4. *Soient $p \geqslant 251$ un nombre premier, $\chi \mod p$ un caractère non principal d'ordre d, $g \in (\mathbb{Z}/p\mathbb{Z})[X]$ un polynôme possédant $s \leqslant p^{1/2}$ racines distinctes dans la clôture algébrique de $\mathbb{Z}/p\mathbb{Z}$ et $E_p = (e_n)_{n\in[1,p]}$ la suite définie par*

$$e_n = \begin{cases} +1 & \text{si } \mathfrak{Re}(\chi(g(n))) \geqslant 0, \\ -1 & \text{sinon.} \end{cases}$$

Alors on a

$$W(E_p) \leqslant 4sp^{1/2}(\log p)^2 + 4p/d$$

et

$$C_2(E_p) \leqslant 5sp^{1/2}(\log 8p)^3 + 8p/d$$

où les termes p/d peuvent être omis lorsque $d > p^{1/2}$.

Si g est irréductible dans $(\mathbb{Z}/p\mathbb{Z})[X]$, alors pour $\ell \geqslant 3$, on a également

$$C_\ell(E_p) \ll \ell sp^{1/2}(\log 8p)^{\ell+1} + p/d.$$

Le chapitre 5 décrit l'étude d'une suite construite selon la répartition du plus grand facteur premier dans les progressions arithmétiques. Ces résultats ont été publié dans le journal *Periodica Mathematica Hungarica* (voir [**51**]). La suite considérée est définie par

$$e_n = \begin{cases} +1, & \text{si } P(n) \equiv +1 \bmod 4 \text{ ou } n = 2^k, \\ -1, & \text{si } P(n) \equiv -1 \bmod 4. \end{cases} \tag{3}$$

Nous pouvons obtenir certaines estimation sur la bonne distribution, avec ou sans condition supplémentaire, présentée dans les théorèmes suivants (voir les Théorèmes 14 et 15) :

THÉORÈME 5. *Soit $E_N = (e_i)_{i\in[1,N]}$ la suite définie par (3). Pour tout $A > 0$ fixé, on a*

$$W(E_N) \ll_A \frac{N}{(\log N)^A}.$$

THÉORÈME 6. *Soit $E_N = (e_i)_{i\in[1,N]}$ la suite définie par (3). S'il n y a aucun zéro de Siegel pour aucun des caractères réels, alors il existe une constante absolue $c > 0$ telle que*

$$W(E_N) \ll N\,\mathrm{e}^{-c(\log N \log_2 N)^{1/4}}.$$

La constante implicite est elle aussi absolue.

Les outils requis pour ces estimations de bonne distribution sont les résultats de Fouvry et Tenenbaum [**19**] sur les entiers friables : l'estimation de leur nombre et leur répartition en progression arithmétique. Malheureusement, la question de l'estimation de la corrélation semble encore hors de portée.

Le dernier chapitre présente quelques résultats numériques. Pour illustrer le chapitre 5, on propose de calculer les valeurs de bonne distribution et comparer avec les majorations issues de la théorie. On évalue également les valeurs sur la mesure de corrélation de petit ordre pour lesquelles on ne possède pas encore de résultat théorique. Ensuite, on établit un programme pour tester la longueur de la plus grande sous-suite pour laquelle on a une coïncidence totale comme il est décrit dans le Théorème 2.

Finalement, comme les suites pseudo-aléatoires constituent un enjeu important en cryptologie, nous croyons utile de développer en annexe quelques aspects importants de la cryptologie. Nous n'avons pas l'intention de donner une introduction globale sur cette science mais donner un survol sur les structures qui peuvent susciter une réflexion mathématique. L'accent est mis sur la cryptographie qui est la branche étudiant la réalisation des systèmes de cryptage et les techniques mathématiques impliquées.

Nous commençons par introduire les notions fondamentales comme le chiffrement et le déchiffrement, l'algorithme et les complexités associées, ainsi que les notions usuelles dans la cryptographie comme les fonctions à sens uniques, les fonctions de hachage, les générateurs pseudo-aléatoires. Ensuite, on approfondit l'étude des chiffrements en les séparant en deux classes : les chiffrements symétriques et asymétriques. Le chiffrement symétrique se distingue en deux sous-classes. Le chiffrement par bloc est le procédé largement utilisé pour chiffrer une grande quantité d'information. Tandis que le chiffrement par chaîne possède un lien étroit avec les générateurs pseudo-aléatoires qui étaient le sujet principal dans les chapitres précédents. Le chiffrement asymétrique est une idée récente qui fait intervenir les outils mathématiques plus fréquemment. Enfin, on expose plusieurs aspects de la cryptanalyse qui est le complémentaire de la cryptographie pour mieux comprendre le

fonctionnement de l'étude dans son ensemble. On termine par donner quelques exemples d'attaques spécifiques sur le chiffrement RSA.

Table des matières

Comment osons-nous parler des lois sur le hasard ?
Le hasard n'est-il pas l'antithèse de tous les lois ?
Bertrand Russell

CHAPITRE 1

Le hasard et ses générations

Le hasard fait partie de notre vie quotidienne, et on le sollicite plus souvent qu'on ne le pense. Les jeux de hasard constituent la première application que l'on peut citer dans le domaine du divertissement. Dans l'industrie ou pour réaliser des sondages, on a besoin d'échantillons choisis « au hasard » pour tester ou trouver un comportement représentatif. Un programme informatique nécessite lui aussi d'être testé avec des paramètres choisis au hasard. Les nombres aléatoires peuvent être utilisés pour réaliser une simulation d'un phénomène physique.

En cryptographie, le hasard est essentiel pour la confection de la clé. On demande souvent le choix d'un grand nombre premier, ou d'un polynôme irréductible dans un corps fini. D'une manière générale, on a besoin de nombres aléatoires, ou d'une suite aléatoire, comme pour la réalisation d'une clé du chiffrement « one-time-pad ». Pour la réalisation pratique en informatique, on considère des suites de nombres rationnels plutôt que des suites de nombres réels, et on se limite à des suites finies.

Bien qu'on puisse regarder ces considérations (la finitude ou la troncature) comme des contraintes pratiques inévitables, ces suites ne possèdent pas exactement les mêmes propriétés liées à l'étude du hasard que les suites infinies. On verra quelques exemples plus loin. En particulier, l'estimation quantitative est privilégiée par rapport à la description qualitative, surtout pour une suite finie. En revanche, on peut se limiter à l'intervalle $[0, 1]$ sans perte de généralité et ne considérer que la distribution uniforme car on dispose de bons outils pour obtenir d'autres loi de distribution à partir de celle-ci. Cependant, la considération sur les suites binaires n'est pas une entrave dans la pratique puisque du point de vue informatique, toute information n'est qu'une suite de 0 et de 1.

1. Qu'est-ce le hasard ?

C'est la première question à se poser dans cette étude. Pourtant, il n'est pas évident de donner une définition pour qualifier le hasard. N'est-il pas assez paradoxal de vouloir déterminer une définition qui est liée à une notion de nature non-déterministe ? Ce n'est

donc pas par hasard que Kolmogorov a établi la théorie des probabilités en esquivant cette question par le biais de la théorie de la mesure. Nous nous contentons ici de donner des conditions nécessaires les plus larges possibles pour discerner le caractère « aléatoire ».

1.1. La normalité. Une approche intéressante a été introduite par É. Borel en 1919 pour étudier les réels dits « normaux » : $x \in [0,1]$ est dit *normal* en base b si pour tout entier positif k, la fréquence d'apparition de tout motif $y_1 \cdots y_k$ $(0 \leqslant y_i < b)$ est la même, c'est-à-dire $1/b^k$. Il a montré que presque tous les réels (au sens de la mesure de Lebesgue) sont normaux en une base b donnée. On peut étendre ce concept (cf. **[32]** p.144) aux suites de réels en introduisant la notion de suite « ∞-distribuée » (selon la terminologie de Knuth, on dira que cette suite *complètement distribuée*). La notion de suite (m,k)-distribuée (pour m, k entiers positifs) peut aussi se faire : Une suite $(X_n) \in [0,1]^{\mathbb{N}}$ est (m,k)-distribuée si

$\forall i \in [\![0, m-1]\!], \forall j \in [\![1,k]\!], \forall u_j, v_j \in [0,1]$, on a

$$\begin{aligned} & \mathbb{P}(u_1 \leqslant X_{mn+i} < v_1, u_2 \leqslant X_{mn+i+1} < v_2, \ldots, u_k \leqslant X_{mn+i+k-1} < v_k) \\ = \; & (v_1 - u_1) \cdots (v_k - u_k). \end{aligned}$$

On retrouve la définition de la suite complètement distribuée qui est alors une suite $(1,k)$-distribuée pour tout $k \in \mathbb{N}$.

Le résultat suivant dû à I. Niven et H. S. Zuckerman montre (cf. **[50]** ou **[32]** p.149 Théorème C) l'importance de la distribution complète : Une suite complètement distribuée est (m,k)-distribuée pour tous entiers positifs m et k. Cette recherche d'uniformité dans la distribution conduit à la notion de discrépance. On se réfère à **[34]** pour une excellente discussion à ce sujet.

1.2. La discrépance. Pour une suite finie, la notion du hasard est encore plus délicate. Sur les suites binaires (constituées de 0 et de 1) finies d'une longueur donnée, la probabilité d'une suite constituée uniquement de 0 n'est-elle pas égale aux autres suites ? Une manière pertinente est de quantifier la « déviation à la normalité » par la *discrépance*, définie pour une suite finie $x_1, \ldots, x_N \in [0,1]$ par

$$D_N(x_1, \ldots, x_N) = \sup_I \left| \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_I(x_n) - \lambda(I) \right|$$

Il est facile de montrer (cf. par exemple **[49]**) que pour $0 \leqslant x_1 \leqslant x_2 \leqslant \cdots \leqslant x_N \leqslant 1$, on a

$$D_N(x_1, \ldots, x_N) = \frac{1}{N} + \max_{1 \leqslant j \leqslant N} \left(\frac{1}{N} - x_j \right) - \min_{1 \leqslant j \leqslant N} \left(\frac{1}{N} - x_j \right)$$

Donc $D_N(x_1, \ldots, x_N) \geqslant 1/N$ et le minimum est atteint lorsque $x_j = j/N$.

On notera la différence avec les suites infinies. W. M. Schmidt [**57**] a démontré qu'il existe une constante absolue $c > 0$ telle que pour toute suite infinie $S \in [0,1]^{\mathbb{N}}$, on ait $D_N(S) \geqslant cN^{-1} \log N$ (la discrépance étant prise sur les N premiers termes de S) On sait que $c = 0,12$ est une valeur admissible. À une constante près, les suites de Van der Corput atteignent cet infimum (cf. [**17**]). La suite de Van der Corput en base $b \geqslant 2$ est ainsi définie : Pour chaque $n \in \mathbb{N}$, à partir de l'écriture unique $n = \sum_{j \in \mathbb{N}} a_j(n) b^j$ avec $0 \leqslant a_j(n) < b$, on introduit la fonction inverse de radical $\phi_b(n) := \sum_{j \in \mathbb{N}} a_j(n) b^{-j-1}$ et la suite de Van der Corput en base $b \geqslant 2$ est $(\phi_b(n))_{n \in \mathbb{N}}$.

Une petite valeur de la discrépance ne suffit pas pour caractériser le hasard. Les termes de la suite de Van der Corput en base 2 se placent dans $[0, 1/2[$ et $[1/2, 1[$ alternativement. Selon la terminologie de Knuth, cette suite n'est pas 2-distribuée.

On peut généraliser la discrépance en dimension supérieure. L'intérêt de l'étude de la discrépance est surtout en dimension $\geqslant 2$ pour la méthode de Monte-Carlo dans le calcul des intégrales par exemple, car on peut exprimer le terme d'erreur par la discrépance. Cette généralisation s'avère utile également pour étudier les suites aléatoires.

1.3. Interlude sur la distribution. On a vu la définition des suites complètement distribuées, qui caractérise les propriétés statistiques intéressantes du hasard. Peut-on trouver une telle suite ? On sait que la réponse est affirmative si l'on ne s'occupe pas de l'aspect pratique de l'usage. Franklin [**20**] a démontré que pour presque tous les nombres transcendants $\theta > 1$, la suite $(\theta^n)_{n \in \mathbb{N}}$ est complètement distribuée. Levin [**37**] a montré ensuite comment choisir α pour un nombre transcendant θ donné de manière que la suite $(\alpha\theta^n)_{n \in \mathbb{N}}$ soit complètement distribuée. Knuth [**31**] avait donné lui-aussi un exemple concret en utilisant une expansion dyadique de fractions.

On se doute que la distribution complète n'est qu'une condition nécessaire dans la caractérisation du hasard. Ce point de vue de départ est néanmoins intéressant car on essaie d'appliquer certains théorèmes de la théorie des probabilités pour justifier le caractère aléatoire : on teste les suites obtenues par une certaine construction. C'est la moindre des choses que l'on doit faire avant d'accepter une quelconque suite trouvée. Cependant, il ne faut pas aller trop loin. Franklin avait proposé de définir qu'une suite est dite aléatoire si elle possède toutes les propriétés vérifiées par une suite des valeurs engendrées par une suite de variables aléatoires indépendantes (suivant la loi uniforme). Cette définition est

trop forte à tel point qu'il n'existe pas une telle suite si l'on interprète la définition d'une certaine manière : Une suite déjà construite explicitement n'est-elle pas vouée à ne pas vérifier certaines propriétés ?

Une autre réflexion porte sur les sous-suites. Ne doit-on pas demander que les sous-suites d'une suite aléatoire soient également complètement distribuées ? On se gardera bien de vouloir exiger que toutes ses sous-suites vérifient cette propriété car la définition sera trop forte comme précédemment. Par exemple, on peut construire une sous-suite monotone qui ne sera plus uniformément distribuée. Ceci est réalisable par un algorithme qui s'arrête si la suite initiale est complètement distribuée. On s'attend à se limiter aux algorithmes « effectifs ». Pour bien appliquer les algorithmes effectifs dans un processeur, il est plus judicieux d'utiliser les types entiers plutôt que les types flottants. Une méthode simple d'obtenir une suite (Y_n) d'entiers entre 0 et $b-1$ à partir d'une suite de réels $(X_n) \in [0,1[^{\mathbb{N}}$ est de poser $Y_n = \lfloor bX_n \rfloor$. C'est dans cette voie que Knuth a élaboré sa définition R6 (cf. [**32**] p.156), qui propose de qualifier « aléatoire » une suite (X_n) si pour tout entier $b \geqslant 2$ et tout algorithme effectif qui permet de déterminer une suite d'entiers strictement croissante de terme général (s_n) à partir des valeurs de $X_{s_0}, \ldots, X_{s_{n-1}}$, toute sous-suite construite par une « règle calculable » depuis $\lfloor bX_{s_n} \rfloor$ est 1-distribuée. Ensuite, il donne un exemple de la suite de rationnels de Wald (cf. [**32**] p.158 théorème W) qui satisfait cette définition pour montrer que cette définition n'est pas futile.

1.4. La complexité. Une autre caractérisation est la complexité de la suite. Cette voie a été explorée par A. N. Kolmogorov [**33**], puis suivi par P. Martin-Löf [**41**] et G. J. Chaitin [**10**]. Kolmogorov considère la complexité d'un objet définissable par un autre en utilisant le plus petit programme (en sa longueur) réalisable par une machine de Turing. Il propose de qualifier « aléatoire » tout objet de complexité maximale. Martin-Löf montre que ces objets possèdent les propriétés probabilistes vérifiées par les suites de variables aléatoires et passent tous les tests statistiques que l'on peut imaginer. Il étend cette définition aux suites infinies et montre que presque toutes les suites sont de complexité maximale. Chaitin propose de définir les suites « sans motif spécial » (patternless) dans le même esprit de maximiser le plus petit programme qui permet de construire cette suite.

On notera qu'on a remplacé la condition « passer tous les tests statistiques » par « tous les tests imaginables » par le biais d'une machine de Turing. Cette démarche de complexité se rapproche de la notion de compressibilité d'une suite. Une suite aléatoire ne doit pas être

compressible à grande échelle. Sinon, cela signifie qu'elle recèle trop de régularités, comme la périodicité, la répétition de motif par réflexion (palindrome), etc. Si l'idée initiale de Kolmogorov semble le pire pour la génération pratique des suites aléatoires, ce concept de compressibilité conduit à une autre voie dans les tests statistiques pour réfuter certaines catégories des suites bénéficiant d'une bonne distribution.

1.5. L'imprévisibilité. Dans ce qui précède, on a discuté la structure de la suite aléatoire, mais il ne faut pas oublier que l'imprévisibilité de la sortie des termes de celle-ci fait une partie indispensable du hasard. Certains phénomènes physiques ont été considérés comme aléatoires de nature et par conséquence non-déterministes de manière intrinsèque. Aléa devient un principe et non plus un effet secondaire indésirable. Il n'y a pas de contradiction à demander l'imprévisibilité sur une suite construite *a posteriori* car on se met ici à la place de l'adversaire et demande à prévoir le prochain terme après l'observation de plusieurs termes de la suite. Toutefois, nous pouvons nous demander s'il est possible de construire un générateur pseudo-aléatoire (voir sections 4.3 et 4.4) qui permet d'obtenir une suite imprévisible sans hypothèse supplémentaire. Ici, l'hypothèse sur la difficulté de certains problèmes mathématiques (pour l'existence de fonctions à sens unique) pour le générateur pseudo-aléatoire semble jouer le même rôle que le principe de l'incertitude dans la physique quantique.

1.6. Résumé. Faisons un petit résumé de la discussion ci-dessus. Le hasard se révèle en deux aspects :

- Être aléatoire dans sa structure. Une suite aléatoire doit être bien distribuée (riche en motifs) mais en même temps il doit être difficile de compresser son information (les motifs sont difficiles à analyser).
- Être imprévisible dans son évolution.

En pratique, il existe différents niveaux d'exigence. Une suite bien distribuée mais prévisible peut être utilisée dans la simulation informatique sans être utile dans la cryptographie (voir section 4.1).

2. Comment obtenir le hasard ?

2.1. Constructions naturelles. Si l'on se contente des suites binaires et qu'on admet qu'une pièce est sans biais, il n'est pas très commode de construire une suite uniformément distribuée en répétant le jet d'une pièce. En outre, on tombe dans un cercle vicieux pour

justifier que la pièce est sans biais. On peut envisager d'utiliser d'autres sources physiques présumées aléatoires, comme l'émission de particules lors d'une radiation, l'instabilité de la fréquence d'un oscillateur, ou le bruit produit dans un circuit électrique. Toutefois, ces sources ne suivent pas forcément la loi uniforme et pourraient donner des résultats biaisés ou corrélés. Alors, on emploie des techniques appropriées pour ôter ces anomalies avant de passer les tests. Cependant, c'est la lenteur de la génération des bits qui est le défaut principal. L'idéal est de générer rapidement en programmant sur un processeur. D'où vient l'idée du générateur de bits pseudo-aléatoires. Les sources naturelles ont néanmoins un intérêt pour l'initialisation de ces générateurs.

2.2. Générateur de bits pseudo-aléatoires. On appelle générateur de bits pseudo-aléatoires (GBPA) tout algorithme déterministe, qui à l'entrée reçoit une suite binaire (appelé la graine) d'une certaine longueur, et donne à la sortie une suite très longue (par rapport à l'entrée) qui a une allure « aléatoire » (appelé la suite pseudo-aléatoire). Ce n'est pas une définition précise qui reprend l'idée de la « simulation du hasard ». On réfère à **[38]** de M. Luby pour plus de détails. Lorsque les données considérées ne sont pas les suites binaires mais les suites en général, on l'appellera générateur (de suites) pseudo-aléatoire.

En pratique, on demande qu'un générateur de bits pseudo-aléatoires possède quelques propriétés supplémentaires. On dit qu'un GBPA *passe tous les tests statistiques en temps polynomial* si aucun algorithme en temps polynomial ne peut distinguer entre une vraie suite aléatoire et la suite pseudo-aléatoire ainsi générée avec une probabilité supérieure à $1/2$ de manière significative. On dit qu'un GBPA passe *le test du prochain bit* s'il n'existe pas d'algorithme en temps polynomial qui permet de prédire, connaissant les n premiers bits de la suite, le $(n+1)$-nième bit de cette suite avec une probabilité supérieure à $1/2$ de manière significative. En fait, Yao **[67]** a montré que ces deux définitions sont équivalentes.

Dans **[38]**, Luby incorpore en fait le test du prochain bit dans sa définition du GBPA. On l'appelle aussi GBPA *cryptographiquement sûr*. Il montre dans le même livre comment construire un GBPA à partir d'une fonction à sens unique. En fait, l'existence de la suite pseudo-aléatoire est aussi difficile à démontrer que l'existence de la fonction à sens unique, et on sait que cette dernière est plus difficile que le problème $\mathbf{NP} \neq \mathbf{P}$.

Au point de vue théorique, on n'a pas besoin d'exiger que la longueur de la suite pseudo-aléatoire à la sortie soit très grande. Un GBPA peut être appliqué récursivement pour obtenir une suite pseudo-aléatoire d'une longueur désirée (cf. Stretch Theorem dans **[38]**).

On ne va pas discuter cet aspect théorique mais donner quelques exemples des générateurs pseudo-aléatoires dans la suite.

3. Les tests statistiques

3.1. Tests selon le type. On sait qu'il vaut mieux se méfier de notre intuition pour juger si une suite est « assez aléatoire » ou non. Les tests statistiques sont des outils objectifs à cette fin. On a distingué plusieurs classes de tests. Dans la première, on recherche la « normalité » d'une suite finie. L'idée est l'emploi du théorème de la limite centrale ou de la loi des grands nombres. Le test de χ^2 a été introduit par K. Pearson pour mieux formuler le test sur une réponse probabiliste. Nous allons discuter dans cette section quelques tests fréquemment utilisés, comme le test d'équidistribution, le test de monotonie, le test de permutation d'ordres, le test de « Poker » et le test de série.

Une autre classe de tests s'oriente vers la compressibilité de la suite. Les algorithmes y ont un rôle prépondérants et ces tests prétendent en général pouvoir se substituer à la plupart des tests probabilistes. On peut citer parmi eux : le test de compressibilité de Lempel-Ziv, le test « universel » de Maurer et le test de complexité linéaire. Nous n'allons pas étudier cet aspect. En revanche, le test de complexité linéaire sera discutée brièvement dans la section 4.2.

Donnons une description simple du déroulement des tests statistiques. On se donne d'abord une hypothèse H (la loi supposée par l'observation) et un *niveau de confiance* $\alpha \in [0,001, 0,05]$. Supposons que X est une variable aléatoire qui suit la loi de χ^2, le seuil correspondant est x_α tel que $\mathbb{P}(X > x_\alpha) = \alpha$. Si X_O est la valeur observée dans l'expérience qui est supposée de même loi que X (l'hypothèse H), On rejette l'hypothèse H si $X_O > x_\alpha$ en disant que la proportion des suites aléatoires qui suivent la même loi mais se comportant ainsi est $\leqslant \alpha$. Voir aussi **[35]** pour une introduction générale et détaillée à ce sujet. On peut trouver également sur le site web de NIST certains logiciels et la description des tests les plus fréquemment utilisés.

3.2. Tests selon la méthode. Toutes les méthodes de tests ci-dessus sont faites pour les suites déjà construites. On peut dire que ce sont les tests *a posteriori*. Un autre point de vue dans l'étude des suites pseudo-aléatoires est de faire les tests *a priori*, c'est-à-dire de donner une estimation de la valeur de ces tests en fonction des paramètres du générateur puisque l'algorithme du générateur est connu d'avance. On les nomme aussi les

tests théoriques par opposition aux tests empiriques précédents qui peuvent être appliqués à n'importe quelle suite. C'est ici que la théorie des nombres joue son rôle.

C. Mauduit et A. Sárközy ont proposé plusieurs mesures dans l'étude des propriétés pseudo-aléatoires dans [**43**]. Les deux principales mesures sont *la mesure de bonne distribution* et *la mesure de corrélation d'ordre* ℓ ($\ell \geqslant 2$ entier) qui sont définies pour une suite binaire à valeurs dans $\{-1, 1\}$, ce qui ne change rien d'essentiel par rapport aux suites binaires usuelles (à valeurs dans $\{0, 1\}$). Soit $E_N = (e_j)_{j=1,\dots,N}$ où $e_j \in \{-1, 1\}$. La mesure de la bonne distribution est

$$W(E_N) = \max_{a,b,t} \left| \sum_{j=0}^{t-1} e_{a+jb} \right|$$

où le maximum est pris pour tous $a, b, t \in \mathbb{N}$ tels que $1 \leqslant a \leqslant a + (t-1)b \leqslant N$. La mesure de corrélation d'ordre ℓ est

$$C_\ell(E_N) = \max_{M,D} \left| \sum_{n=1}^{M} e_{n+d_1} e_{n+d_2} \cdots e_{n+d_\ell} \right|$$

où le maximum est pris pour tous $M \in \mathbb{N}$ et $D = (d_1, \dots, d_\ell) \in \mathbb{N}^\ell$ tels que $0 \leqslant d_1 < \cdots < d_\ell \leqslant N - M$. J. Rivat et A. Sárközy ont donné quelques majorations de certains tests statistiques par ces deux mesures dans l'article [**54**]. C'est sur ces mesures qu'on développe les sujets d'études dans les chapitres suivants.

3.3. Quelques tests de normalité. Pour effectuer ces tests, le nombre de termes de la suite doit dépasser une certaine longueur (dépendant du test).

3.3.1. *Test d'équidistribution.* On teste si la suite est uniformément répartie dans $[0, 1]$. Pour une suite de réels, c'est la discrépance qu'on mesure ici. Au point de vue informatique, on ne considère que des rationnels après une certaine approximation ce qui nous permet de procéder à un test de χ^2. Dans le cas binaire, ce test, appelé également le *test de fréquence* , se réduit au comptage du nombre de 0 et de 1. Une suite aléatoire de longueur N doit contenir une quantité proche de $N/2$ de 0 et de 1. Si l'on note n_0 (resp. n_1) le nombre de 0 (resp. 1) et $X = \frac{(n_0 - n_1)^2}{N}$, X doit suivre la loi de χ^2 de degré 1. C'est le premier test à effectuer avant de chercher une autre anomalie de la suite.

3.3.2. *Test de monotonie.* (« Run test »)

Pour une suite de réels dans $[0, 1]$, on considère les nombres de termes consécutifs formant une sous-suite (finie) monotone. Il existe une petite difficulté technique du fait que ces quantités ne sont pas indépendantes qu'on ne peut pas appliquer directement un

test de χ^2 (cf. **[32]** p.65 pour plus de détails). Dans le cas de la suite binaire, il suffit de compter le nombre de termes dans un bloc de 1 et de 0 alternativement. Le test est alors plus simple.

3.3.3. *Test de permutation d'ordres.* On se donne une suite de réels dans $[0, 1]$ et un entier $t \geqslant 2$. Pour les t termes consécutifs de la suite, il y a $t!$ possibilités d'ordre à considérer. On forme la suite en paquets de t termes successifs puis compte le nombre de motif de ces ordres. On peut appliquer un test de χ^2 avec la probabilité $1/t!$ pour chaque ordre. La comparaison du motif peut se faire rapidement par un autre algorithme. Si la suite est binaire, ce test peut être considéré comme un cas particulier du test de Poker.

3.3.4. *Test de Poker.* C'est un test de combinaison sur les suite des entiers (par exemple). On suppose que le cardinal de l'ensemble des entiers est petit et on regroupe par paquet de t termes successifs de la suite, puis compare le nombre de chaque combinaison possible. Le nom du test était désigné pour le cas $t = 5$. Pour simplifier l'algorithme de comparaison, on peut modifier le test en ne tenant pas en compte le modèle des entiers considérés (et pas les combinaisons qui en résultent). Par exemple, supposons que $t = 5$ et le cardinal de l'ensemble des entiers est $\geqslant t$, alors il y a sept motifs :

$$aaaaa; aaaab; aaabb; aaabc; aabbc; aabcd; abcde.$$

Ces comptages permettent de réaliser un test de χ^2 direct. Sur une suite binaire, il vaut mieux considérer les permutations, pour ne pas trop réduire le nombre de motifs.

3.3.5. *Test de série.* On se donne un entier $s \geqslant 2$ et on considère les points $\mathbf{x_n} = (x_n, x_{n+1}, \ldots x_{n+s-1})$ puis on étudie la distribution de ces points en dimension s. C'est un test sur la dépendance éventuelle des terme adjacents de la suite. Niederreiter s'est intéressé à la discrépance (voir **[48]**) de ces points générés par le générateur à congruence linéaire (voir 4.1). On peut aussi faire quelques tests de χ^2. En notant $\mathbf{X_n}$ les vecteurs aléatoires correspondant, on peut soit se restreindre à s familles (notées par $i = 0, \ldots, s-1$) de vecteurs $\mathbf{Y_{i_j}} := \mathbf{X_{js+i}}$ pour éviter la dépendance des vecteurs $\mathbf{X_n}$ causée par le chevauchement, soit reprendre les valeurs statistiques sur $\mathbf{X_n}$ mais faire certaines différentiations discrètes pour éliminer la dépendance, afin de réaliser un test de χ^2 valide (cf. **[22]** et **[40]**).

4. Les générateurs pseudo-aléatoires

On a compris que la méthode la plus rapide pour obtenir une suite « aléatoire » est par la voie algorithmique. Von Neumann avait proposé la méthode du « carré au milieu » :

Supposons qu'on veut créer une suite de dix chiffres. On choisit un nombre de dix chiffres au hasard puis on calcule son carré et choisit les dix chiffres au milieu de ce nombre. On réitère ce procédé. Le défaut de cette méthode est que la suite n'est pas très « aléatoire », surtout lorsqu'un zéro apparaît au milieu car il va continuer à se propager dans la suite.

Knuth (cf.[**32**] p.4) a présenté un autre algorithme ad hoc qui a été écrit avec des opérations déroutantes afin de brouiller les pistes à toute analyse. Ce générateur est destiné à produire une large famille de suites pseudo-aléatoires mais on remarque aussitôt qu'on obtient très vite une seule suite périodique à partir de certain rang. La moralité de ces anecdotes est que l'algorithme ne doit pas être choisi au hasard.

Cependant, on trouve encore des générateurs non sûrs qui continuent à être utilisé, comme l'algorithme ANSI X9.17 (cf. par exemple [**47**] p.173) approuvé par FIPS (Federal Information Processing Standard) dans la confection de clés pour **DES**, car il est plus rapide qu'un algorithme cryptographiquement sûr. On ne discutera pas de ces générateurs là. Les quatre exemples qui suivent sont le générateur à congruence linéaire, le générateur feed-back linéaire à registres décalés, le générateur BBS (Blum-Blum-Shub) et le générateur RSA. A part le premier, les trois autres sont des GBPA.

4.1. Le générateur à congruence linéaire. C'est un générateur de suites d'entiers compris entre 0 et m pour un entier positif m donné. Par une division par m, on obtient une suite des rationnels de $[0,1]$ (avec le dénominateur commun m). La graine est un 4-uplet (x_0, m, a, c) avec $m \in \mathbb{N}^*$ et $0 \leqslant x_0, a, c < m$. On pose $\forall n \in \mathbb{N}$,

$$x_{n+1} \equiv ax_n + c \bmod m. \tag{4}$$

Cette méthode de construction a été proposée par D. H. Lehmer [**36**]. Elle est algorithmiquement simple et permet d'obtenir une suite pseudo-aléatoire intéressante si la graine est bien choisie. Pour des raisons pratiques, on choisit habituellement $m = 2^N$ ou 10^N. L'entier m doit être grand car la suite est toujours périodique (à partir de certain rang) de période $\leqslant m$. Le théorème suivant n'est pas difficile à établir (cf. [**29**] ou [**32**] p.16), et il précise les cas où la période maximum est atteinte :

THÉORÈME 7. *La suite définie par (4) a pour période minimale m si et seulement si les trois conditions suivantes sont remplies :*

- $(c, m) = 1$;
- *Pour tout nombre premier $p|m$, on a $p|(a-1)$;*
- *Si $4|m$, alors $4|(a-1)$.*

Ainsi décrit, le générateur à congruence linéaire est un bon candidat pour engendrer une large famille des suites pseudo-aléatoires. On sait que ce générateur passe a priori le test d'équidistribution avec peu de contraintes sur la graine (cf. [**48**]). D'autres études intéressantes comme la discrépance sont aussi faite dans le même article. R. R. Coveyou et R. D. MacPherson ont introduit le *test spectral* pour ce générateur. Dans le livre de Knuth [**32**] chapitre 3.3.4, on considère les points $\mathbf{x_n} = (x_n, x_{n+1}, \dots x_{n+s-1})$ comme dans le test de série et on remarque que ces points forment un réseau dans l'espace. G. Marsaglia [**39**] a calculé la distance maximale des hyperplans (voisins) qui permettent de couvrir tous ces points. C'est une interprétation géométrique du test spectral. Cette régularité de la répartition des points met en doute également le caractère aléatoire supposé de ces suites, mais il ne faut pas oublier que ces points sont corrélés et la suite reste acceptable si la distance des hyperplans n'est pas trop grande. Niederreiter a montré que le générateur passe le test de série s'il passe le test spectral et si une autre condition technique est satisfaite. (cf. [**32**] théorème N p.109).

Ce générateur possède des propriétés statistiques intéressantes mais il n'est pas un générateur cryptographiquement sûr : la suite générée est prévisible en observant une petite partie de la suite. Plumstead [**52**] a présenté une méthode efficace pour retrouver a, m et c. Considérons la suite $y_n = x_n - x_{n-1}$, on a $y_{n+1} = ay_n \bmod m$. Supposons qu'on a pu observer $t+2$ termes successifs de x_n et on trouve que le pgcd des $y_n, \dots, y_{n+t}$, noté d, divise y_{n+t+1}. On sait trouver les $u_1, \dots, u_t \in \mathbb{Z}$ tels que $d = \sum_{i=1}^{t} u_i y_{n+i}$. En multipliant par a, on voit que a est probablement égal à $\sum_{i=1}^{t} u_i \frac{y_{n+i+1}}{d}$. Une fois que a est trouvé, on cherche ensuite m encore par une considération du pgcd en utilisant a et y_n. On en déduit c en revenant à la suite (x_i).

Ceci met en lumière qu'une suite pseudo-aléatoire peut être utile dans une application (comme la simulation) mais pas dans une autre (comme dans la cryptographie). On peut penser à généraliser cette méthode et à rendre plus compliquée la génération de la suite. Par exemple, on peut poser

$$x_{n+2} \equiv ax_{n+1}^2 + bx_n + c \bmod m$$

puisqu'une fonction linéaire est peu « aléatoire » ; ou encore une généralisation aux équations homogènes due à Tausworthe [**61**] :

$$x_{n+s} \equiv a_{s-1}x_{n+s-1} + \cdots a_0 x_n \bmod m. \tag{5}$$

Cependant, Boyar [**5**] a montré que ces deux méthodes ne sont pas cryptographiquement sures en suivant la même démarche que Plumstead.

4.2. Le générateur feed-back linéaire à registres décalés. Les suites pseudo-aléatoires construites par (5) possèdent une structure beaucoup plus riche et font le sujet d'études de nombreux articles (cf. [**48**]). Lorsque $m = 2$, le générateur est appelé le générateur feed-back linéaire à registres décalés. Ainsi dénommé, ce générateur atteste la facilité de l'implantation sur un processeur : les x_i défilent sur les registres du processeur à chaque étape. De plus, on obtient une suite bien distribuée si les a_i sont bien choisis. Par exemple, pour $s = 4$, $a_0 = a_1 = 1$, $a_2 = a_3 = 0$, et les valeurs initiales $x_0 = x_1 = x_2 = x_3 = 1$, on obtient une suite de période 15. C'est la période maximale qu'on peut espérer.

Dans le cas général, la période maximale est $2^s - 1$ car cela signifie que le s-uplet $(x_i, \ldots, x_{i+s-1})$ qui défile a pris tous les motifs possibles sauf $(0, \ldots, 0)$. On sait comment obtenir une telle suite (remarquons que toutes les suites de période maximale sont semblables à une translation près, puisque tout motif doit y figurer sur le défilement). Exprimons la formule de récurrence par une matrice :

$$\begin{pmatrix} x_{n+1} \\ x_{n+2} \\ \vdots \\ \vdots \\ \vdots \\ x_{n+s} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix} \begin{pmatrix} x_n \\ x_{n+1} \\ \vdots \\ \vdots \\ \vdots \\ x_{n+s-1} \end{pmatrix}$$

$$= U \begin{pmatrix} x_n \\ x_{n+1} \\ \vdots \\ x_{n+s-1} \end{pmatrix}$$

$$= U^{n+1} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{s-1} \end{pmatrix}$$

Le polynôme caractéristique de la matrice U, à un signe près, est

$$P(X) = X^s - a_{s-1}X^{s-1} - \cdots - a_0. \tag{6}$$

Si P est irréductible dans $(\mathbb{Z}/2\mathbb{Z})[X]$, alors P définit sur $(\mathbb{Z}/2\mathbb{Z})[X]$ par passage au quotient, un corps d'extension à 2^s éléments dans lequel il se factorise complètement en

racines simples. La période de la suite (x_n) est égale à l'ordre de U, soit l'ordre de n'importe quelle racine de P. On appelle *polynôme primitif* tout polynôme minimal d'un élément primitif. Avec un raisonnement plus abouti, on peut montrer que la suite (x_n) définie par (5) est de période $2^s - 1$ si et seulement si le polynôme caractéristique associé P défini par (6) est primitif.

On obtient encore une méthode simple pour générer des suites pseudo-aléatoires longue qui possèdent des propriétés statistiques intéressantes, mais on se rend compte rapidement que la suite n'est pas imprévisible. En effet, supposons qu'on puisse estimer la valeur s, alors il nous suffit de disposer $2s$ termes consécutifs de la suite pour retrouver les valeurs a_i en résolvant les s équations linéaires à s inconnus.

Cette observation nous dit également que la suite contient trop peu d'information. Elle peut être très bien compressée : s valeurs suffisent pour déterminer $2^s - 1$ valeurs par une récurrence linéaire. Le degré s du polynôme caractéristique correspondant P est appelé aussi *la complexité linéaire* de la suite. Une exigence naturelle d'une suite aléatoire est qu'elle possède une complexité linéaire grande (comparable à sa longueur). Pour calculer la complexité linéaire d'une suite binaire quelconque, on dispose de l'algorithme *Berlekamp-Massey* (cf. par exemple [**47**] p.200).

4.3. Le générateur de BBS. Les générateurs qu'on a vu jusqu'à présent donnent des suites prévisibles en observant une petite portion de leur sortie, donc ils ne sont pas cryptographiquement sûrs. Voici un générateur proposé par L. Blum, M. Blum et M. Shub dans [**4**] qui passe le test du prochain bit sous l'hypothèse que le problème **RQ** est difficile (voir la section 1.2 dans l'annexe sur « la cryptologie »).

Choisissons deux nombres p et q congrus à 3 modulo 4. Comme -1 n'est pas un résidu quadratique modulo p, l'application $x \mapsto x^2$ est une bijection sur l'ensemble des résidus quadratiques modulo p. On a la même conclusion pour q. Posons $n = pq$, d'après le théorème chinois, on déduit que l'application $x \mapsto x^2$ est une bijection sur l'ensemble des résidus quadratiques modulo n. Notons f son inverse. On se donne un résidu quadratique modulo n quelconque a_0. On définit la suite $(a_m)_{m\in\mathbb{N}}$ par

$$\forall m \in \mathbb{N}^*, \quad a_m = f(a_{m-1})$$

(f est calculable puisqu'on connaît p et q). On note enfin x_m le dernier bit de a_m correspondant, c'est-à-dire $x_m = \mathrm{Par}(a_m)$, la fonction de la parité ou le reste de la division par

2. La suite (x_m) est appelée la suite pseudo-aléatoire de Blum-Blum-Shub (avec la graine (p, q, a_0)), abrégé en BBS.

Nous allons voir que si une telle suite est prévisible, le problème de **RQ** sera résolu avec autant d'effort. Supposons qu'on dispose d'un oracle qui à l'entrée reçoit n ainsi que les k bits consécutifs $x_m, x_{m+1}, \ldots, x_{m+k-1}$, et fournit à la sortie la valeur x_{m+k} avec une probabilité $\geqslant 1/2 + \varepsilon$. Prenons maintenant un $\alpha \in \mathbb{Z}/n\mathbb{Z}$ vérifiant $\left(\frac{\alpha}{n}\right) = 1$ (symbole de Jacobi) et on souhaite déterminer si α est un résidu quadratique ou non. On pose alors $x_{m+k-1} = \mathrm{Par}(\alpha^2)$. On remarque que $x_{m+k} = \mathrm{Par}(\alpha)$ si et seulement si α est un résidu quadratique. On calcule alors les carrés successifs du α^2 et on pose $x_{m+i} = \mathrm{Par}(\alpha^{2(k-i)})$ (la connaissance de n suffit pour le calcul). On demande ensuite à l'oracle de déterminer la valeur x_{m+k} avec une probabilité $\geqslant 1/2+\varepsilon$, ce qui détermine si α est un résidu quadratique ou non avec la même probabilité.

Il est clair qu'en pratique on peut générer la suite en calculant d'abord les carrés puis inverser le sens de la suite pour retrouver le sens usuel en Occident. Une autre façon d'économiser le temps de calcul est de remarquer les choix spécifiques sur p et q donnent alors $f(a) = a^{\frac{(p-1)(q-1)+4}{8}}$. On obtient finalement une garantie de l'imprévisibilité de la suite sous hypothèse de la difficulté du problème **RQ**. On paie aussi le prix pour la lenteur de la génération de la suite.

4.4. Le générateur de RSA. Dans le chiffrement de RSA, déterminer la parité du message m est aussi difficile que de trouver m lui-même. Cette constatation permet de créer un générateur du même genre que celui de BBS.

Reprenons les notations dans le chiffrement RSA. On pose $n = pq$ et on choisit e au hasard tel que $1 < e < \phi(n)$. On choisit $m_0 \in \mathbb{Z}/n\mathbb{Z}$ au hasard et on construit la suite m_i en posant

$$\forall i \in \mathbb{N}^*,\ m_i = m_{i-1}^e \bmod n.$$

On définit $x_i = \mathrm{Par}(m_i)$ comme précédemment. Alors, (x_i) est une suite binaire imprévisible sous hypothèse que le problème **Fct** est difficile. En fait, les Vazirani [**65**] ont également démontré que la sécurité de la suite de BBS s'obtient sous cette hypothèse, ce qui améliore le résultat de [**4**].

CHAPITRE 2

Des suites pseudo-aléatoires

Nous avons vu de nombreuses manières pour qualifier une suite de pseudo-aléatoire. Ici, on se limitera à l'étude de deux propriétés pseudo-aléatoires proposées par C. Mauduit et A. Sárközy décrites pour la première fois dans l'article [**43**]. Pour une suite binaire finie $E_N = (e_n)_{n\in[1,N]} \in \{-1,+1\}^N$, on définit la mesure de bonne distribution par

$$W(E_N) = \max_{a,b,t} \left| \sum_{j=0}^{t-1} e_{a+jb} \right|$$

où le maximum est pris pour tous $a, b, t \in \mathbb{N}^*$ tels que $1 \leqslant a+b \leqslant a+(t-1)b \leqslant N$. Parallèlement, la mesure de corrélation d'ordre ℓ est

$$C_\ell(E_N) = \max_{M,D} \left| \sum_{n=0}^{M} e_{n+d_1} e_{n+d_2} \cdots e_{n+d_\ell} \right|$$

où le maximum est pris pour tous $M \in \mathbb{N}$ et $D = (d_1, \ldots, d_\ell) \in \mathbb{N}^\ell$ tels que $0 < d_1 < \cdots < d_\ell \leqslant N - M$.

Ainsi, au lieu de chercher des irrégularités sur toutes les sous-suites possibles, la mesure de bonne distribution examine les sous-suites en progression arithmétique. Pour une suite binaire infinie, la notion des suites complètement distribuées correspond à celle des suites ayant des « petites valeurs » sur la mesure de corrélation (pour un tronçon fini) à tout ordre ℓ. Tandis que pour une suite finie, l'estimation ne se fait qu'à un sens : Pour $0 < k < N$, en notant

$$N_k(E_N) := \max_{X\in\{-1,+1\}^k} \max_{0<M\leqslant N-k} \left| \sharp\{n \in [0, M] : (e_{n+1}, \ldots, e_{n+k}) = X\} - M/2^k \right|,$$

on a (cf. Prop. 1 de [**43**]) :

$$N_k(E_N) \leqslant \max_{1\leqslant\ell\leqslant k} |C_\ell(E_N)|.$$

En revanche, on peut obtenir d'autres informations via les corrélations comme par exemple l'apparition de motifs. La complexité $f(k, E_N)$ d'une suite E_N est le nombre de différents motifs de longueur k qui apparaissent dans cette suite. La complexité maximale est alors $f(k, E_N) = 2^k$ (pour $k \leqslant (\log N)/(\log 2)$). Une suite pseudo-aléatoire doit avoir

la complexité maximale, au moins pour les petites valeurs de k. Le Théorème 6 dans [**7**] donne une condition suffisante lorsque $k \leqslant N/2$ pour atteindre la complexité maximale : Il suffit que

$$\forall \ell \in [1, k], \quad C_\ell(E_N) \leqslant \frac{N}{2^{k+1}}.$$

On cherche donc a priori des suites ayant des « petites valeurs » pour la mesure de bonne distribution et la mesure de corrélation d'ordre ℓ (au moins pour les petits ℓ). On remarque qu'il est possible de fabriquer une suite ayant $W(E_N)$ petite mais $C_2(E_N)$ grande : Il suffit de faire une concaténation sur elle-même. (On pose $e_{n+N} = e_n$ pour $n \leqslant N$, alors $W(E_{2N})$ reste petite mais $C_2(E_{2N}) \geqslant N$.) Ainsi, il est naturel de demander que les deux mesures soient petites pour qualifier une suite pseudo-aléatoire de « bonne ». Quels sont les ordres de grandeur auxquels on peut s'attendre ? Les Théorèmes 1 et 2 dans [**9**] nous donnent une idée : pour tout $\varepsilon \in]0, 1[$, et tout entier $\ell \geqslant 2$, il existe $\delta > 0$ et $N_0 \in \mathbb{N}$ tels que pour tout $N \geqslant N_0$, avec une probabilité $> 1 - \varepsilon$, on a les encadrements

$$\delta N^{1/2} < W(E_N) < 6(N \log N)^{1/2} \tag{7}$$

et

$$\delta N^{1/2} < C_\ell(E_N) < 5(\ell N \log N)^{1/2}. \tag{8}$$

Donc, le bon ordre de grandeur de $W(E_N)$ et $C_\ell(E_N)$ est approximativement $N^{1/2}$, à une constante multiplicative et un facteur logarithmique près, comme on peut s'y attendre en vertu du théorème de la limite centrale.

Dans le même article [**9**], les liens entre les corrélations C_k et C_ℓ pour des valeurs différentes de k et ℓ ont été étudiés partiellement (cf. Théorème 4 et 5) et on montre qu'il n'est pas suffisant de restreindre l'étude sur quelques valeurs de ℓ. On peut aussi citer [**27**] et [**25**] pour d'autres relations entre ces différentes mesures.

La question de savoir si $C_2(E_N) \gg \sqrt{N}$ est restée ouverte pendant très longtemps, et a finalement été résolue par N. Alon, Y. Kohayakawa, C. Mauduit, C.-G. Moreira et V. Rödl [**2**] qui ont montré que pour tout $1 \leqslant k \leqslant N$, on a

$$C_{2k}(E_N) \geqslant \sqrt{\frac{1}{2}\left\lfloor \frac{N}{(2k+1)} \right\rfloor}$$

pour tout $E_N \in \{-1,+1\}^N$, puis les mêmes auteurs [**3**] ont montré que pour presque toute suite $E_N \in \{-1,+1\}^N$, on a

$$C_k(E_N) \asymp \sqrt{N \log \binom{N}{k}}.$$

1. Quelques exemples de construction

1.1. Des constructions « naturelles ». Une construction simple de suite binaire est d'écrire la suite des entiers naturels en base 2, en prenant chaque chiffre comme élément de la suite :

$$1,1,0,1,1,1,0,0,1,0,1,\ldots$$

En remplaçant 0 par -1, on obtient la suite de Champernowne binaire qui peut être étudiée dans notre cadre. C'est une suite 2-normale, pourtant sa mesure de bonne distribution et sa mesure de corrélation à l'ordre 2 sont grandes (cf. Théorème 1 dans [**44**]) :

Pour $N \geqslant 17$, on a

$$W(E_N) \geqslant \frac{N}{32 \log N}$$

et

$$C(E_N) \geqslant \frac{1}{48} N.$$

Une autre possibilité est d'utiliser la somme des chiffres en base 2, c'est la suite de Thue-Morse qui peut être définie également en posant :

$$\forall n \in \mathbb{N}, \quad e_0 = 1, \quad e_{2n} = e_n, \quad e_{2n+1} = -e_n.$$

On obtient dans ce cas un meilleur contrôle sur $W(E_N)$ mais $C(E_N)$ reste très large, d'après le Théorème 2 dans le même article [**44**] :

Pour $N \geqslant 5$, on a

$$W(E_N) \leqslant 2(1+\sqrt{3}) N^{(\log 3)/\log 4}$$

et

$$C(E_N) \geqslant \frac{1}{12} N.$$

1.2. Des constructions utilisant la partie fractionnaire. Une idée simple est de considérer une suite dans $\mathbb{R} \setminus \mathbb{N}$, puis on espère que sa partie fractionnaire est relativement bien distribuée. Erdős avait proposé de fixer un irrationnel α et considérer la suite $e_n = 2 \cdot \mathbb{1}_{[0,1/2[}(\{n\alpha\}) - 1$. Comme une fonction linéaire n'est pas très aléatoire, on peut aussi remplacer $n\alpha$ par $n^k\alpha$ pour $k \geqslant 2$. Notons dans cette section la suite $E_{N,k} = (e_{1,k}, \ldots, e_{N,k})$ définie par

$$e_{n,k} = 2 \cdot \mathbb{1}_{[0,1/2[}(\{n^k\alpha\}) - 1. \tag{9}$$

Certaines hypothèses sur α sont nécessaires pour étudier les suites définies dans (9). C'est sous l'hypothèse suivante que C. Mauduit et A. Sárközy ont étudié ce sujet dans [**45**] et [**46**]

($\diamondsuit$): Le développement en fraction continue régulière $\alpha = [a_0; a_1; \ldots]$ est tel que les a_i sont bornés (par K disons).

Sous l'hypothèse ($\diamondsuit$), on a alors l'estimation suivante (cf. Théorèmes 1, 2 et 3 dans [**45**]) :

$$W(E_{N,2}) \ll_K N^{3/5}(\log N)^{1/5}.$$

Les estimations sur les corrélations sont plus compliquées. Les mêmes auteurs ont obtenu certains résultats lors des études sur la généralisation pour $k \geqslant 3$ (cf. [**46**]). Ils définissent d'abord une fonction croissante σ_k qui est de l'ordre

$$\frac{3}{2}k^2(\log k + O(\log_2 k))$$

au voisinage de l'infini. Ils donnent également quelques estimations explicites de σ_k pour les petites valeurs de k. Dans cette perspective, ils démontrent qu'on a :

$$\forall \varepsilon > 0, \exists N_0 = N_0(K, k, \varepsilon), \forall N \geqslant N_0,$$

$$W(E_{N,k}) \ll N^{1-1/\sigma_k+\varepsilon}.$$

Si de plus $k \geqslant 2\ell + 1$, alors il existe également $N_1 = N_1(K, k, \varepsilon)$ tel que pour tous $N \geqslant N_1$, on ait

$$C_\ell(E_{N,k}) \ll N^{1-1/\sigma_k+\varepsilon}.$$

(cf. Théorèmes 1 et 2 dans [**46**]).

Il est très important de noter que la condition $k \geqslant 2\ell + 1$ nous empêche d'avoir une petite valeur sur les corrélation à l'ordre petit sans recourir à un k plus grand. Dans [**45**],

on exhibe un exemple pour $k = 2$ et $\alpha = \sqrt{1601}$ satisfaisant la condition ($\diamondsuit$), mais pour une infinité de N, on a $C_2(E_{N,2}) \gg N$.

En revanche, on peut obtenir une estimation qualitative pour les corrélations à certains vecteurs D fixés. En effet, dans ces deux articles cités dans cette section, il a été démontré qu'en considérant $(e_{n,k})$ comme une suite infinie, pour $d \in \mathbb{N}^*$ on a

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e_{n,k} e_{n+d,k} = 0$$

et pour $0 < d_1 < \cdots < d_{2\ell}$, on a

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e_{n,k} e_{n+d_1,k} \cdots e_{n+d_{2\ell},k} = 0.$$

Cela montre le rôle spécial joué par les « grands vecteurs » dans l'étude quantitative de la définition de la corrélation.

1.3. Des constructions avec les fonctions multiplicatives. On peut aussi considérer les fonctions multiplicatives à valeurs dans $\{-1,+1\}$. Prenons $\Omega(n)$ le nombre de nombres premiers de n comptés avec la multiplicité qui est une fonction additive. En posant $\lambda(n) = (-1)^{\Omega(n)}$, on obtient la fonction de Liouville qui peut être étudiée dans notre cadre. On notera

$$L_N = (\lambda(n))_{n\in[1,N]}$$

D'une part, l'estimation sur la bonne distribution est obtenue dans le Théorème 1 de [**7**]. On a $\forall A > 0$

$$W(L_N) \ll_A \frac{N}{(\log N)^A}$$

Sous l'hypothèse de Riemann généralisée, on a une amélioration sensible (voir la remarque de la section 5 dans [**9**]) : $\forall \varepsilon > 0$

$$W(L_N) \ll_\varepsilon N^{3/4+\varepsilon}$$

D'autre part, l'estimation sur $C_\ell(L_N)$ est encore une question qui semble hors de portée. Même une estimation faible comme

$$\liminf_{x\to\infty} \frac{1}{x} \sum_{n\leqslant x} \lambda(n)\lambda(n+1) < 1$$

n'est pas encore démontrée (cf. [**15**]). Néanmoins, on peut obtenir une estimation de l'ordre de $O(x)$ pour certains vecteurs D particuliers comme $D = (0, d, 2d, \ldots, kd)$, et ce pour toute

fonction complètement multiplicative à valeurs dans $\{-1,+1\}$ (cf. [**7**]). En particulier, on a

$$\sum_{n\leqslant x}\lambda(n)\lambda(n+d)\lambda(n+2d)\leqslant\begin{cases}\frac{7}{9}x+O(\log x) & \text{si } d \text{ pair},\\ \\ \frac{2}{3}x+O(\log x) & \text{si } d \text{ impair}\end{cases}$$

et

$$\sum_{n\leqslant x}\lambda(n)\lambda(n+d)\geqslant\begin{cases}-\frac{1}{3}x+O(\log x) & \text{si } d \text{ impair},\\ \\ -\frac{2}{3}x+O(\log x) & \text{si } d \text{ pair}.\end{cases}$$

Devant ces difficultés, on se ramène à poser d'autres questions. Par exemple, quelle est la complexité de cette suite ? Est-il possible de donner certaines estimations si l'on utilise la fonction de Liouville tronquée définie par

$$\lambda_y(p^\alpha)=\begin{cases}(-1)^\alpha & \text{si } p\leqslant y,\\ 1 & \text{si } p>y.\end{cases}$$

Ou encore la complexité de cette suite tronquée ? La complexité de la suite infinie $(\lambda(n))_{n\in\mathbb{N}}$ au sens où les motifs à toute longueur apparaissent une infinité de fois peut s'en déduire au prix de l'hypothèse H de Schinzel (cf. Théorème 7 dans [**7**]). Toutefois, une estimation sans hypothèse supplémentaire est possible sur la fonction tronquée : Notons $L_N(y)=(\lambda_y(n))_{n\in[1,N]}$, d'après le Théorème 1 dans [**8**], pour $y\geqslant 2$ donné, la complexité $f(k,L_N(y))$ est de l'ordre de $k^{\pi(y)}$ lorsque N est assez grand.

Revenons à la suite L_N. La suite $\Big(\lambda\big(P(n)\big)\Big)_{n\in\mathbb{N}}$ peut-elle être constante à partir d'un certain rang, pour un polynôme $P\in\mathbb{Z}[X]$ donné, comme des calculs numériques de $W(L_N)$ et de $C_\ell(L_N)$ l'ont suggéré. Certaines formes de polynômes ont été étudiées dans [**8**]. On sait par exemple, quelque soit $(a,b)\in\mathbb{N}\times\mathbb{N}^*$ donné, la suite $\big(\lambda(a+bn)\big)_{n\in\mathbb{N}}$ ne sera pas constante à partir de certain rang.

Comme la fonction λ est complètement multiplicatives, une généralisation a été faite pour tous les fonctions multiplicatives à valeurs dans $\{-1,+1\}$ pour l'étude des fonctions tronquées dans [**12**]. Outre certaines améliorations des estimations, on peut porter l'attention sur l'uniformité de y dans ces résultats. On peut citer parmi eux :

$\exists x_0\in\mathbb{R},\quad \forall x\geqslant x_0,\quad 3\leqslant y\leqslant x^{\frac{1}{50\log_2 x}}$,

$$\left|\sum_{n\leqslant x}\lambda_y(n)\lambda_y(n+1)\right|\ll\frac{x}{(\log y)^9}$$

où la constante implicite est absolue. (cf. Corollaire 4 dans [**12**].)

1.4. Autres constructions. Erdős avait proposé d'étudier la suite définie par

$$e_n = \begin{cases} +1 & \text{si } P(n+1) > P(n), \\ -1 & \text{sinon} \end{cases}$$

où $P(n)$ désigne le plus grand facteur premier de n. L'étude générale semble inaccessible, J. Rivat a pu obtenir quelques résultats pour la suite modifiée suivante :

$$e'_n = \begin{cases} +1 & \text{si } P_y(n+1) > P_y(n), \\ -1 & \text{sinon} \end{cases}$$

où $y \in \mathbb{N}$ est fixé et $P(n)$ désigne le plus grand facteur premier p de n avec $p \leqslant y$.

Dans [**53**], il a démontré que pour $3 \leqslant y \leqslant \exp(\dfrac{\log x}{100 \log_2 x})$, on a

$$\left| \sum_{n \leqslant x} e'_n e'_{n+1} + \frac{x}{3} \right| \ll \frac{x}{(\log x)^8}.$$

Si de plus, on a $(a, b) \in \mathbb{N} \times \mathbb{N}^*$ tel que $(a(a+1), b) = 1$, alors

$$\left| \sum_{\substack{j \in \mathbb{N} \\ 1 \leqslant a+jb \leqslant x}} e'_{a+jb} \right| \ll \frac{x}{(\log x)^{10}}$$

Dans une autre direction semblable à celle de la section 1.2, on peut étudier la suite définie par

$$e_n = 2 \cdot \mathbb{1}_{[0,1/2[}\big(\{n^c\}\big) - 1$$

où $c \in]1, \infty[\setminus \mathbb{N}$.

Là encore, d'un côté, l'estimation sur la corrélation à l'ordre de 2 est une question difficile, mais on peut obtenir certains résultats sur les « petites » valeurs de d (cf. Théorème 2 dans [**42**]). De l'autre côté, l'estimation sur la bonne distribution peut se faire à l'aide de l'inégalité d'Erdős-Turán et de la méthode de Van der Corput sur les sommes exponentielles. Les auteurs démontrent alors :

$$W(E_N) \ll N^{1 - \frac{\lceil c \rceil - c}{2^{\lceil c \rceil} - 1}}.$$

2. Avec les caractères de Dirichlet

2.1. Un exemple d'une « bonne » suite pseudo-aléatoire. Cette série d'études avait commencé par la publication de l'article [**43**] dans lequel on a défini plusieurs mesures sur les suites pseudo-aléatoires et on a construit une « bonne » suite avec le symbole de

Legendre. Choisissons un nombre premier suffisamment grand et notons $N = p - 1$, on définit

$$e_n = \left(\frac{n}{p}\right). \tag{10}$$

Alors, les résultats qu'on peut traduire du Théorème 1 dans [**43**] est

$$W(E_N) \ll N^{1/2} \log N$$

et pour tout entier $\ell \geqslant 2$,

$$C_\ell(E_N) \ll \ell N^{1/2} \log N$$

Les démonstrations de ces estimations résultent d'un théorème de André Weil [**66**] dans l'étude des sommes des caractères. Dans la section suivante, on va détailler ces résultats et ses conséquences, en se référant à la monographie de W. Schmidt [**56**].

2.2. Utilisation du symbole de Legendre. Nous avons vu la possibilité d'obtenir une « bonne » suite pseudo-aléatoire dans la section 2.1. Ce n'est pas très raisonnable pour les applications si l'on ne peut pas fournir une large famille des suites pseudo-aléatoires (hormis les différents p) satisfaisant une semblable estimation. L'idée naturelle est de remplacer n par certain polynôme $f(n)$:

$$e_n = \begin{cases} \left(\frac{f(n)}{p}\right) & \text{si} \quad p \nmid f(n), \\ 1 & \text{si} \quad p \mid f(n). \end{cases} \tag{11}$$

Ici, on prend $N = p$.

Dans [**44**], des polynômes de permutation ont été considérés. Mauduit et Sárközy ont réussi à obtenir une estimation semblable sous hypothèse que le polynôme de permutation f n'admet que de zéros de multiplicité impaires. Cependant, à part des polynômes de Dickson (qui sont déjà totalement classés) et autres exemples de polynômes « triviaux » (comme les polynômes linéaires non nuls ou certains monômes), on connaît peu sur la détermination des polynômes de permutation, bien que tous les fonctions bijectives dans un corps fini soient les polynômes de permutation.

Il importe donc d'élargir la famille des polynômes « admissibles ». Remarquons d'abord que les polynômes ayant de racines multiples ne sont pas intéressants à considérer sous le symbole de Legendre. On supposera donc

(γ) $f \in (\mathbb{Z}/p\mathbb{Z})[X]$ est un polynôme de degré $k < p$ n'ayant que de racines simples dans $\overline{\mathbb{Z}/p\mathbb{Z}}$.

L'étude sur la mesure de bonne distribution ne nécessite pas d'autre d'hypothèse supplémentaire, on a (cf. [**23**]) :

$$W(E_p) < 10kp^{1/2}\log p.$$

En revanche, les estimations sur les corrélations sont plus délicates. Outre la condition (γ), on a besoin d'autres conditions pour que le triplet (k, ℓ, p) soit « admissible » (voir [**23**] pour plus de détail de cette définition). Dans ce même article [**23**], les trois conditions suivantes permettent d'obtenir l'admissibilité :

(i) $\ell = 2$;
(ii) $(4\ell)^k < p$;
(iii) 2 est une racine primitive dans $(\mathbb{Z}/p\mathbb{Z})^*$.

Dans ce cas, on a

$$C_\ell(E_p) < 10k\ell p^{1/2}\log p.$$

La condition (iii) est plus intéressante car elle permet d'estimer les corrélations de grand ordre pour un polynôme de grand degré. Cependant, on ne dispose pas de bon moyen pour déterminer ces nombres premiers. On ne sais d'ailleurs pas s'il existe une infinité de nombres premiers p tels que 2 soit une racine primitive de $(\mathbb{Z}/p\mathbb{Z})^*$, mais la conjecture d'Artin prévoit qu'ils existent avec une densité positive. D'après un résultat de Heath-Brown sur la conjecture d'Artin (cf. [**28**]), on sait néanmoins qu'il existe une infinité de nombres premiers dont la plus petite racine primitive est $\leqslant 5$.

2.3. Sommes des caractères dans un corps fini. Le résultat important qu'on va utiliser dans la suite est la Proposition 7. J'ai donné une forme améliorée sur la constante car il importe d'avoir une petite constante lorsqu'on a besoin de résultat numérique.

Soit χ un caractère multiplicatif d'ordre d sur le corps fini $\mathbb{F}_q$. On sait qu'il existe un nombre premier p et $r \in \mathbb{N}^*$ tels que $q = p^r$ et $d|(q-1)$. Il est facile de voir que

$$\sum_{x\in\mathbb{F}_q} \chi(x) = \begin{cases} q & \text{si } \chi = \chi_0, \\ 0 & \text{sinon.} \end{cases}$$

L'estimation devient plus difficile lorsque le x dans χ est remplacé par une expression polynomiale $f(x)$. Un polynôme (à plusieurs indéterminées) est dit *absolument irréductible* s'il est irréductible sur tout corps d'extension algébrique de $\mathbb{F}_q$. D'un côté, on sait que les deux conditions $(\alpha 1)$ et $(\alpha 2)$ sont équivalentes (cf. lemme 2C dans [**56**] p.11) :

$(\alpha 1)$ $Y^d - f(X)$ est absolument irréductible.

($\alpha 2$) Si $f(X) = a(X - x_1)^{d_1} \cdots (X - x_s)^{d_s}$ est la factorisation de f dans la clôture algébrique $\overline{\mathbb{F}}_q$ avec $x_i \neq x_j$ lorsque $i \neq j$, alors $(d, d_1, d_2, \ldots, d_s) = 1$.

De l'autre côté, soit $g \in \mathbb{F}_q[X]$ un polynôme de degré n, on sait que la condition ($\beta 1$) implique ($\beta 2$) (cf. Théorème 1B dans **[56]** p.92) :

($\beta 1$) $0 < n < q$ et $(n, q) = 1$.

($\beta 2$) $Z^q - Z - g(X)$ est absolument irréductible.

On suppose désormais que χ est non principal. Le premier résultat important est (cf. Théorème 2C dans **[56]** p.43)

PROPOSITION 2. *Soit $\chi \neq \chi_0$ un caractère multiplicatif d'ordre d sur $\mathbb{F}_q[X]$. Si $f \in \mathbb{F}_q[X]$ admet s racines distinctes et vérifie la condition ($\alpha 2$), alors on a*

$$\left| \sum_{x \in \mathbb{F}_q} \chi\big(f(x)\big) \right| \leqslant (s-1)q^{1/2}.$$

On a une estimation analogue pour les caractères additifs (cf. Théorème 2E dans **[56]** p.44) :

PROPOSITION 3. *Soient $\Psi \neq \Psi_0$ un caractère additif sur $\mathbb{F}_q$ et $g \in \mathbb{F}_q[X]$ un polynôme de degré n vérifiant ($\beta 2$), alors on a*

$$\left| \sum_{x \in \mathbb{F}_q} \Psi\big(g(x)\big) \right| \leqslant (n-1)q^{1/2}. \tag{12}$$

Ainsi, si q est un nombre premier, tout polynôme non constant vérifie (12).

Le troisième résultat est qu'on a le même type d'estimation pour une « somme hybride ». On a comme précédemment (voir aussi Théorème 2G dans **[56]** p.45),

PROPOSITION 4. *Soient $\chi \neq \chi_0$ un caractère multiplicatif d'ordre d et $\Psi \neq \Psi_0$ un caractère additif sur $\mathbb{F}_q$, $f \in \mathbb{F}_q[X]$ possède s racines distinctes dans $\overline{\mathbb{F}}_q$ et vérifie la condition ($\alpha 2$), $g \in \mathbb{F}_q[X]$ est de degré n vérifiant la condition ($\beta 2$). Alors on a*

$$\left| \sum_{x \in \mathbb{F}_q} \chi\big(f(x)\big)\Psi\big(g(x)\big) \right| \leqslant (s+n-1)q^{1/2}.$$

D'après ce qui précède, nous pouvons maintenant formuler cette proposition utile :

PROPOSITION 5. *Soient p un nombre premier, $\chi \neq \chi_0$ un caractère de Dirichlet d'ordre d sur $\mathbb{F}_p$ et $f \in \mathbb{F}_q[X]$ un polynôme tel que sa factorisation dans $\overline{\mathbb{F}}_p$ s'écrit $f(X) = a(X - x_1)^{d_1} \cdots (X - x_s)^{d_s}$ vérifiant*

$$(d, d_1, d_2, \ldots, d_s) = 1.$$

Alors pour $a \in \mathbb{F}_p$, on a

$$\left| \sum_{x \in \mathbb{F}_q} \chi(f(x)) \, \mathrm{e}\left(\frac{ax}{p}\right) \right| \leqslant sp^{1/2}.$$

DÉMONSTRATION. En effet, si $a = 0$, c'est le Proposition 2, sinon $x \to \mathrm{e}\left(\frac{ax}{p}\right)$ est un caractère additif non principal. Le polynôme X vérifie la condition $(\beta 1)$ et l'estimation voulue découle du Proposition 4. □

Pour passer aux sommes incomplètes, on utilise la proposition suivante :

PROPOSITION 6. *Soit $g : \mathbb{Z} \to \mathbb{C}$ une fonction périodique de période $m \in \mathbb{Z}$. Pour $(X, Y) \in \mathbb{R} \times \mathbb{R}^+$, on a*

$$\left| \sum_{X < n \leqslant X+Y} g(n) \right| \leqslant \frac{Y+1}{m} \left| \sum_{n=1}^{m} g(n) \right| + \sum_{\substack{-\frac{m}{2} < h \leqslant \frac{m}{2} \\ h \neq 0}} \frac{1}{m \sin\left(\frac{\pi}{m}|h|\right)} \left| \sum_{n=1}^{m} g(n) \, \mathrm{e}\left(\frac{hn}{m}\right) \right|. \tag{13}$$

DÉMONSTRATION. On a

$$\frac{1}{m} \sum_{-\frac{m}{2} < h \leqslant \frac{m}{2}} \mathrm{e}\left(\frac{h}{m}(a-b)\right) = \delta_{ab}.$$

Donc

$$\begin{aligned}
\left|\sum_{X<n\leqslant X+Y} g(n)\right| &= \frac{1}{m}\left|\sum_{X<n\leqslant X+Y}\sum_{s=1}^{m} g(s)\sum_{-\frac{m}{2}<h\leqslant\frac{m}{2}} \mathrm{e}\left(\frac{h}{m}(n-s)\right)\right| \\
&\leqslant \frac{1}{m}\sum_{-\frac{m}{2}<h\leqslant\frac{m}{2}}\left|\sum_{s=1}^{m}\mathrm{e}\left(-\frac{hs}{m}\right)g(s)\sum_{X<n\leqslant X+Y}\mathrm{e}\left(\frac{hn}{m}\right)\right| \\
&\leqslant \frac{1}{m}\left|\sum_{s=1}^{m} g(s)\sum_{X<n\leqslant X+Y} 1\right| \\
&\quad + \frac{1}{m}\sum_{\substack{-\frac{m}{2}<h\leqslant\frac{m}{2}\\ h\neq 0}}\left|\sum_{s=1}^{m}\mathrm{e}\left(-\frac{hs}{m}\right)g(s)\sum_{X<n\leqslant X+Y}\mathrm{e}\left(\frac{hn}{m}\right)\right| \\
&\leqslant \frac{Y+1}{m}\left|\sum_{s=1}^{m} g(s)\right| + \frac{1}{m}\sum_{\substack{-\frac{m}{2}<h\leqslant\frac{m}{2}\\ h\neq 0}}\frac{1}{\sin\left(\frac{\pi}{m}|h|\right)}\left|\sum_{s=1}^{m}\mathrm{e}\left(-\frac{hs}{m}\right)g(s)\right| \\
&\leqslant \frac{Y+1}{m}\left|\sum_{n=1}^{m} g(n)\right| + \sum_{\substack{-\frac{m}{2}<h\leqslant\frac{m}{2}\\ h\neq 0}}\frac{1}{m\sin\left(\frac{\pi}{m}|h|\right)}\left|\sum_{n=1}^{m} g(n)\,\mathrm{e}\left(\frac{hn}{m}\right)\right|
\end{aligned}$$

□

Remarque: Si l'on utilise la convexité de sin sur $[0,\pi/2]$ tel que

$$\sin\left(\frac{\pi}{m}|h|\right) \geqslant \frac{2|h|}{m},$$

on retrouve une forme usuelle :

$$\left|\sum_{X<n\leqslant X+Y} g(n)\right| \leqslant \frac{Y+1}{m}\left|\sum_{n=1}^{m} g(n)\right| + \sum_{1\leqslant h\leqslant\frac{m}{2}}\frac{1}{h}\left|\sum_{n=1}^{m} g(n)\,\mathrm{e}\left(\frac{hn}{m}\right)\right|.$$

On peut finalement obtenir la proposition fondamentale suivante (qui améliore la constante initialement obtenue dans le Théorème 2 de [**43**]) pour la suite de notre études sur les suites pseudo-aléatoires :

PROPOSITION 7. *Soient $p \geqslant 5$ un nombre premier, χ et f comme dans la Proposition 5, alors pour $(X,Y) \in \mathbb{R}\times\mathbb{R}^+$ avec $0 < Y \leqslant p$, on a*

$$\left|\sum_{X<n\leqslant X+Y}\chi\left(f(n)\right)\right| \leqslant 2sp^{1/2}\log p.$$

DÉMONSTRATION. D'après les Propositions 5 et 6, on a

$$\left|\sum_{X<n\leqslant X+Y}\chi(f(n))\right| \leqslant \frac{p+1}{p}\left|\sum_{n=1}^{p}\chi(f(n))\right| + \sum_{\substack{-\frac{p}{2}<h\leqslant\frac{p}{2}\\ h\neq 0}}\frac{1}{p\sin\left(\frac{\pi}{p}|h|\right)}\left|\sum_{n=1}^{p}\chi(f(n))\,\mathrm{e}\left(\frac{hn}{p}\right)\right|$$

$$\leqslant \big(1+1/p\big)sp^{1/2} + sp^{1/2}\sum_{\substack{-\frac{p}{2}<h\leqslant\frac{p}{2}\\ h\neq 0}}\frac{1}{p\sin\left(\frac{\pi}{p}|h|\right)} \tag{14}$$

Pour estimer la sommation sur $1/\sin$, on compare avec son intégrale :

$$\begin{aligned}\sum_{\substack{-\frac{p}{2}<h\leqslant\frac{p}{2}\\ h\neq 0}}\frac{1}{\sin\left(\frac{\pi}{p}|h|\right)} &\leqslant \sum_{h=1}^{p-1}\frac{1}{\sin\left(\frac{\pi h}{p}\right)}\\ &\leqslant \int_{\frac{1}{2}}^{p-\frac{1}{2}}\frac{dh}{\sin\left(\frac{\pi h}{p}\right)} = \frac{p}{\pi}\log\tan\frac{\pi h}{2p}\,\Big|_{\frac{1}{2}}^{p-\frac{1}{2}}\\ &\leqslant \frac{2p}{\pi}\log\cot\frac{\pi}{4p}\\ &\leqslant \frac{2p}{\pi}\log\frac{4p}{\pi}\end{aligned}$$

En reportant dans (14), on obtient

$$\begin{aligned}\left|\sum_{X<n\leqslant X+Y}\chi(f(n))\right| &\leqslant \big(1+1/p+\frac{2}{\pi}\log\frac{4p}{\pi}\big)sp^{1/2}\\ &\leqslant 2sp^{1/2}\log p \qquad\qquad (\text{si } p\geqslant 5).\end{aligned}$$

□

2.4. Un autre exemple utilisant les caractères de Dirichlet. On avait utilisé le symbole de Legendre qui est un caractère réel. Remarquons qu'un caractère de Dirichlet χ en général correspond à un élément primitif g de $(\mathbb{Z}/p\mathbb{Z})^*$ et qui définit ses valeurs par la puissance d'indice d'une racine primitive $\omega\in\mathbb{U}_{p-1}$ correspondante :

$$\chi(n)=\omega^{\mathrm{ind}_g\, n}$$

avec $g^{\mathrm{ind}_g\, n}\equiv n \bmod p$.

Étant donné un nombre premier $p \in \mathbb{N}$ et un élément primitif $g \in (\mathbb{Z}/p\mathbb{Z})^*$, on peut construire une autre suite en posant $N = p - 1$ et

$$e_n = \begin{cases} +1 & \text{si} \quad 1 \leqslant \operatorname{ind}_g n \leqslant \frac{p-1}{2}, \\ -1 & \text{si} \quad \frac{p+1}{2} \leqslant \operatorname{ind}_g n \leqslant p-1. \end{cases}$$

On obtient une suite pseudo-aléatoire satisfaisante si la répartition des indices sont « relativement » uniforme, ce qui a été démontrée par Sárközy dans **[60]** :

$$W(E_N) \ll N^{1/2}(\log N)^2$$

et

$$C_\ell(E_N) \ll \ell 8^k N^{1/2}(\log N)^{\ell+1}.$$

Dans le cas du symbole de Legendre de la section 2.1, on avait construit la suite en considérant la parité des indices. Cet exemple de construction a été généralisé dans **[24]** et **[26]**, suivant la démarche similaire que décrit la section 2.2.

CHAPITRE 3

Sur les questions probabilistes

1. Un comportement des corrélations de petit ordre

Pour poursuivre l'étude des propriétés stochastiques des mesures W et C_ℓ, nous abordons deux questions probabilistes posées par A. Sárközy. Considérons $(e_n)_{n\in\mathbb{N}}$ comme une suite de variables aléatoires indépendantes suivant la loi de Bernoulli symétrique :

$$\mathbb{P}(e_n = -1) = \mathbb{P}(e_n = +1) = \frac{1}{2}. \tag{15}$$

Alors est-il vrai que la quantité

$$\lim_{N\to\infty} \left(\max_{N<m<2N} \frac{|\sum_{n=1}^{m} e_n|}{\sqrt{m}} \right)$$

est presque sûrement infinie ? Est-il vrai que la quantité

$$\limsup_{N\to\infty} \frac{\left|\sum_{n=1}^{N} e_n e_{n+1}\right|}{\sqrt{N}}$$

est presque sûrement infinie ?

Plus généralement, pour des entiers $0 < d_1 < \cdots < d_k$ fixés, on aimerait connaître l'ordre de grandeur des sommes

$$\sum_{n=1}^{N} e_{n+d_1} \cdots e_{n+d_k}.$$

Nous pouvons répondre à cette dernière question grâce au théorème suivant :

THÉORÈME 8. *Soit $(e_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes suivant la loi de Bernoulli symétrique définie par* (15). *Alors presque sûrement*

$$\limsup_{N\to\infty} \frac{\sum_{n=1}^{N} e_{n+d_1} \cdots e_{n+d_k}}{\sqrt{2N \log_2 N}} = 1, \tag{16}$$

et

$$\liminf_{N\to\infty} \frac{\sum_{n=1}^{N} e_{n+d_1} \cdots e_{n+d_k}}{\sqrt{2N \log_2 N}} = -1. \tag{17}$$

Pour obtenir ce résultat, on commence par remarquer une propriété intéressante de la loi de Bernoulli réelle :

PROPOSITION 8. *Soit $I \subset \mathbb{N}$ et $(Y_i)_{i\in I}$ une famille de variables aléatoires suivant une loi de Bernoulli réelle quelconque. Alors les $(Y_i)_{i\in I}$ sont indépendantes dans leur ensemble si et seulement si pour tout $J \subset I$ fini, on a*

$$\mathbb{E}(\prod_{j\in J} Y_j) = \prod_{j\in J} \mathbb{E}(Y_j). \tag{18}$$

Remarque: La condition (18) est presque minimale. On pourrait imaginer améliorer cette condition par une condition « en chaîne » qui demande de vérifier l'égalité (pour $I = [1, N]$) dans le cas $Y_1, Y_1Y_2, Y_1Y_2Y_3, \ldots, Y_1Y_2\cdots Y_N$. Cependant il existe un contre exemple dès $N = 3$: Considérons $\Omega = \{a_1, a_2, \ldots, a_{12}\}$ muni de la probabilité uniforme, et

$$\begin{aligned} I_1 &= \{a_1, a_2, a_3, a_7, a_9, a_{11}\}, \\ I_2 &= \{a_1, a_2, a_4, a_7, a_8, a_{10}\}, \\ I_3 &= \{a_1, a_3, a_5, a_7, a_8, a_9\}. \end{aligned}$$

Posons $X_i = \mathbb{1}_{I_i}$ pour $i = 1, 2, 3$. Alors les X_i suivent la loi de Bernoulli symétrique et on vérifie les égalités

$$\mathbb{E}(X_1X_2) = \mathbb{E}(X_1)\mathbb{E}(X_2) = 0, \quad \mathbb{E}(X_1X_2X_3) = \mathbb{E}(X_1)\mathbb{E}(X_2)\mathbb{E}(X_3) = 0,$$

mais X_1 et X_3 ne sont pas indépendantes, puisque I_1 et I_3 ne le sont pas.

DÉMONSTRATION. Il s'agit de montrer que la condition (18) est suffisante pour assurer l'indépendance des $(Y_i)_{i\in I}$. Notons a_i et $\bar{a}_i$ les deux valeurs réelles que les Y_i peuvent prendre, et $\Omega_i = \{a_i, \bar{a}_i\}$. Par convention on écrit $\bar{\bar{x}} = x$ pour $x \in \Omega_i (i \in I)$. Soit $J \subset I$ fini, avec $J \neq \emptyset$. Alors pour tout $(y_i)_{i\in J} \in \prod_{i\in J} \Omega_i$, on obtient en développant le produit

$$\prod_{j\in J}(Y_j - y_j) = \sum_{k+k'=|J|} \sum_{j_1<\cdots<j_k} \sum_{j'_1<\cdots<j'_{k'}} (-1)^k y_{j_1}\cdots y_{j_k}\, Y_{j'_1}\cdots Y_{j'_{k'}}$$

où les indices de sommation sont des éléments de J qui vérifient

$$\{j_1, \cdots, j_k\} \cap \{j'_1, \cdots, j'_k\} = \emptyset.$$

En appliquant la linéarité de l'espérance et l'hypothèse (18), on obtient

$$\mathbb{E}(\prod_{j\in J}(Y_j - y_j)) = \sum_{k+k'=|J|} \sum_{j_1<\cdots<j_k} \sum_{j'_1<\cdots<j'_{k'}} (-1)^k y_{j_1}\cdots y_{j_k}\mathbb{E}(Y_{j'_1})\cdots\mathbb{E}(Y_{j'_{k'}}).$$

En factorisant

$$\mathbb{E}(\prod_{j\in J}(Y_j - y_j)) = \prod_{j\in J}(\mathbb{E}(Y_j) - y_j) = \prod_{j\in J}\mathbb{E}((Y_j - y_j)).$$

D'après la définition de l'espérance, en éliminant les termes nuls, l'égalité précédente s'écrit

$$\left(\prod_{j\in J}(\bar{y_j} - y_j)\right)\mathbb{P}(\forall j \in J, Y_j = \bar{y_j}) = \prod_{j\in J}(\bar{y_j} - y_j)\ \mathbb{P}(Y_j = \bar{y_j}),$$

et en simplifiant le facteur non nul $\prod_{j\in J}(\bar{y_j} - y_j)$, on a établi pour tout $J \subset I$ fini

$$\mathbb{P}(\forall j \in J, Y_j = \bar{y_j}) = \prod_{j\in J}\mathbb{P}(Y_j = \bar{y_j}),$$

ce qui est la définition de l'indépendance de la famille $(Y_i)_{i\in I}$. □

LEMME 1. *Soit $(e_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes suivant la loi de Bernoulli symétrique définie par* (15). *Pour des entiers $k \geqslant 1$, et $0 < d_1 < \cdots < d_k$, la suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ définie pour tout $n \in \mathbb{N}$ par*

$$X_n = e_{n+d_1} \cdots e_{n+d_k} \tag{19}$$

est une suite de variables aléatoires indépendantes, qui suivent la loi de Bernoulli symétrique.

DÉMONSTRATION. Il est clair que $X_n \in \{-1, +1\}$, et pour $\varepsilon = \pm 1$, on a

$$\begin{aligned}\mathbb{P}(X_n = \varepsilon) &= \mathbb{P}(e_{n+d_1} = \varepsilon)\ \mathbb{P}(e_{n+d_2} \cdots e_{n+d_k} = 1) \\ &\quad + \mathbb{P}(e_{n+d_1} = -\varepsilon)\ \mathbb{P}(e_{n+d_2} \cdots e_{n+d_k} = -1) \\ &= \tfrac{1}{2}\ \mathbb{P}(e_{n+d_2} \cdots e_{n+d_k} = 1) + \tfrac{1}{2}\ \mathbb{P}(e_{n+d_2} \cdots e_{n+d_k} = -1),\end{aligned}$$

donc $\mathbb{P}(X_n = -1) = \mathbb{P}(X_n = 1)$. Ainsi les X_n suivent la loi de Bernoulli symétrique définie par (15) et $\mathbb{E}(X_n) = 0$. D'après la Proposition 8, il suffit donc de montrer que pour tout ensemble fini $I \subset \mathbb{N}$,

$$\mathbb{E}(\prod_{i\in I} X_i) = \prod_{i\in I}\mathbb{E}(X_i) = 0.$$

Soit $I \subset \mathbb{N}$ fini. On note $D = \{d_1, \ldots, d_k\}$ et on écrit

$$\prod_{i\in I} X_i = \prod_{j\in I+D} e_j^{\nu_j}$$

où ν_j représente le nombre d'occurences de e_j dans le produit des X_i. Maintenant comme $e_j^2 = 1$, on peut éliminer les termes pour lesquels ν_j est pair, et on obtient

$$\prod_{i\in I} X_i = \prod_{j\in J} e_j$$

avec $J \subset I + D$. De plus $J \neq \emptyset$ car si on note i_1 le plus petit élément de I, alors le facteur $e_{i_1+d_1}$ n'apparaît que dans X_1, donc $i_1 + d_1 \in J$. Comme les $(e_j)_{j\in J}$ sont indépendantes, on a

$$\mathbb{E}(\prod_{i\in I} X_i) = \mathbb{E}(\prod_{j\in J} e_j) = \prod_{j\in J} \mathbb{E}(e_j) = 0.$$

□

Pour achever la démonstration du Théorème 8, nous allons faire appel au célèbre résultat suivant démontré par Khintchine en 1924 :

THÉORÈME 9 (Loi du log itéré). *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires qui suivent la loi de Bernoulli symétrique définie par* (15). *Si on note $S_N = \sum_{n=1}^{N} X_n$, alors presque sûrement on a les égalités*

$$\limsup_{N\to\infty} \frac{S_N}{\sqrt{2N\log_2 N}} = 1, \quad \liminf_{N\to\infty} \frac{S_N}{\sqrt{2N\log_2 N}} = -1.$$

DÉMONSTRATION. Voir par exemple [**18**] p.204. □

FIN DU PREUVE DU THÉORÈME 8. En appliquant le Théorème 9 aux variables aléatoires X_n définies par (19) (qui satisfont les hypothèses requises d'après le Lemme 1), on obtient la conclusion du Théorème 8. □

2. Un comportement de marche aléatoire

La première question posée par A. Sárközy est plus délicate. Il n'est pas évident que la limite existe donc on s'intéresse à la limite inférieure. Après quelques réflexions, nous avons été convaincu que même si on remplace le 2 par toute autre quantité finie, on n'aura pas une quantité presque sûrement infinie, sans pour autant formuler une démonstration valide. G. Tenenbaum a reformulé cette question après avoir pris connaissance de ce problème et nous a fait l'honneur d'autoriser la reproduction de sa démonstration ici.

THÉORÈME 10 (Tenenbaum). *Soient $A > 0$, $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes centrées, de variance 1, vérifiant $\sup_n |X_n| \leqslant A$. Posons $S_n = \sum_{m=1}^{n} X_m$.*

Soit encore $f : \mathbb{N}^* \to \mathbb{R}^+$ une fonction croissante. La quantité

$$\liminf_{N\to\infty} \max_{N\leqslant n\leqslant f(N)N} \frac{|S_n|}{\sqrt{n}} \tag{20}$$

est presque sûrement finie si

$$\limsup_{N\to\infty} \frac{f(N)(\log_2 N)}{\log N} < \infty,$$

alors qu'elle est presque sûrement infinie si

$$\lim_{N\to\infty} \frac{\log f(N)}{\log_2 N} = \infty.$$

PREUVE D'UN CAS PARTICULIER DU THÉORÈME 10. Comme on s'intéresse essentiellement à la loi de Bernoulli symétrique, ce qui simplifie certains étapes de démonstration, nous montrerons seulement le Théorème 10 dans le cas particulier où $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires indépendantes suivant la loi de Bernoulli symétrique.

Comme principaux outils, la démonstration utilisera le lemme de Borel-Cantelli et une inégalité de Kolmogorov. Borel-Cantelli nous affirme que (voir par exemple [**18**] p.200 ou [**21**] p.231) :

LEMME 2. *Soit $(A_n)_{n\in\mathbb{N}}$ une suite d'évènements.*
Si $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) < \infty$, alors $\mathbb{P}(\limsup_{n\in\mathbb{N}} A_n) = 0$.
Si $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) = \infty$ et que les A_n sont indépendants deux à deux, alors $\mathbb{P}(\limsup_{n\in\mathbb{N}} A_n) = 1$.

L'inégalité suivante est dûe à Kolmogorov (voir [**18**] p.234 ou [**21**] p.253) :

LEMME 3. *Soit S_n la somme comme au début de cette section. Pour tous $a > 0$ et $n \geqslant 1$, on a*

$$\mathbb{P}\big(\max_{1\leqslant k\leqslant n} S_k \geqslant a\big) \leqslant \mathrm{e}^{-a^2/(2n)}.$$

Par symétrie, on a alors

$$\mathbb{P}\big(\max_{1\leqslant k\leqslant n} |S_k| \geqslant a\big) \leqslant 2\,\mathrm{e}^{-a^2/(2n)}.$$

Montrons d'abord que si $f(N) \ll \frac{\log N}{\log_2 N}$, alors la quantité (20) est finie.

Posons pour $j \geqslant 1$

$$N_j = \lfloor j \log_2 j \rfloor^j \text{ et } H_j = 2^{\lfloor (\log j)/\log 2 \rfloor}.$$

Soit $k > 0$. Nous considérons trois suites d'évènements dont le terme général est défini comme suit pour chaque $j \geqslant 1$:

$$A_j := \Big\{ \big|S_{H_j N_j}\big| \leqslant 2\sqrt{N_{j+1}} \Big\},$$

$$B_j := \bigcap_{0 \leqslant t \leqslant (\log j)/\log 2} B_{j,t} \text{ avec } B_{j,t} := \Big\{ \max_{2^t N_{j+1} \leqslant n \leqslant 2^{t+1} N_{j+1}} \big|S_{2^{t+1} N_{j+1}} - S_n\big| \leqslant k\sqrt{2^t N_{j+1}} \Big\},$$

$$C_j := \Big\{ \big|S_{N_{j+1}} - S_{H_j N_j}\big| \leqslant k\sqrt{N_{j+1}} \Big\}.$$

D'après l'inégalité de Kolmogorov, on a

$$\mathbb{P}\Big(\max_{2^t N_{j+1} \leqslant n \leqslant 2^{t+1} N_{j+1}} \big|S_{2^{t+1} N_{j+1}} - S_n\big| > k\sqrt{2^t N_{j+1}} \Big) \leqslant 2\,\mathrm{e}^{-k^2/2}.$$

Comme B_j est une intersection d'évènements indépendants, on alors

$$\mathbb{P}(B_j) \geqslant (1 - 2\,\mathrm{e}^{-k^2/2})^{(\log j)/\log 2}.$$

Cela implique, pour k assez grand et pour $j \geqslant j_0(k)$,

$$\mathbb{P}(B_j) \geqslant 2/\sqrt{j}.$$

De même, l'inégalité de Markov appliquée à C_j, donne pour $j \geqslant j_1(k)$

$$\mathbb{P}(C_j) \geqslant 1/2.$$

D'où

$$\mathbb{P}(B_j C_j) = \mathbb{P}(B_j)\mathbb{P}(C_j) \geqslant 1/\sqrt{j},$$

et

$$\sum_{j \geqslant 1} \mathbb{P}(B_j C_j) = \infty.$$

Comme les $B_j C_j$ sont indépendants deux à deux, le lemme de Borel-Cantelli implique que p.s. une infinité de $B_j C_j$ sont réalisés.

Pour les évènements A_j, on a pour j assez grand

$$\frac{N_{j+1}}{H_j N_j} \geqslant 2 \log N_{j+1}.$$

Les majorations des complémentaires A_j obtenus par l'inégalité de Markov donnent alors une série convergente. D'après le lemme de Borel-Cantelli, on en conclut qu'avec la probabilité 1 tous les A_j sauf au plus un nombre fini sont réalisés.

Pour ces valeurs de j on peut donc écrire pour tout $n \in [N_{j+1}, H_{j+1}N_{j+1}]$, $2^h N_{j+1} < n \leqslant 2^{h+1}N_{j+1}$,

$$\begin{aligned} |S_n| &\leqslant |S_{H_jN_j}| + |S_{N_{j+1}} - S_{H_jN_j}| + \sum_{0\leqslant t<h} |S_{2^{t+1}N_{j+1}} - S_{2^tN_{j+1}}| + |S_n - S_{2^hN_{j+1}}| \\ &\leqslant \Big(2 + k + k\sum_{0\leqslant t\leqslant h} 2^{t/2}\Big)\sqrt{N_{j+1}} \\ &\leqslant 8k\sqrt{n}. \end{aligned}$$

Comme $H_j \gg (\log N_j)/\log_2 N_j$, cela nous permet de conclure que la quantité (20) est presque sûrement finie.

Supposons à présent que $f(N) = (\log N)^{\xi(N)}$ où $\xi(N) \to \infty$ et montrons que (20) est p.s infinie. Étant donné une fonction $k \mapsto \lambda_k$ tendant vers l'infini avec k, nous posons

$$N_k^* = k^{\lambda_k k}.$$

Si la croissance de λ_k est assez lente, on a $N_{k+1}^* \ll N_k^*(\log N_k^*)^{\lambda_k+1}$. Donc, pour N assez grand, au moins un intervalle de la forme $J_k := [N_k^*, N_{k+1}^*]$ est contenu dans $[N, Nf(N)]$. Nous pouvons ainsi nous limiter à montrer que l'on a p.s.

$$Y_k := \max_{n\in J_k} \frac{|S_n|}{\sqrt{n}} \to \infty \quad (k \to \infty).$$

Soit $u \in \mathbb{R}$, l'inégalité $Y_k(w) \leqslant u$ implique que pour tout $t \in [1, \lambda_k \log k]$, on a

$$\begin{aligned} |S_{\mathrm{e}^t N_k^*} - S_{\mathrm{e}^{t-1} N_k^*}| &\leqslant \sqrt{\mathrm{e}^{t-1} N_k^*} + \sqrt{\mathrm{e}^t N_k^*} \\ &\leqslant \frac{(\sqrt{\mathrm{e}}+1)u}{\sqrt{\mathrm{e}-1}}\sqrt{(\mathrm{e}^t - \mathrm{e}^{t-1})N_k^*} \end{aligned}$$

Comme $\frac{(\sqrt{\mathrm{e}}+1)}{\sqrt{\mathrm{e}-1}} > 2$, l'inégalité de Kolmogorov appliquée à chaque t donne alors

$$\mathbb{P}(Y_k \leqslant u) \leqslant \left(1 - 2\,\mathrm{e}^{-u^2}\right)^{\lambda_k \log k - 1}.$$

Finalement,

$$\sum_{k\geqslant 1} \mathbb{P}(Y_k \leqslant u) < \infty.$$

Comme u est arbitraire, d'après le lemme de Borel-Cantelli, Y_k tend vers l'infini presque sûrement. □

CHAPITRE 4

Autour des caractères de Dirichlet

1. Sur la coïncidence des suites définies par le symbole de Legendre

La construction définie par (11) nous permet d'obtenir une grande famille de « bonnes » suites pseudo-aléatoires. Une question naturelle est de demander : étant donné deux polynômes non constants f et g, quand est-ce que les deux suites de symboles de Legendre $\left(\frac{f(n)}{p}\right)$ et $\left(\frac{g(n)}{p}\right)$ coïncident ? Combien de coïncidences consécutives sur les valeurs de ces deux suites doit-on observer pour avoir la coïncidence totale ? Quelles sont alors les relations liant f et g dans ce cas ? On constate que si

$$f(x) = (x+a)^2 \text{ et } g(x) = (x+b)^2$$

avec $a \neq b$, alors les deux suites sont trivialement égales à la suite constante 1. En revanche, si l'on choisit c un résidu quadratique modulo p et on pose $g = cf$, alors ces deux suites ne coïncident que lorsque $p \mid f(n)$. Hormis ce type de situation, on peut obtenir une réponse partielle aux questions précédentes :

THÉORÈME 11. *Soit $p \geqslant 5$ un nombre premier, f, g deux polynômes à coefficients dans $(\mathbb{Z}/p\mathbb{Z})$ de degrés respectifs k et ℓ avec $k, \ell \leqslant \sqrt{p}$. Si les deux suites $\left(\frac{f(n)}{p}\right)$ et $\left(\frac{g(n)}{p}\right)$ ont $> 3(k+\ell)p^{1/2}\log p$ valeurs coïncidentes consécutives, alors ces deux suites sont identiques. De plus, si f et g sont unitaires et ne contiennent pas de facteur carré dans leur factorisation dans $(\mathbb{Z}/p\mathbb{Z})[X]$, alors $f = g$.*

Remarque: Notons que la condition $k, \ell \leqslant \sqrt{p}$ n'est pas contraignante puisque la conclusion devient triviale dès que k ou $\ell \geqslant \sqrt{p}$. On remarque également que ce résultat est trivial si $p < 12907$.

L'outil essentiel pour cette démonstration est le Lemme 7. Cependant, on va énoncer un lemme qui possède son propre intérêt et qui permet de mettre en lumière le raisonnement. Ce lemme a été inspiré par le Lemme 1 de [**1**].

LEMME 4. *Soit p un nombre premier et F un corps commutatif de caractéristique p. Soit $P \in F[X]$ un polynôme de degré $< p$. Alors on a une décomposition unique dans $F[X]$*

$$P = Q^2 R \tag{21}$$

où $Q, R \in F[X]$ avec Q unitaire et R ne contenant pas de racine multiple dans $\overline{F}$ (la clôture algébrique de F).

DÉMONSTRATION. Comme F est un corps commutatif, $F[X]$ est un anneau factoriel et on a une décomposition unique de P sous la forme

$$P = aS_1^{2\alpha_1+1} \cdots S_s^{2\alpha_s+1} T_1^{2\beta_1} \cdots T_t^{2\beta_t} \tag{22}$$

où $a \in F$, $\alpha_i, \beta_i \in \mathbb{N}^*$, S_i et T_i sont des polynômes unitaires irréductibles à coefficients dans F et sont premiers entre eux.

Posons

$$Q = S_1^{\alpha_1} \cdots S_s^{\alpha_s} T_1^{\beta_1} \cdots T_t^{\beta_t}, \quad R = aS_1 \cdots S_s,$$

alors on obtient l'expression (21). Cette décomposition est unique puisque l'expression (22) l'est.

Il reste à démontrer que R ne contient pas de racine multiple dans $\overline{F}$. Lorsque $i \neq j$, S_i et S_j ne peuvent pas avoir une racine commune dans $\overline{F}$ car sinon, d'après l'identité de Bézout, leur pgcd (S_i, S_j) aurait un facteur non trivial de S_i (et de S_j) qui est supposé irréductible. En plus, si S_i possède une racine multiple dans $\overline{F}$, alors comme $deg(S_i) < p$, on a $S_i' \neq 0$, ce qui implique que $deg\big((S_i, S_i')\big) \geqslant 1$. Ceci est impossible pour la même raison puisque $(S_i, S_i') \mid S_i$ et S_i est supposé irréductible. Finalement R n'admet que de racines simples dans $\overline{F}$. □

DÉMONSTRATION DU THÉORÈME 11. Supposons que les deux suites $\left(\frac{f(n)}{p}\right)$ et $\left(\frac{g(n)}{p}\right)$ possèdent plus de $3(k+\ell)p^{1/2}\log p$ valeurs coïncidentes consécutives.

Posons $P = fg$, un polynôme de degré $k + \ell < p$. D'après le lemme 4, on peut écrire

$$P = Q^2 R$$

avec $Q, R \in (\mathbb{Z}/p\mathbb{Z})[X]$ tels que Q est unitaire et R ne possède pas de racine multiple dans la clôture algébrique de $(\mathbb{Z}/p\mathbb{Z})[X]$.

Nous allons montrer que R est une constante. Supposons le contraire. Alors, d'un côté, comme le symbole de Legendre est un caractère de Dirichlet d'ordre 2, la proposition 7

nous assure que pour tous $X, Y \in \mathbb{R}$ avec $0 < Y \leqslant p$, on a

$$\left| \sum_{X<n\leqslant X+Y} \left(\frac{R(n)}{p}\right) \right| \leqslant 2(k+\ell)p^{1/2}\log p.$$

De l'autre côté, choisissons $X+1$ comme le début de la coïncidence des deux suites $\left(\frac{f(n)}{p}\right)$ et $\left(\frac{g(n)}{p}\right)$. D'après l'hypothèse, nous pouvons prendre $2(k+\ell)p^{1/2}\log p < Y \leqslant p$ tel que pour $X < n \leqslant X+Y$ et n n'est ni une racine de f ni une racine de g dans $\mathbb{Z}/p\mathbb{Z}$, on ait

$$1 = \left(\frac{f(n)}{p}\right)\left(\frac{g(n)}{p}\right) = \left(\frac{h(n)}{p}\right) = \left(\frac{R(n)}{p}\right),$$

ce qui implique

$$\left| \sum_{X<n\leqslant X+Y} \left(\frac{R(n)}{p}\right) \right| > 3(k+\ell)p^{1/2}\log p - (k+\ell) \geqslant 2(k+\ell)p^{1/2}\log p.$$

Cette contradiction prouve que R doit être une constante qui est un résidu quadratique de $\mathbb{Z}/p\mathbb{Z}$ (puisqu'il y a au moins une valeur coïncidente). Finalement, les deux suites de symboles de Legendre sont identiques. Dans le cas où f et g sont unitaires et ne contenant pas de carré dans leur factorisation dans $(\mathbb{Z}/p\mathbb{Z})[X]$, par l'unicité de l'écriture sur Q et R, on peut conclure que $f = Q = g$. □

Remarque: Supposons que f et g sont deux polynômes unitaires de $\mathbb{Z}/p\mathbb{Z}[X]$, nous venons de montrer que les trois affirmations suivantes sont équivalentes :

- f et g ont la même parité sur le degré de tous les facteurs dans leur décomposition dans $\mathbb{Z}/p\mathbb{Z}[X]$.
- Les deux suites $\left(\frac{f(n)}{p}\right)$ et $\left(\frac{g(n)}{p}\right)$ ont plus de $2(k+\ell)p^{1/2}\log p$ valeurs identiques consécutives.
- Les deux suites $\left(\frac{f(n)}{p}\right)$ et $\left(\frac{g(n)}{p}\right)$ sont identiques.

2. Du symbole de Legendre aux caractères de Dirichlet

Introduction. Une autre idée pour généraliser la construction (10) est de remplacer le symbole de Legendre par un caractère $\chi \neq \chi_0$ quelconque. E. Fouvry a proposé d'étudier la suite définie par

$$\mathrm{e}_n = \begin{cases} +1 & \text{si } \mathfrak{Re}(\chi(n)) \geqslant 0, \\ -1 & \text{sinon} \end{cases} \tag{23}$$

pour un caractère de Dirichlet χ mod p d'ordre d donné.

Nous pouvons démontrer que cela constitue un autre bon candidat parmi les suites pseudo-aléatoires. En fait, lorsque p est assez grand, nous avons les estimations suivantes :

THÉORÈME 12. *Soient $p \geqslant 2999$ un nombre premier, $\ell \geqslant 2$ un entier et $\chi \bmod p$ un caractère non principal d'ordre $d \geqslant 2$. Soit $E_p = (e_i)_{i\in[1,p]}$ la suite définie par (23). Alors, on a*

$$W(E_p) \leqslant 2p^{1/2}(\log p)^2 + 4p/d$$

et

$$C_\ell(E_p) \leqslant 3\ell p^{1/2}(\log 8p)^{\ell+1} + 4\ell p/d.$$

Lorsque $d > \sqrt{p}$, on peut supprimer les termes contenant p/d dans les deux estimations.

On peut se demander s'il existe beaucoup de caractères d'ordre supérieur à $\sqrt{p}$. Comme les caractères forment un groupe isomorphe au groupe additif $\mathbb{Z}/(p-1)\mathbb{Z}$, il y a

$$\varphi(p-1) \gg \frac{p}{\log_2 p}$$

éléments d'ordre maximal égal à $p-1$. Plus précisément, comme $\mathbb{Z}/(p-1)\mathbb{Z}$ est cyclique, pour tout $d \mid (p-1)$, il y a exactement $\varphi(d)$ caractères d'ordre d. Donc, l'ordre moyen d'un caractère modulo p est

$$F(p-1) := \frac{1}{p-1}\sum_{d\mid(p-1)} \varphi(d)d.$$

La fonction F est multiplicative et on peut aussi écrire pour un entier n quelconque,

$$F(n) = \sum_{kd=n} \frac{\varphi(d)}{k} = \varphi * \frac{1}{J}(n).$$

Donc

$$\begin{aligned} F(p^v) &= \frac{1}{p^v} + \sum_{i=1}^{v} \frac{p^i - p^{i-1}}{p^{v-i}} \\ &= p^{-v} + p^{-v}\sum_{i=1}^{v}(p^{2i} - p^{2i-1}) \\ &= p^{-v}\sum_{i=0}^{2v}(-p)^i \\ &= p^v\left(1+\frac{1}{p}\right)^{-1}\left(1+\frac{1}{p^{2v+1}}\right). \end{aligned}$$

Finalement,

$$F(n) = n \prod_{p^v \| n} \left(1 + \frac{1}{p}\right)^{-1} \left(1 + \frac{1}{p^{2v+1}}\right).$$

Comme le produit infini $\prod_{p^v \| n} \left(1 + \frac{1}{p^{2v+1}}\right)$ est convergent (puisque $v \geqslant 1$), on a alors

$$F(n) \asymp n \prod_{p \mid n} \left(1 + \frac{1}{p}\right)^{-1}.$$

Si $p-1$ est un entier égal à son noyau, l'ordre minimal $\frac{p}{\log_2 p}$ trouvé ci-dessus est atteint à une constante près. Si $p = 2^k + 1$, alors l'ordre moyen est proche de p.

3. Une approximation trigonométrique de Vaaler

Dans l'article **[64]**, J-D. Vaaler présente des inégalités extrémales utiles en analyse, obtenues à l'aide de la fonction de Beurling ou de ses dérivées. La fonction de Beurling, popularisée par Selberg, est une fonction entière définie par :

$$B(z) = \left(\frac{\sin \pi z}{\pi}\right) \left\{ \sum_{n \geqslant 0} (z-n)^{-2} - \sum_{m \geqslant 1} (z+m)^{-2} + \frac{2}{z} \right\}.$$

Un des résultats présentés par Vaaler est une bonne approximation des fonctions normalisées à variation bornée. Le lemme suivant reprend le cas des fonctions périodiques. C'est une partie du Théorème 19 dans ce même article **[64]** (p.211). Avec les mêmes notations que Vaaler, on définit

$$\hat{J}(t) = \begin{cases} 1 & \text{si } t = 0, \\ \pi t(1 - |t|) \cot(\pi t) + |t| & \text{si } 0 < |t| < 1, \\ 0 & \text{si } |t| \geqslant 1, \end{cases}$$

et

$$\hat{K}(t) = \max(1 - |t|, 0).$$

Pour $H \in \mathbb{N}$, on pose

$$j_H(x) = \sum_{n=-H}^{H} \hat{J}\left(\frac{n}{H+1}\right) \mathrm{e}(nx),$$

et

$$k_H(x) = \sum_{n=-H}^{H} \hat{K}\left(\frac{n}{H+1}\right) \mathrm{e}(nx).$$

Notre lemme s'énonce ainsi :

LEMME 5. *Soit $f : \mathbb{R} \to \mathbb{C}$ une fonction normalisée, de variation bornée sur tout intervalle fermé et 1-périodique. On note $V_f(x)$ la variation totale de f sur $[-\frac{1}{2}, x]$, alors on a*

$$|f(x) - f * j_H(x)| \leqslant (2H+2)^{-1}(dV_f) * k_H(x) \tag{24}$$

où

$$f * j_H(x) = \sum_{n=-H}^{H} \hat{f}(n)\hat{J}\left(\frac{n}{H+1}\right) \mathrm{e}(nx)$$

et

$$(dV_f) * k_H(x) = \sum_{n=-H}^{H} \widehat{dV_f}(n)\hat{K}\left(\frac{n}{H+1}\right) \mathrm{e}(nx)$$

sont des notations de convolution.

A l'aide de ce lemme, nous pouvons démontrer le résultat suivant qui est crucial pour notre étude :

LEMME 6. *Soit $f : \mathbb{U} \to \mathbb{Z}$ telle que*

$$f(z) = \begin{cases} +1 & si \quad |\arg z| \leqslant \pi/2, \\ -1 & si \quad \pi/2 < |\arg z| \leqslant \pi. \end{cases} \tag{25}$$

Alors, pour tout $H \in \mathbb{N}^*$, il existe deux fonctions

$$\phi_H(z) = \sum_{k=-H}^{H} a_H(k) z^k$$

et

$$\psi_H(z) = \sum_{k=-H}^{H} b_H(k) z^k,$$

vérifiant

$$a_H(0) = 0, \quad |b_H(0)| \leqslant \frac{2}{H}, \tag{26}$$

et pour $k \in [-H, H] \setminus \{0\}$

$$|a_H(k)| \leqslant \frac{1}{|k|}, \quad |b_H(k)| \leqslant \frac{2}{H}, \tag{27}$$

telles qu'on ait

$$\forall z \in \mathbb{U}, |f(z) - \phi_H(z)| \leqslant \psi_H(z). \tag{28}$$

DÉMONSTRATION. Appliquons le lemme 5 à la fonction F définie par

$$F(x) = \begin{cases} 1 & \text{si} \quad |x| < \frac{1}{4}, \\ 0 & \text{si} \quad |x| = \frac{1}{4}, \\ -1 & \text{si} \quad \frac{1}{4} < |x| \leqslant \frac{1}{2}. \end{cases}$$

F est 1-périodique, de variation bornée. Ses coefficients de Fourier se calculent facilement : $\hat{F}(0) = 0$, pour $n \neq 0$, on a

$$\hat{F}(n) = \frac{\sin(\pi n/2)}{\pi n/2}.$$

De même, on obtient

$$\widehat{dV_F}(n) = 4\cos(\pi n/2).$$

Donc

$$j_H(x) = \sum_{\substack{k=-H \\ k\neq 0}}^{H} \frac{\sin(\pi k/2)}{\pi k/2} \hat{J}\left(\frac{k}{H+1}\right) \mathrm{e}(kx)$$

et

$$(dV_F) * k_H(x) = 4 \sum_{k=-H}^{H} \cos(\pi k/2) \hat{K}\left(\frac{k}{H+1}\right) \mathrm{e}(kx).$$

Posons

$$\tilde{\phi}_H(x) = F * j_H(x), \quad \tilde{\psi}_H(x) = (2H+2)^{-1}(dV_F) * k_H(x).$$

D'après (24), on déduit que

$$\left|F(x) - \tilde{\phi}_H(x)\right| \leqslant \tilde{\psi}_H(x).$$

Puisque $\tilde{\phi}_H$ et $\tilde{\psi}_H$ sont des fonctions continues, on peut changer F par G, la fonction 1-périodique définie par

$$G(x) = \begin{cases} 1 & \text{if } |x| \leqslant 1/4, \\ -1 & \text{if } 1/4 < |x| \leqslant 1/2 \end{cases}$$

qui vérifie la même inégalité :

$$\left|G(x) - \tilde{\phi}_H(x)\right| \leqslant \tilde{\psi}_H(x).$$

À ce stade, on réécrit cette inégalité en posant $z = \mathrm{e}(x) \in \mathbb{U}$, de sorte que

$$\phi_H(z) = \tilde{\phi}_H(x), \quad \psi_H(z) = \tilde{\psi}_H(x).$$

On remarque que $f(z) = G(x)$, les fonctions $\hat{J}$, $\hat{K}$ sont majorées par 1, ce qui permet de conclure. □

4. Démonstration du Théorème

4.1. Preuve de l'estimation de bonne distribution. Par construction on a pour $n < p$

$$f(\chi(n)) = e_n. \tag{29}$$

L'idée suivante est d'appliquer la proposition 7. Toutefois, l'inégalité de Pólya-Vinogradov (cf. par exemple [**13**] p. 135) suffit pour estimer $W(E_N)$. En effet, on a le lemme suivant :

LEMME 7. *Soient $p \geqslant 3$ un nombre premier et $a, b, t \in \mathbb{N}^*$ tels que $1 \leqslant a \leqslant a+(t-1)b \leqslant p$. Supposons que χ est un caractère non principal modulo p, alors on a*

$$\left|\sum_{j=0}^{t-1} \chi(a+jb)\right| \leqslant p^{1/2} \log p.$$

DÉMONSTRATION. Notons b' l'inverse de b modulo p. Il suffit d'écrire

$$\begin{aligned}\left|\sum_{j=0}^{t-1} \chi(a+jb)\right| &= \left|\sum_{j=0}^{t-1} \chi(ab'+j)\chi(b)\right| \\ &= \left|\sum_{j=0}^{t-1} \chi(ab'+j)\right| \\ &\leqslant p^{1/2}\log p\end{aligned}$$

où la dernière inégalité est due à Pólya et Vinogradov. □

Soient $a, b, t \in \mathbb{N}^*$ tels que $1 \leqslant a \leqslant a+(t-1)b \leqslant p-1$. Appliquons le lemme 6, on a pour tout $H \in \mathbb{N}^*$

$$\left|\sum_{j=0}^{t-1} (f(\chi(a+jb)) - \phi_H(\chi(a+jb)))\right| \leqslant \sum_{j=0}^{t-1} \psi_H(\chi(a+jb)),$$

où ϕ_H, ψ_H sont définie dans ce même lemme.

Donc

$$\left|\sum_{j=0}^{t-1} e_{a+jb}\right| \leqslant \sum_{k=-H}^{H} |a_H(k)| \left|\sum_{j=0}^{t-1} \chi^k(a+jb)\right| + \sum_{k=-H}^{H} |b_H(k)| \left|\sum_{j=0}^{t-1} \chi^k(a+jb)\right|. \tag{30}$$

D'après (26) et (27), on a

$$\sum_{k=-H}^{H} |a_H(k)| \left|\sum_{j=0}^{t-1} \chi^k(a+jb)\right| \leqslant \sum_{\substack{k=-H \\ k\neq 0, d\nmid k}}^{H} \frac{1}{|k|} \left|\sum_{j=0}^{t-1} \chi^k(a+jb)\right| + \sum_{\substack{k=-H \\ k\neq 0, d\mid k}}^{H} \frac{p}{|k|} \tag{31}$$

et

$$\sum_{k=-H}^{H} |b_H(k)| \left|\sum_{j=0}^{t-1} \chi^k(a+jb)\right| \leqslant \frac{2}{H} \sum_{\substack{k=-H \\ k\neq 0, d\nmid k}}^{H} \left|\sum_{j=0}^{t-1} \chi^k(a+jb)\right| + \frac{2}{H} \sum_{\substack{k=-H \\ d\mid k}}^{H} p. \tag{32}$$

Lorsque $d \nmid k$, χ^k est non principal et le lemme 7 s'applique pour $\chi^k(a+jb)$. Sinon, on donne une majoration triviale. Choisissons $H = d-1$ si $d \leqslant p^{1/2}$ et $H = \lfloor p^{1/2} \rfloor$ dans le cas contraire afin de limiter le nombre des estimations triviales. En reportant cela dans (30), on obtient

$$\left|\sum_{j=0}^{t-1} e_{a+jb}\right| \leqslant 2(\log H + 1)p^{1/2}\log p + 4p^{1/2}\log p + 2p/H. \tag{33}$$

On voit que le choix de $H \asymp p^{1/2}$ permet d'optimiser la taille de ces trois termes et donne une bonne estimation lorsque $d > p^{1/2}$.

Donc, pour $p \geqslant 2999$, on a

$$\begin{aligned} W(E_p) &\leqslant p^{1/2}(\log p)^2 + 6p^{1/2}\log p + 4p/d \\ &\leqslant 2p^{1/2}(\log p)^2 + 4p/d. \qquad (\text{ puisque } \log p \geqslant 6) \end{aligned} \tag{34}$$

D'ailleurs, si $d > p^{1/2}$, le terme $2p/H$ du membre droit de (33) est remplacé par $2p^{1/2}\log p$. Pour la même raison numérique, on obtient la majoration finale $2p^{1/2}(\log p)^2$, c'est-à-dire qu'on peut omettre le terme $4p/d$ lorsque $d > p^{1/2}$.

4.2. Preuve de l'estimation des corrélations. Montrons d'abord un lemme facile

LEMME 8. *Soient $z_1, \ldots, z_r, z'_1, \ldots, z'_r \in \mathbb{U}$, alors on a*

$$\left|\prod_{i=1}^{r} z_i - \prod_{i=1}^{r} z'_i\right| \leqslant \sum_{i=1}^{r} |z_i - z'_i|.$$

DÉMONSTRATION. Il suffit d'écrire

$$\begin{aligned} & |z_1\cdots z_r - z'_1\cdots z'_r| \\ = \; & |(z_1-z'_1)z_2\cdots z_r + z'_1(z_2-z'_2)z_3\cdots z_r + \cdots + z'_1\cdots z'_r(z_r-z'_r)| \\ \leqslant \; & |z_1-z'_1| + \cdots + |z_r-z'_r|. \end{aligned} \tag{35}$$

□

Soient $M, d_1, \dots, d_\ell$ les entiers vérifiant $0 \leqslant d_1 < d_2 < \cdots \leqslant d_\ell < p - M$. En utilisant le lemme 6 et en appliquant l'inégalité (35), on obtient

$$\begin{aligned}
&\left| \sum_{n=1}^{M} e_{n+d_1} \cdots e_{n+d_\ell} - \sum_{n=1}^{M} \phi_H(\chi(n+d_1)) \cdots \phi_H(\chi(n+d_\ell)) \right| \\
\leqslant\ & \sum_{n=1}^{M} \sum_{i=1}^{\ell} |e_{n+d_i} - \phi_H(\chi(n+d_i))| \\
\leqslant\ & \sum_{i=1}^{\ell} \sum_{n=1}^{M} \psi_H(\chi(n+d_i)).
\end{aligned}$$

Désignons I_3 le dernier terme de l'inégalité précédente. Pour tous entiers U et V, on a

$$\sum_{n=U}^{V} \psi_H(\chi(n)) = \sum_{k=-H}^{H} b_H(k) \sum_{n=U}^{V} \chi^k(n).$$

Supposons que $0 \leqslant V - U < p$. Si $d \nmid k$, on peut utiliser le lemme 7 lors de la sommation sur n. En tenant compte de (26) et (27), on obtient

$$\left| \sum_{n=U}^{V} \psi_H(\chi(n)) \right| \leqslant 4p^{1/2} \log p + \sum_{\substack{k=-H \\ d \mid k}}^{H} \frac{2p}{H} \leqslant 4p^{1/2} \log p + 2p/H$$

si $H < d$. Donc

$$|I_3| \leqslant 4\ell p^{1/2} \log p + 2\ell p/H. \tag{36}$$

lorsque $H < d$. En fait, ce dernier terme est négligeable par rapport au premier si H est assez grand.

Il nous reste à estimer le terme principal, à savoir

$$\begin{aligned}
I_4 &:= \sum_{n=1}^{M} \phi_H(\chi(n+d_1)) \cdots \phi_H(\chi(n+d_\ell)) \\
&= \sum_{n=1}^{M} \sum_{\substack{(k_1,\dots,k_\ell) \in [-H,H]^\ell \\ \forall i \in [1,\ell], k_i \neq 0}} a_H(k_1) \cdots a_H(k_\ell) \chi^{k_1}(n+d_1) \cdots \chi^{k_\ell}(n+d_\ell) \\
&= \sum_{\substack{(k_1,\dots,k_\ell) \in [-H,H]^\ell \\ \forall i \in [1,\ell], k_i \neq 0}} a_H(k_1) \cdots a_H(k_\ell) \sum_{n=1}^{M} \chi^{k_1}(n+d_1) \cdots \chi^{k_\ell}(n+d_\ell). \qquad (37)
\end{aligned}$$

Pour tout $k \in \mathbb{Z}$, on note $\bar{k}$ l'unique entier $0 \leqslant \bar{k} < d$ tel que $k \equiv \bar{k} \bmod d$. Pour tout ℓ-uplet $K = (k_1, \dots, k_\ell)$, on note D_K le pgcd de $\bar{k_1}, \dots, \bar{k_\ell}$. Ici, on a $1 \leqslant D_K \leqslant d - 1$.

Posons $k_i' = \bar{k_i}/D_K$, alors

$$\begin{aligned}\chi^{k_1}(n+d_1)\cdots\chi^{k_\ell}(n+d_\ell) &= \chi^{\bar{k_1}}(n+d_1)\cdots\chi^{\bar{k_\ell}}(n+d_\ell)\\ &= \chi^{D_K}\left((n+d_1)^{k_1'}\cdots(n+d_\ell)^{k_\ell'}\right)\end{aligned}$$

et les caractères χ^{D_K} sont non principaux mais d'ordre

$$\delta_K = \frac{d}{(d, D_K)} > 1.$$

On a en outre $(\delta_K, k_1', \ldots, k_\ell') = 1$. La proposition 7 nous assure alors

$$\left|\chi^{k_1}(n+d_1)\cdots\chi^{k_\ell}(n+d_\ell)\right| \leqslant 2\ell p^{1/2}\log p.$$

En tenant compte de (26) et (27), on reporte cette estimation dans (37) et obtient

$$\begin{aligned}|I_4| &\leqslant \sum_{\substack{(k_1,\ldots,k_\ell)\in[-H,H]^\ell\\ \forall i\in[1,\ell], d\nmid k_i}} \frac{1}{|k_1|}\cdots\frac{1}{|k_\ell|}\left|\sum_{n=1}^{M}\chi^{D_K}\left((n+d_1)^{k_1'}\cdots(n+d_\ell)^{k_\ell'}\right)\right|\\ &\quad+\sum_{\substack{(k_1,\ldots,k_\ell)\in[-H,H]^\ell\\ \forall i\in[1,\ell], k_i\neq 0, d\mid k_i}}\frac{1}{|k_1|}\cdots\frac{1}{|k_\ell|}M\\ &\leqslant 2\ell(2\log H+2)^\ell p^{1/2}\log p + p\left(\sum_{\substack{k=-H\\ k\neq 0\\ d\mid k}}^{H}\frac{1}{|k|}\right)^\ell\end{aligned}$$

En faisant le même choix pour H comme pour l'estimation de bonne distribution, on fait disparaître le dernier terme de l'inégalité précédente, et on obtient alors numériquement pour $p \geqslant 2999$, $0 \leqslant d_1 < d_2 < \cdots \leqslant d_\ell \leqslant p - M$

$$\begin{aligned}\left|\sum_{n=1}^{M} e_{n+d_1}\cdots e_{n+d_\ell}\right| &\leqslant |I_3| + |I_4| + 1\\ &\leqslant 4\ell p^{1/2}\log p + 4\ell p/d + 2\ell p^{1/2}(\log 8p)^{\ell+1} + 1 \qquad (38)\\ &\leqslant 3\ell p^{1/2}(\log 8p)^{\ell+1} + 4\ell p/d.\end{aligned}$$

Finalement,

$$C_\ell(E_p) \leqslant 3\ell p^{1/2}(\log 8p)^{\ell+1} + 4\ell p/d.$$

Lorsque $d > p^{1/2}$, le terme $4\ell p/d$ dans (38) qui correspond au terme $2\ell p/H$ dans (36) est remplacé par $2\ell p^{1/2}\log p$, ce qui est absorbé par le terme contenant $(\log 8p)^{\ell+1}$ (pour $\ell \geqslant 2$) lorsque $p \geqslant 2999$. Ainsi, nous pouvons omettre ce terme.

5. Passons aux polynômes

La suite définie par (23) étudiée dans la section précédente est une suite pseudo-aléatoire intéressante lorsque l'ordre de χ est plus grand que $\sqrt{p}$. À l'instar de (11), nous pouvons nous demander si le même type d'estimation pourrait avoir lieu lorsqu'on utilise les polynômes. À savoir, si on choisit $g \in (\mathbb{Z}/p\mathbb{Z})[X]$ et on généralise (23) en posant

$$\mathrm{e}_n = \begin{cases} +1 & \text{si } \mathfrak{Re}(\chi(g(n))) \geqslant 0, \\ -1 & \text{sinon.} \end{cases} \tag{39}$$

Certes, sans hypothèse supplémentaire sur g, l'étude semble impossible. Nous nous proposons de démontrer ici le théorème suivant :

THÉORÈME 13. *Soient $p \geqslant 251$ un nombre premier, $\chi \bmod p$ un caractère non principal d'ordre d, $g \in (\mathbb{Z}/p\mathbb{Z})[X]$ un polynôme possédant $s \leqslant \sqrt{p}$ racines distinctes dans la clôture algébrique de $\mathbb{Z}/p\mathbb{Z}$ et $E_p = (e_n)_{n\in[1,p]}$ la suite définie par (39). Alors on a*

$$W(E_p) \leqslant 4sp^{1/2}(\log p)^2 + 4p/d$$

et

$$C_2(E_p) \leqslant 5sp^{1/2}(\log 8p)^3 + 8p/d$$

où les termes p/d peuvent être omis lorsque $d > \sqrt{p}$.

Si g est irréductible dans $(\mathbb{Z}/p\mathbb{Z})[X]$, alors pour $\ell \geqslant 3$, on a également

$$C_\ell(E_p) \ll \ell sp^{1/2}(\log 8p)^{\ell+1} + p/d.$$

Ce théorème montre la possibilité d'une classe très large de suites pseudo-aléatoires. La démonstration se décompose en trois parties, où l'on reprend les notations de la section 2.

5.1. Estimation sur $W(E_p)$. Nous avons besoin de la proposition 7 cette fois-ci pour estimer $W(E_p)$. Écrivons la factorisation de g dans sa clôture algébrique :

$$g(X) = c(X - x_1)\cdots(X - x_s)$$

avec $c \in \mathbb{Z}/p\mathbb{Z}$ et $\forall i \in [1, s]$, $x_i \in \overline{\mathbb{Z}/p\mathbb{Z}}$.

Soient $a, b, t \in \mathbb{N}^*$ tels que $1 \leqslant a \leqslant a + (t-1)b \leqslant p$. Alors

$$h(X) := g(a + bX) = cb^s\big(X - b^{-1}(x_1 - a)\big) \cdots \big(X - b^{-1}(x_s - a)\big)$$

est aussi un polynôme possédant s racines distinctes (les opérations sont dans $\mathbb{Z}/p\mathbb{Z}$), et donc h s'annule au plus s fois dans $\mathbb{Z}/p\mathbb{Z}$.

D'après (28) et (29), nous pouvons écrire

$$\left| \sum_{\substack{j=0 \\ h(j)\neq 0}}^{t-1} \big(e_{a+jb} - \varphi_H\big(\chi(h(j))\big)\big) \right| \leqslant \sum_{\substack{j=0 \\ h(j)\neq 0}}^{t-1} \psi_H(\chi(h(j))).$$

En tenant compte de (26) et (27), on a

$$\begin{aligned} \left| \sum_{\substack{j=0 \\ h(j)\neq 0}}^{t-1} \varphi_H(\chi(h(j))) \right| &\leqslant \sum_{k=-H}^{H} \frac{1}{|k|} \left| \sum_{\substack{j=0 \\ h(j)\neq 0}}^{t-1} \chi^k(h(j)) \right| \\ &\leqslant \sum_{\substack{k=-H \\ k\neq 0, d\nmid k}}^{H} \frac{1}{|k|} \left| \sum_{j=0}^{t-1} \chi^k(h(j)) \right| + \sum_{\substack{k=-H \\ k\neq 0, d\mid k}}^{H} \frac{p}{|k|} \end{aligned}$$

De même, on a

$$\left| \sum_{\substack{j=0 \\ h(j)\neq 0}}^{t-1} \psi_H(\chi(h(j))) \right| \leqslant \sum_{\substack{k=-H \\ k\neq 0, d\nmid k}}^{H} \frac{2}{H} \left| \sum_{j=0}^{t-1} \chi^k(h(j)) \right| + \sum_{\substack{k=-H \\ k\neq 0, d\mid k}}^{H} \frac{2p}{H}$$

Pour la même raison, on choisit $H = d-1$ si $d \leqslant p^{1/2}$ et $H = \lfloor p^{1/2} \rfloor$ sinon. En utilisant la proposition 7, on trouve alors

$$\begin{aligned} \left| \sum_{j=0}^{t-1} e_{a+jb} \right| &\leqslant s + \left| \sum_{\substack{j=0 \\ h(j)\neq 0}}^{t-1} \varphi_H(\chi(h(j))) \right| + \left| \sum_{\substack{j=0 \\ h(j)\neq 0}}^{t-1} \psi_H(\chi(h(j))) \right| \\ &\leqslant 2(\log p^{1/2} + 1) \cdot 2sp^{1/2} \log p + 4sp^{1/2} \log p + 2p/H \qquad (40) \\ &\leqslant 4sp^{1/2}(\log p)^2 + 4p/d \qquad (\text{si } p \geqslant 251). \end{aligned}$$

Lorsque $d > p^{1/2}$, on peut remplacer le terme $2p/H$ de (40) par $2p^{1/2}\log p$, ce qui nous permet d'omettre numériquement le terme $4p/d$ à la fin.

Finalement, on obtient l'estimation voulue sur $W(E_p)$.

5.2. Estimation sur $C_2(E_p)$. Nous avons besoin du lemme suivant pour l'estimation de $C_2(E_p)$.

LEMME 9. $\forall A, B \subsetneq \mathbb{Z}/p\mathbb{Z},\ \sharp A > 0, \sharp B = 2,\ \exists c \in \mathbb{Z}/p\mathbb{Z},\ \exists!(a,b) \in A \times B, a+b=c.$

DÉMONSTRATION. En fait, c'est la partie (i) du Théorème 2 dans [**23**] mais elle n'a pas été énoncé de cette manière. Re-démontrons rapidement ici.

Écrivons $B = \{b, b+d\}$ avec $d > 0$. Soit $a \in A$, il existe $T \in \mathbb{N}^*$, tel que $a + Td \notin A$ mais $a + (T-1)d \in A$. Posons

$$\begin{aligned} c &= a + b + Td \\ &= a + (T-1)d + b + d \end{aligned}$$

On va voir que c est un entier qui s'exprime de manière unique par la somme des deux éléments respectives de A et B. Supposons qu'il existe $(a', b') \in A \times B$ tel que $c = a' + b'$ avec $b' \neq b + d$ alors $b' = b$ et $a' = a + Td \in A$ ce qui contredit la définition de T. □

Nous sommes en mesure de d'estimer $C_2(E_p)$. Soient $d_1, d_2 \in \mathbb{N}^*$ avec $d_1 < d_2$. On dit que deux polynômes P et Q de $(\mathbb{Z}/p\mathbb{Z})[X]$ sont équivalents s'il existe $a \in \mathbb{Z}/p\mathbb{Z}$ tel que $P(X) = Q(X+a)$. C'est évidemment une relation d'équivalence et on peut ainsi regrouper les facteurs irréductibles de g :

$$g(X) = \prod_{j \in I} \left(\prod_{i \in I_j} P_i \right).$$

(Les polynômes $P_i \in (\mathbb{Z}/p\mathbb{Z})[X]$ sont irréductibles dans $\mathbb{Z}/p\mathbb{Z}[X]$, $I \neq \emptyset$ et chaque I_j est une classe d'équivalence.)

Remarque: On remarque que les P_i ont tous des racines simples et les polynômes appartenant à différentes classes ne peuvent pas posséder des racines dont la différence est un élément de $\mathbb{Z}/p\mathbb{Z}$:

Sinon, il existerait deux polynômes irréductibles P, Q qui ne sont pas équivalents et $(x,y) \in \overline{\mathbb{Z}/p\mathbb{Z}} \times \mathbb{Z}/p\mathbb{Z}$ tels que $P(x) = Q(x+y) = 0$. Alors $\tilde{Q}(X) := Q(X+y) \in \mathbb{Z}/p\mathbb{Z}[X]$ et $\tilde{Q} \neq P$ mais le pgcd de P et $\tilde{Q}$ est de degré $\geqslant 1$, ce qui contredirait l'irréductibilité de P.

Posons

$$L := \{n \in \mathbb{Z}/p\mathbb{Z} : g(n+d_1)g(n+d_2) \equiv 0 \bmod p\}.$$

Alors le cardinal de L est $\leqslant 2s$. Lorsque $n \notin L$, on a

$$e_{n+d_1}e_{n+d_2} = f\big(\chi(g(n+d_1))\big)f\big(\chi(g(n+d_2))\big).$$

D'après (28) et (35), on a $\forall H \in \mathbb{N}^*$

$$\begin{aligned}(41)\qquad &\left|\sum_{\substack{n=0\\ n\notin L}}^{M}\Big(e_{n+d_1}e_{n+d_2} - \varphi_H\big(\chi(g(n+d_1))\big)\varphi_H\big(\chi(g(n+d_2))\big)\Big)\right| \\ &\leqslant \sum_{\substack{n=0\\ n\notin L}}^{M}\psi_H\big(\chi(g(n+d_1))\big) + \sum_{\substack{n=0\\ n\notin L}}^{M}\psi_H\big(\chi(g(n+d_2))\big)\end{aligned}$$

La sommation sur $\psi_H\big(\chi(g(n+d_i))\big)$ se fait en appliquant directement la proposition 7 car g n'admet que des racines simples :

$$\begin{aligned}\sum_{\substack{n=0\\ n\notin L}}^{M}\psi_H\big(\chi(g(n+d_i))\big) &\leqslant \sum_{k=-H}^{H}\frac{2}{H}\left|\sum_{n=0}^{M}\chi^k(g(n+d_i))\right| \\ (42)\qquad &\leqslant 8sp^{1/2}\log p + \sum_{\substack{k=-H\\ d\,|\,k}}^{H}\frac{2p}{H}\end{aligned}$$

Pour traiter la sommation portant sur φ_H, d'une part on peut écrire avec les notations dans la section 2

$$\begin{aligned}&\left|\sum_{\substack{n=0\\ n\notin L}}^{M}\varphi_H\big(\chi(g(n+d_1))\big)\varphi_H\big(\chi(g(n+d_2))\big)\right| \\ &\leqslant \sum_{\substack{k_1,k_2=-H\\ k_1,k_2\neq 0}}^{H}\frac{1}{|k_1k_2|}\left|\sum_{n=0}^{M}\chi\big(g^{\overline{k_1}}(n+d_1)g^{\overline{k_2}}(n+d_2)\big)\right|.\end{aligned}$$

D'autre part, on peut écrire pour g :

$$g(X) = \prod_{j\in I}\left(\prod_{a_i\in A_j}P_j(X+a_i)\right)$$

où $A_j \subset \mathbb{Z}/p\mathbb{Z}$ vérifient $0 < \sharp A_j < p$.

Posons

$$\begin{aligned} h_{k_1,k_2}(X) : &= g^{\overline{k_1}}(X+d_1)g^{\overline{k_2}}(X+d_2) \\ &= \prod_{j\in I}\left(\prod_{(a_i,a_i')\in A_j^2} P_j^{\overline{k_1}}(X+a_i+d_1)P_j^{\overline{k_2}}(X+a_i'+d_2)\right). \end{aligned}$$

Pour un j fixé, on peut regrouper les P_j tels que $a_i + d_1 = a_i' + d_2$ pour former un polynôme irréductible élevé à la puissance $\overline{k_1} + \overline{k_2}$. Appliquons le lemme 9 à chaque A_j et $B = \{d_1, d_2\}$, on déduit que pour chaque classe j, il existe un polynôme irréductible P_j élevé à la puissance $\overline{k_1}$ ou $\overline{k_2}$ (donc tous ne sont pas de puissance $\overline{k_1} + \overline{k_2}$). Ainsi, le pgcd de ces puissances est égal à celui de $\overline{k_1}$ et $\overline{k_2}$, que l'on note D_K.

En écrivant

$$h_{k_1,k_2}(X) = c(X-a_1)^{t_1}\cdots(X-a_v)^{t_v}$$

avec $v \leqslant 2s$ (les a_i, t_j, c dépendent de k_1 et k_2), on sait que le pgcd des $t_1, \ldots, t_v$ est égal à D_K et $D_K = 0$ si et seulement si $d|k_1$ et $d|k_2$. Lorsque $1 \leqslant D_K \leqslant d-1$, on peut poser

$$\widetilde{h}_{k_1,k_2}(X) = c(X-a_1)^{t_1/D_K}\cdots(X-a_v)^{t_v/D_K}$$

tel que $\widetilde{h}_{k_1,k_2}^{D_K} = h_{k_1,k_2}$ et $(t_1/D_K, \ldots, t_v/D_K) = 1$. On peut donc appliquer la proposition 7 sur $\widetilde{h}_{k_1,k_2}$:

$$\begin{aligned} \sum_{\substack{k_1,k_2=-H\\ k_1,k_2\neq 0}}^{H} \frac{1}{|k_1k_2|}\left|\sum_{n=0}^{M}\chi\big(h_{k_1,k_2}(n)\big)\right| &\leqslant \sum_{\substack{k_1,k_2=-H\\ k_1,k_2\neq 0}}^{H} \frac{1}{|k_1k_2|}\left|\sum_{n=0}^{M}\chi^{D_K}\big(\widetilde{h}_{k_1,k_2}(n)\big)\right| \\ &\leqslant \sum_{\substack{k_1,k_2=-H\\ k_1,k_2\neq 0\\ d|k_1,d|k_2}}^{H} \frac{p}{|k_1k_2|} + \sum_{\substack{k_1,k_2=-H\\ k_1,k_2\neq 0\\ d\nmid k_1,d\nmid k_2}}^{H} \frac{1}{|k_1k_2|}\cdot 4sp^{1/2}\log p \\ &\leqslant p\left(\sum_{\substack{k=1\\ d|k}}^{H}\frac{2}{k}\right)^2 + \big(2(\log H+1)\big)^2\cdot 4sp^{1/2}\log p \end{aligned}$$

Compte tenu de (42), on choisit $H = d-1$ si $d \leqslant p^{1/2}$ et $H = \lfloor p^{1/2} \rfloor$ sinon. Alors, on obtient

$$\begin{aligned}\left|\sum_{\substack{n=0\\ n\notin L}}^{M} \varphi_H\big(\chi(g(n+d_1))\big)\varphi_H\big(\chi(g(n+d_2))\big)\right| &\leqslant \big(2(\log p^{1/2}+1)\big)^2 \cdot 4sp^{1/2}\log p \\ &\leqslant 4sp^{1/2}(\log 8p)^3\end{aligned}$$

et

$$\begin{aligned}\sum_{\substack{n=0\\ n\notin L}}^{M} \psi_H\big(\chi(g(n+d_i))\big) &\leqslant 8sp^{1/2}\log p + 2p/H \\ &\leqslant 8sp^{1/2}\log p + 4p/d.\end{aligned}$$

Donc

$$\begin{aligned}\left|\sum_{\substack{n=0\\ n\notin L}}^{M} e_{n+d_1}e_{n+d_2}\right| &\leqslant \sharp L + \left|\sum_{\substack{n=0\\ n\notin L}}^{M} \varphi_H\big(\chi(g(n+d_1))\big)\varphi_H\big(\chi(g(n+d_2))\big)\right| \\ &\qquad + \sum_{i=1}^{2}\sum_{\substack{n=0\\ n\notin L}}^{M} \psi_H\big(\chi(g(n+d_i))\big) \\ &\leqslant 2s + 4sp^{1/2}(\log 8p)^3 + 16sp^{1/2}\log p + 8p/d \\ &\leqslant 5sp^{1/2}(\log 8p)^3 + 8p/d \qquad (\text{si } p \geqslant 251).\end{aligned}$$

Lorsque $d > p^{1/2}$, on peut remplacer le terme $2p/H$ de (43) par $2p^{1/2}\log p$, ce qui nous permet d'omettre numériquement le terme $8p/d$ à la fin.

Finalement, on obtient l'estimation voulue sur $C_2(E_p)$.

5.3. Le cas $\ell \geqslant 3$. Faute de technique supplémentaire, nous devons supposer que g est irréductible pour estimer les corrélations d'ordre $\geqslant 3$. On remarque que dans ce cas, les racines de g dans $\overline{\mathbb{Z}/p\mathbb{Z}}$ sont conjuguées et aucune de leur différence est dans $\mathbb{Z}/p\mathbb{Z}$ (voir la remarque de la section 5.2). On peut donc reprendre la même idée comme précédemment (voir l'estimation de $C_\ell(E_p)$ dans la section 2) pour démontrer que :

$$C_\ell(E_p) \ll \ell sp^{1/2}(\log 8p)^{\ell+1} + p/d.$$

CHAPITRE 5

Sur une construction utilisant le plus grand facteur premier

Une idée de construction de suites binaires suggérée par Erdős consiste à comparer le nombre de facteurs premiers d'un entier n respectivement congrus à $+1$ ou -1 modulo 4. Notons $\omega^+(n)$ (resp. $\omega^-(n)$) le nombre des facteurs premiers différents de la forme $4k+1$ (resp. $4k-1$) de n. On peut alors poser

$$e_n^+ = \begin{cases} +1, & \text{si } \omega^+(n) \geqslant \omega^-(n), \\ -1, & \text{si } \omega^+(n) < \omega^-(n), \end{cases} \tag{43}$$

et

$$e_n^- = \begin{cases} +1, & \text{si } \omega^+(n) > \omega^-(n), \\ -1, & \text{si } \omega^+(n) \leqslant \omega^-(n). \end{cases} \tag{44}$$

Le problème de ces constructions est qu'on pense que parmi les entiers $n \leqslant \mathbb{N}$, il y en a $\gg \frac{N}{\log_2 N}$ qui vérifient $\omega^+(n) = \omega^-(n)$ (cf. **[16]**). Ainsi, on n'obtiendra jamais de « bonne » suite pseudo-aléatoire de cette manière.

Une autre idée suggérée par A. Sárközy dans le même esprit consiste à considérer la congruence du plus grand facteur de n modulo 4. On pose alors

$$e_n = \begin{cases} +1, & \text{si } P(n) \equiv +1 \bmod 4 \text{ ou } n = 2^k, \\ -1, & \text{si } P(n) \equiv -1 \bmod 4, \end{cases} \tag{45}$$

La contribution des entiers égaux à une puissance de 2 est marginale, et l'équirépartition des nombres premiers dans les classes -1 et $+1$ modulo 4 donnent l'espoir d'obtenir une « bonne » suite. Nous avons démontré dans **[51]** les Théorèmes 14 et 15 suivants :

THÉORÈME 14. *Soit $E_N = (e_i)_{i\in[1,N]}$ la suite définie par (45). Pour tout $A > 0$ fixé, on a*

$$W(E_N) \ll_A \frac{N}{(\log N)^A}.$$

Ce premier résultat est inconditionnel. Au prix d'une hypothèse supplémentaire classique en théorie des nombres, nous pouvons cependant obtenir une estimation plus forte. Soit $q > 2$ un entier. Dans l'étude des fonctions L de Dirichlet, on sait qu'il y a au plus un

zéro exceptionnel pour le produit

$$\prod_{\substack{\chi \bmod q \\ \chi \neq \chi_0}} L(s,\chi)$$

dans la région

$$\left\{\sigma + \imath\tau \in \mathbb{C} : \sigma \geqslant 1 - \frac{c_0}{\log(q(1+|\tau|))}\right\} \tag{46}$$

qui sera un réel (c_0 est une constante absolue). C'est ce que l'on appelle le zéro de Siegel. Le caractère correspondant χ est nécessairement réel et non principal (voir par exemple Davenport **[13]** p.93 : la région décrite dans (6) du Théorème et l'inégalité (7) qui s'en suit, qui est dû à Landau). En général, on pense qu'il n'y a aucun zéro de Siegel quel que soit q. C'est une hypothèse plus faible que l'hypothèse de Riemann. Sous cette hypothèse, on peut améliorer sensiblement l'estimation précédente :

THÉORÈME 15. *Soit $E_N = (e_i)_{i\in[1,N]}$ la suite définie par (45). Si aucun des caractères réels n'admet de zéro de Siegel, alors il existe une constante absolue $c > 0$ telle que*

$$W(E_N) \ll N\,\mathrm{e}^{-c(\log N \log_2 N)^{1/4}}, \tag{47}$$

La constante implicite est aussi absolue.

Les outils principaux pour ces preuves sont des estimations sur les entiers friables. On donnera une preuve qui permet d'obtenir les deux théorèmes précédents car l'idée de preuve reste la même, avec ces lemmes ainsi présentés.

Le fait de construire la suite par (45) en considérant les classes de congruences modulo 4 n'est pas important. On peut très bien construire une suite en distinguant si le plus grand facteur des suites des entiers est congruent à $+1$ ou -1 modulo 3 et on obtient ainsi une simplification de la construction (pas besoin de s'occuper spécialement des entiers égaux à une puissance de 2) mais la conclusion sera la même. On peut aussi considérer les classes de congruence modulo 8 suivant que le plus grand facteur est congruent à un résidu quadratique (la classe de 1 et 7) ou non (la classe de 3 et 5). Là encore, les facteurs premiers pairs jouent un rôle négligeable et on peut les attribuer à n'importe quelle classe dans la définition pour obtenir la même conclusion, mais avec une construction différente. Voir aussi le chapitre 6 pour une illustration numérique de cette discussion.

L'estimation de la corrélation $C_2(E_N)$ (et *a fortiori* de $C_k(E_N)$) semble être trop difficile, donc nous nous contenterons de présenter des résultats numériques dans le chapitre 6.

1. Lemmes préliminaires

Pour $2 \leqslant y \leqslant x$, on note $\Psi(x,y)$ le nombre des entiers naturels $n \leqslant x$ tels que $P(n) \leqslant y$. De la même manière, pour $(a,b) \in \mathbb{N} \times \mathbb{N}^*$ avec $a < b$, on note $\Psi(x,y;a,b)$ le nombre des entiers $n \leqslant x$ congrus à a modulo b tels que $P(n) \leqslant y$. Avec la méthode de Rankin, G. Tenenbaum a démontré une majoration uniforme pour $\Psi(x,y)$:

LEMME 10. *Pour $2 \leqslant y \leqslant x$, on a*

$$\Psi(x,y) \ll x \exp\left(-\frac{\log x}{2\log y}\right), \qquad (2 \leqslant y \leqslant x). \tag{48}$$

Comme la démonstration n'est pas très longue, je l'inclus ici.

DÉMONSTRATION. (cf. **[63]** p. 106-107) On peut supposer $y \geqslant 11$ car dans le cas contraire, on a $\Psi(x,y) \ll (\log x)^4$. Sous cette hypothèse on peut écrire, pour tout $\alpha > 0$,

$$\Psi(x,y) \leqslant x^{3/4} + \sum_{\substack{x^{3/4} < n \leqslant x \\ P(n) \leqslant y}} \left(\frac{n}{x^{3/4}}\right)^{\alpha}.$$

Pour chaque $\alpha \in [0,1]$, on peut définir une fonction multiplicative en posant $f_\alpha(p^\nu) = p^{\alpha\nu}$ lorsque $p \leqslant y$ et $f_\alpha(p^\nu) = 0$ pour $p > y$. Alors, en voyant l'expression de $f * \mu$, on est ramené à définir $g_\alpha(p^\nu) = p^{\alpha\nu}(1 - p^{-\alpha})$ pour $p \leqslant y$ et $g_\alpha(p^\nu) = 0$ pour $p > y$ de sorte que $g_\alpha * 1$ est une fonction multiplicative positive qui vérifie $g_\alpha * 1(n) = f_\alpha(n)$ lorsque $P(n) \leqslant y$

Notons $u = \frac{\log x}{\log y}$ et choisissons $\alpha = \frac{2}{3\log y}$, qui permet d'obtenir

$$x^{-3\alpha/4} = \mathrm{e}^{-\frac{1}{2}u} \quad \text{et} \quad x^{3/4} \leqslant x\,\mathrm{e}^{-\frac{1}{2}u} \quad (\text{puisque } y \geqslant 9).$$

Alors, on a

$$\begin{aligned}
\Psi(x,y) &\leqslant x^{3/4} + x^{-3\alpha/4} \sum_{x^{3/4} < n \leqslant x} f_\alpha(n) \\
&\leqslant x\,\mathrm{e}^{-\frac{1}{2}u} + \mathrm{e}^{-\frac{1}{2}u} \sum_{n \leqslant x} \sum_{d \mid n} g_\alpha(n) \\
&\leqslant x\,\mathrm{e}^{-\frac{1}{2}u} + x\,\mathrm{e}^{-\frac{1}{2}u} \sum_{P(d) \leqslant y} g_\alpha(d)/d.
\end{aligned}$$

Comme g_α est multiplicative, on a

$$\begin{aligned}\sum_{P(d)\leqslant y} g_\alpha(d)/d &\leqslant \prod_{p\leqslant y}\sum_{\nu=0}^{\infty} g_\alpha(p^\nu)/p^\nu \leqslant \prod_{p\leqslant y}\sum_{\nu=0}^{\infty} p^{\nu(\alpha-1)}(1-p^{-\alpha})\\ &\leqslant \prod_{p\leqslant y}\sum_{\nu=0}^{\infty} \alpha\log p\left(\frac{\mathrm{e}^{2/3}}{p}\right)^\nu \ll \exp(\alpha\sum_{p\leqslant y}\frac{\log p}{p})\\ &\ll 1,\end{aligned}$$

où dans la dernière inégalité on a utilisé l'estimation de Mertens. Finalement, on obtient l'estimation voulue. □

Cette estimation élémentaire ne sera pas suffisante dans certaines parties de notre preuve. Avec un travail plus élaboré, il est possible d'obtenir un résultat plus précis :

LEMME 11 (de Bruijn-Tenenbaum). *Notons*

$$Z = \frac{\log x}{\log y}\log\left(1+\frac{y}{\log x}\right) + \frac{y}{\log y}\log\left(1+\frac{\log x}{y}\right), \tag{49}$$

alors on a uniformément pour $2 \leqslant y \leqslant x$

$$\log\Psi(x,y) = Z\left\{1+O\left(\frac{1}{\log y}+\frac{1}{\log_2 2x}\right)\right\}. \tag{50}$$

L'estimation (50) a été obtenue dans [**6**] avec le terme d'erreur supplémentaire

$$O\left(\frac{1}{1+\frac{\log x}{\log y}}\right)$$

dans l'accolade. La formule asymptotique (50) est établie dans [**62**] (p.363).

Le lemme suivant sur $\Psi(x,y;a,b)$ est du type « Siegel-Walfisz » qui permet de donner une estimation intéressante uniformément sur les variables dans un domaine assez large.

On remarque si $d = \mathrm{pgcd}(a,b)$ vérifiant $P(d) \leqslant y$, alors on a

$$\Psi(x,y;a,b) = \Psi(x/d,y;a/d,b/d).$$

On introduit également $\Psi_b(x,y)$ qui désigne le nombre de $n \leqslant x$ qui vérifiant $\mathrm{pgcd}(n,b) = 1$ et $P(n) \leqslant y$.

LEMME 12. *Soit A un réel positif. Il existe des constantes positives c_1, c_2 (dépendent au plus de A) tels que uniformément pour x, y, a, b satisfaisant la condition*

$$\exp\left(c_1(\log_2 x)^2\right) \leqslant y \leqslant x, \quad \mathrm{pgcd}(a,b) = 1, \quad 1 < b \leqslant (\log x)^A,$$

on ait

$$\Psi(x,y;a,b)=\frac{\Psi_b(x,y)}{\varphi(b)}\left(1+O(\exp(-c_2\sqrt{\log y}))\right) \tag{51}$$

Si aucun des caractères réels n'admet de zéro de Siegel, alors la même estimation est valide pour le domaine $1<b\leqslant\exp(c_2\sqrt{\log y})$.

Ce résultat est dû à É. Fouvry et G. Tenenbaum. Le lemme précédent résume le Théorème 5 dans [**19**] décrit dans (1.28) p. 455. (voir aussi la formule (1.16)). La validité de cette estimation pour la région $1<b\leqslant\exp(c_2\sqrt{\log y})$ est une conséquence de (1.26) qui permet d'obtenir ce même théorème.

Dans la suite, on fixe A un réel positif et désigne par c_i des constantes positives qui dépendent au plus de A.

Les lemmes suivants seront utiles pour démontrer les Théorèmes 14 et 15 en même temps. Ils ne sont pas présentés sous une forme optimale.

LEMME 13. *Pour tous réels* $x\geqslant 2$, $\lambda>0$ *et tous entier* $b\in[2,x]$, *on a*

$$\sum_{\substack{p\mid b\\ p>\exp(\lambda(\log x)^{4/7})}}\Psi\left(\frac{x}{p},p\right)\ll x\ (\log x)\exp\left(-\lambda(\log x)^{4/7}\right).$$

DÉMONSTRATION. Il suffit d'écrire trivialement $\Psi(x/p,p)\leqslant x/p$ et de constater le nombre de nombres premiers divisant b est $\ll\log x$. □

LEMME 14. *Pour tous réels* $x\geqslant 10$, $\frac{1}{2}\leqslant\lambda\leqslant\frac{1}{\sqrt{2}}$, *on a*

$$\sum_{p\leqslant\exp(\lambda(\log x)^{4/7})}\Psi\left(\frac{x}{p},p\right)\ll x\ (\log\log x)\exp\left(-\frac{(\log x)^{3/7}}{2\lambda}\right).$$

DÉMONSTRATION. Les restrictions sur λ and x nous garantissent que $2\leqslant p\leqslant x/p$ lorsque $p\leqslant\exp(\lambda(\log x)^{4/7})$. Cette même condition implique également

$$\frac{\log x/p}{\log p}=\frac{\log x}{\log p}-1\geqslant\frac{(\log x)^{3/7}}{\lambda}-1.$$

Ainsi, d'après le lemme 10, on a

$$\Psi\left(\frac{x}{p},p\right)\ll\frac{x}{p}\exp\left(-\frac{(\log x)^{3/7}}{2\lambda}\right).$$

L'estimation classique sur les nombres premiers $\sum_{p\leqslant x} p^{-1} = O(\log\log x)$, nous donne alors

$$\sum_{p\leqslant \exp(\lambda(\log x)^{4/7})} \Psi\left(\frac{x}{p}, p\right) \ll x(\log\log x)\exp\left(-\frac{(\log x)^{3/7}}{2\lambda}\right).$$

□

LEMME 15. *Pour les réels* $\frac{1}{2} \leqslant c \leqslant \frac{1}{\sqrt{3}}$ *et* $2 \leqslant b \leqslant x$, *on a*

$$\sum_{p\,|\,b} \Psi\left(\frac{x}{p}, p\right) \ll x\ \exp(-c(\log x)^{3/7}).$$

DÉMONSTRATION. Il suffit de diviser la somme suivant que $\log p \leqslant \lambda(\log x)^{4/7}$ ou $\log p > \lambda(\log x)^{4/7}$ puis appliquer ces deux derniers lemmes avec $\lambda = 1/\sqrt{2}$. □

2. Transformation du problème

Soient $a, b, M \in \mathbb{N}$ tels que $0 \leqslant a < b \leqslant M$. Il suffit d'estimer la somme

$$\sum_{\substack{1\leqslant n\leqslant M\\ n\equiv a \bmod b}} e_n.$$

À cette fin, on écrit $n = pm$ avec p premier et $P(m) \leqslant p$. On constate que $e_{pm} = e_p$, donc la somme ci-dessus est égale à

$$\sum_{p\leqslant M} e_p \sum_{\substack{1\leqslant m\leqslant M/p\\ P(m)\leqslant p\\ pm\equiv a \bmod b}} 1. \tag{52}$$

Soit $\frac{1}{2} \leqslant c_3 \leqslant \frac{1}{\sqrt{3}}$. La contribution des termes pour $p \leqslant z := \exp(c_3(\log M)^{4/7})$ est facile à estimer :

$$\left|\sum_{p\leqslant z} e_p \sum_{\substack{1\leqslant m\leqslant M/p\\ P(m)\leqslant p\\ pm\equiv a \bmod b}} 1\right| \leqslant \sum_{p\leqslant z} \Psi\left(M/p, p\right) \ll M\exp(-c_3(\log M)^{3/7}),$$

où la dernière majoration s'obtient grâce au lemme 14 avec $\lambda = 1/\sqrt{2}$. On peut supposer désormais

$$p > z = \exp(c_3(\log M)^{4/7}). \tag{53}$$

Puisque p est premier, on a soit $p \mid b$ ou $(p,b)=1$. Pour le premier cas, la condition de divisibilité est tellement forte qu'on peut enlever la condition $pm \equiv a \bmod b$ et on écrit

$$\left| \sum_{\substack{z<p\leqslant M \\ p\mid b}} e_p \sum_{\substack{1\leqslant m\leqslant M/p \\ P(m)\leqslant p \\ pm\equiv a \bmod b}} 1 \right| \leqslant \sum_{p\mid b} \Psi\left(M/p,p\right) \ll M\exp(-c_3(\log M)^{3/7}),$$

où la dernière majoration s'obtient grâce au lemme 15.

Lorsque $(p,b)=1$, on note $\overline{p} \in [1,b-1]$ l'inverse de $p \bmod b$. Alors $pm \equiv a \bmod b$ équivaut à $m \equiv a\overline{p} \bmod b$, et on a

$$\sum_{\substack{1\leqslant n\leqslant M \\ n\equiv a \bmod b}} e_n = \sum_{\substack{z<p\leqslant M \\ \operatorname{pgcd}(p,b)=1}} e_p\, \Psi\left(M/p,p;a\overline{p},b\right) + O(M\exp(-c_3(\log M)^{3/7})). \tag{54}$$

Soit $d=(a,b)$. D'après $(p,b)=1$, on déduit que $(\overline{p},b)=1$ et

$$\operatorname{pgcd}(a\overline{p}/d, b/d) = 1. \tag{55}$$

Si $P(d)>p$, alors $\Psi\left(M/p,p;a\overline{p},b\right)=0$. Sinon, on a

$$\Psi\left(M/p,p;a\overline{p},b\right) = \Psi\left(\frac{M}{dp},p;a\overline{p}/d,b/d\right).$$

D'ailleurs, pour M suffisamment grand, d'après (53), on a

$$p \geqslant \exp(c_1(\log_2 M)^2). \tag{56}$$

Premièrement, supposons que

$$2 \leqslant b \leqslant (\log M)^A, \tag{57}$$

Comme les conditions (55) et (56) sont satisfaites, on peut appliquer (51) pour obtenir

$$\begin{aligned} \sum_{\substack{1\leqslant n\leqslant M \\ n\equiv a \bmod b}} e_n &= \frac{1}{\varphi(b/d)} \sum_{\substack{z<p\leqslant M \\ \operatorname{pgcd}(p,b)=1}} e_p\, \Psi_{b/d}\left(\frac{M}{dp},p\right)\left(1+O(\exp(-c_2\sqrt{\log z}))\right) \\ &\quad + O(M\exp(-c_3(\log M)^{3/7})) \\ &= \frac{1}{\varphi(b/d)} \sum_{\substack{z<p\leqslant M \\ \operatorname{pgcd}(p,b)=1}} e_p\, \Psi_{b/d}\left(\frac{M}{dp},p\right) + O(M\,\exp(-c_4(\log M)^{2/7})). \end{aligned} \tag{58}$$

Deuxièmement, on suppose que pour

$$(\log M)^A \leqslant b \leqslant \exp(c_5(\log M)^{2/7}), \tag{59}$$

le zéro de Siegel n'existe pas pour aucun $L(s,\chi)$ où χ désigne un réel caractère modulo b.

Lorsque c_5 est suffisamment petit, par exemple $c_5 = c_2c_3$, alors d'après (53) et (59), on obtient la condition $b \leqslant \exp(c_2\sqrt{\log p})$. Par conséquent, l'estimation est encore valide puisque on peut encore appliquer (51).

À ce stade, on peut enlever la condition $(p, b) = 1$. Puisque d'après le lemme 15, on a

$$\sum_{p\,|\,b} \Psi_{b/d}\left(\frac{M}{dp}, p\right) \leqslant \sum_{p\,|\,b} \Psi\left(\frac{M}{dp}, p\right) \ll M \exp(-c_3(\log M)^{3/7}).$$

Finalement, nous avons obtenu soit sous l'hypothèse (57), soit sous l'hypothèse (59) avec une condition supplémentaire sur la non-existence du zéro de Siegel, la formule suivante :

$$\sum_{\substack{1\leqslant n\leqslant M\\ n\equiv a \bmod b}} e_n = \frac{1}{\varphi(b/d)} \sum_{z<p\leqslant M} e_p\ \Psi_{b/d}\left(\frac{M}{dp}, p\right) + O(M\ \exp(-c_4(\log M)^{2/7})). \tag{60}$$

Bien que nous soyons obligé d'imposer différentes conditions sur b comme (57) et (59) à cause notre manque de connaissance sur la répartition des nombres premiers en progression arithmétiques, nous pouvons poursuivre notre estimation sur la bonne distribution grâce à cette formule.

3. Estimation du terme principal

L'idée primordiale est d'utiliser la compensation des changements de signe de e_p dans le terme principal de la formule (60). À cette fin, nous réécrivons (60)

$$\sum_{z<p\leqslant M} e_p\ \Psi_{b/d}\left(\frac{M}{dp}, p\right) = \sum_{z<p\leqslant M} e_p \sum_{\substack{m\leqslant \frac{M}{dp}\\ P(m)\leqslant p\\ \mathrm{pgcd}(m,b/d)=1}} 1 = \sum_{\substack{m\\ mP(m)\leqslant M/d\\ \mathrm{pgcd}(m,b/d)=1}} \sum_{\substack{\max(z,P(m))<p\\ p\leqslant M/dm}} e_p. \tag{61}$$

D'après une version simple de l'estimation des nombres premiers en progression arithmétique (voir aussi (11) p. 123 du livre de Davenport [**13**]), il existe une constante absolue c_6 telle que pour tout réel $x \geqslant 2$, on ait

$$\sum_{\substack{p\leqslant x\\ p\equiv 1 \bmod 4}} 1 = \tfrac{1}{2}\,\mathrm{Li}(x) + O(x\exp(-c_6\sqrt{\log x})),$$

$$\sum_{\substack{p\leqslant x\\ p\equiv 3 \bmod 4}} 1 = \tfrac{1}{2}\,\mathrm{Li}(x) + O(x\exp(-c_6\sqrt{\log x})).$$

Donc, pour tout réel $x \geqslant 1$, on a

$$\sum_{p\leqslant x} e_p \ll x\exp(-c_6\sqrt{\log x}),$$

En reportant dans (61), on obtient

$$\begin{aligned}\sum_{z<p\leqslant M} e_p\,\Psi_{b/d}\left(\frac{M}{dp},p\right) &\ll \sum_{\substack{m\\ mP(m)\leqslant M/d}} \frac{M}{dm}\exp\left(-c_6\sqrt{\log\frac{M}{dm}}\right)\\ &\ll \sum_{mP(m)\leqslant M}\frac{M}{m}\exp\left(-c_6\sqrt{\log\frac{M}{m}}\right).\end{aligned}$$

Lorsque $b=1$, nous pouvons écrire

$$\sum_{1\leqslant n\leqslant M} e_n = \sum_{\substack{1\leqslant m\leqslant M/2\\ P(m)m\leqslant M}}\sum_{\substack{P(m)\leqslant p\leqslant M\\ p\equiv 1 \bmod 4}} 1 - \sum_{\substack{1\leqslant m\leqslant M/2\\ P(m)m\leqslant M}}\sum_{\substack{P(m)\leqslant p\leqslant M\\ p\equiv -1 \bmod 4}} 1 + \sum_{\substack{1\leqslant m\leqslant M/2\\ P(m)=2}} 1$$

pour obtenir la compensation de signes et on majore ce dernier terme par $\log M$.

Donc, soit pour b dans cette région

$$1\leqslant b\leqslant (\log M)^A,$$

soit pour une région plus large qui inclus celle de (59) mais avec l'hypothèse supplémentaire sur la non-existence de zéro de Siegel, nous avons l'estimation suivante :

$$\sum_{\substack{1\leqslant n\leqslant M\\ n\equiv a \bmod b}} e_n \ll \frac{1}{\varphi(b/d)}\sum_{mP(m)\leqslant M}\frac{M}{m}\exp\left(-c_6\sqrt{\log\frac{M}{m}}\right) + M\exp\left(-c_5(\log M)^{2/7}\right). \tag{62}$$

Soit $y\geqslant 1$ un paramètre à déterminer ultérieurement. On divise la somme du membre droit de l'inégalité ci-dessus suivant que $M/m\leqslant y$ ou $M/m>y$. Alors

$$\begin{aligned}\sum_{mP(m)\leqslant M}\frac{M}{m}\exp\left(-c_6\sqrt{\log\frac{M}{m}}\right) &\leqslant \sum_{P(m)\leqslant M/m\leqslant y} y + \sum_{\substack{M/m>y\\ mP(m)\leqslant M}}\frac{M}{m}\exp\left(-c_6\sqrt{\log y}\right)\\ &\leqslant y\Psi(M,y) + M(\log M)\exp\left(-c_6\sqrt{\log y}\right) \qquad (63)\end{aligned}$$

Nous allons utiliser le lemme 11. Dans l'expression de Z, appliqué avec $x=M$, si y croît suffisamment rapidement de sorte que $\frac{y}{\log M}\to\infty$ lorsque $M\to\infty$, on a

$$\begin{aligned}Z &= \frac{\log M}{\log y}\left(\log\left(1+\frac{y}{\log M}\right)+\frac{y}{\log M}\log\left(1+\frac{\log M}{y}\right)\right)\\ &\ll \frac{\log M}{\log y}\left(\log y - \log_2 M + C\right)\end{aligned}$$

où $C>1$ est une constante fixe.

Pour optimiser l'estimation de (63), avec la considération de (50), on choisit alors $y = \exp(\frac{1}{2}\sqrt{\log M \log_2 M})$, ce qui permet d'obtenir lorsque M est suffisamment grand

$$\begin{aligned}\sum_{mP(m)\leqslant M} \frac{M}{m} \exp\left(-c_6\sqrt{\log\frac{M}{m}}\right) &\leqslant M\exp\left(-\frac{1}{2}\sqrt{\log M}\left(\sqrt{\log_2 M} - \frac{C}{\sqrt{\log_2 M}}\right)\right)\\ &\quad + M(\log M)\exp\left(-c_6'(\log M\log_2 M)^{1/4}\right)\\ &\leqslant M\exp\left(-c_7(\log M\log_2 M)^{1/4}\right).\end{aligned}$$

En reportant cette majoration dans (62), on en déduit

$$\sum_{\substack{1\leqslant n\leqslant M\\ n\equiv a \bmod b}} e_n \ll M\ \exp(-c_8(\log M\log_2 M)^{1/4}), \tag{64}$$

Si $b > (\log M)^A$, alors en comptant le nombre de termes, on a trivialement

$$\sum_{\substack{1\leqslant n\leqslant M\\ n\equiv a \bmod b}} e_n \ll \frac{M}{(\log M)^A}$$

et la conclusion du Théorème 14 s'ensuit avec (64).

Si nous supposons que la non-existence de zéro de Siegel quel que soit le caractère réel, alors nous avons l'estimation (64) dans la région décrit par (59). Dans ce cas, lorsque $b > \exp(c_5(\log M)^{2/7})$, nous avons trivialement

$$\sum_{\substack{1\leqslant n\leqslant M\\ n\equiv a \bmod b}} e_n \ll M\ \exp(-c_5(\log M)^{2/7}).$$

Finalement, nous obtenons la conclusion du Théorème 15.

CHAPITRE 6

Les résultats numériques

Nous donnons dans ce chapitre quelques résultats numériques relatifs aux chapitres 4 et 5. D'une part, cela nous permet de mieux comprendre les résultats théoriques obtenus et leurs possibles améliorations. D'autre part, pour les questions qui sont encore hors de portée, nous obtenons une esquisse des réponses attendues.

1. Sur les constructions inspirées d'Erdős

Regardons d'abord la suite E_N définie par (45) dans le chapitre 5. Cette suite s'obtient facilement en criblant par les nombres premiers un tableau constitué d'entiers. Ensuite, les algorithmes qui permettent de calculer la mesure de bonne distribution et sur la mesure de corrélation d'ordre 2 s'exécutent avec une complexité $O(N^2)$ (où N désigne la longueur de la suite). Les graphes reproduits dans les pages suivantes ont été réalisés grâce au logiciel *gnuplot*.

La figure 1 compare l'allure de la courbe $W(E_N)$ et celle de $S(E_N)$ où $S(E_N)$ désigne la fonction $S(n)$ de sommation

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

On a introduit un signe négatif pour que la courbe présentée dans cette figure soit du même côté que $W(E_N)$. En effet, outre que les 50 premiers valeurs, la somme des e_i est toujours négative dans l'intervalle représenté sur ce graphe. On constate alors une prépondérance apparente des plus grands facteurs des entiers naturels qui sont congruents à $-1 \bmod 4$ sur ceux qui sont congruents à $1 \bmod 4$. En fait, ce type de phénomène a été observé par Tchebychev en 1853 sur la distribution des nombres premiers. En admettant quelques hypothèses supplémentaires (voir plus de détails dans **[55]**), on peut estimer la proportion de certaines classes de congruences par rapport aux autres (au sens de densité logarithmique). Plus précisément, soit $q \in \{4,\ p^\alpha,\ 2p^\alpha\}$ avec p un nombre premier impair et α un entier

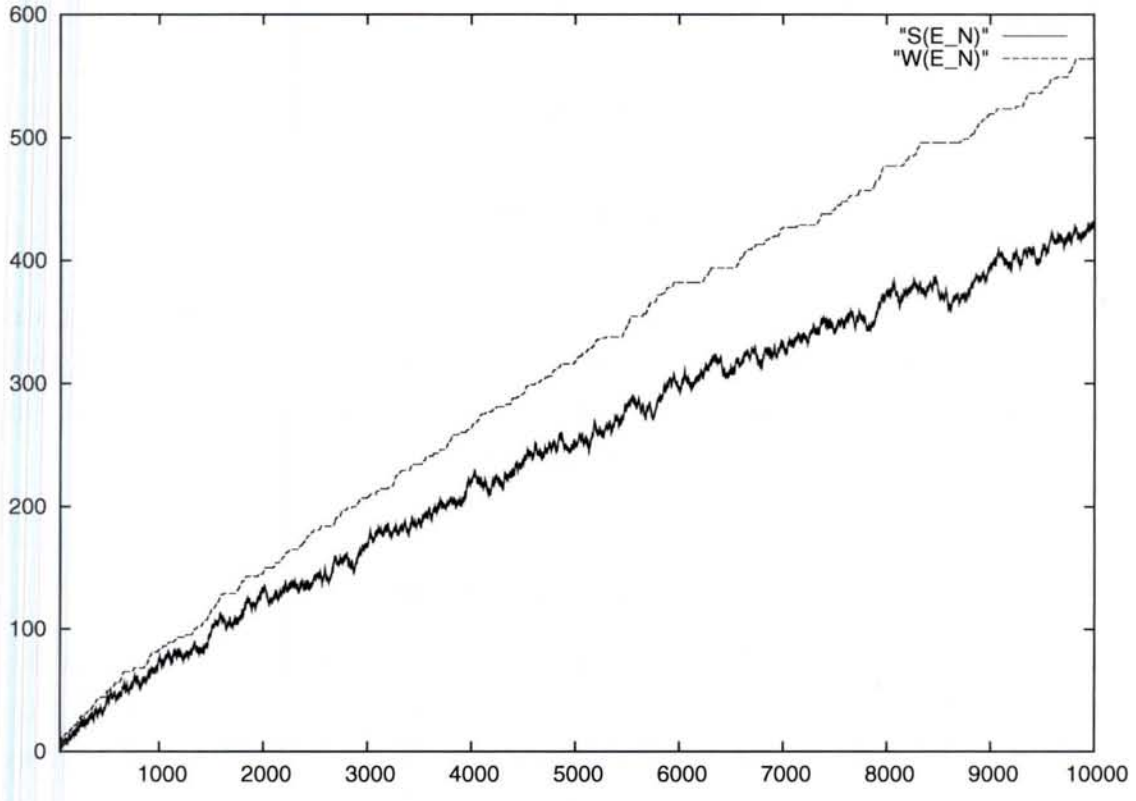

FIG. 1. Graphe de $S(E_N)$ et de $W(E_N)$

positif. On définit les ensembles suivants :

$$\begin{aligned} P_{q;N,R} &= \{n \geqslant 2 : \pi_N(n,q) > \pi_R(n,q)\}, \\ P_{q;R,N} &= \{n \geqslant 2 : \pi_R(n,q) > \pi_N(n,q)\}, \end{aligned}$$

où $\pi_R(n,q)$ (resp. $\pi_N(n,q)$) désigne le nombre de résidus quadratiques (resp. résidus non quadratiques) premiers $p \leqslant n$ modulo q. On appelle *hypothèse de grande simplicité* l'hypothèse selon laquelle pour tout caractère primitif χ de Dirichlet donné, les abscisses τ des zéros de χ sur la droite verticale $\sigma = 1/2$ du plan complexe sont linéairement indépendantes sur $\mathbb{Q}$. Sous cette hypothèse et celle de Riemann généralisée, on peut alors montrer qu'il y a toujours un biais dans la densité logarithmique de $P_{q;N,R}$ par rapport à celle de $P_{q;R,N}$. (Voir **[55]**.)

D'après un résultat de Littlewood, on sait que les deux ensembles $P_{q;N,R}$ et $P_{q;R,N}$ sont infinis. Cependant, trouver un élément explicite peut demander beaucoup de patience. Pour $q = 4$, Leech a trouvé le premier entier appartenant à $P_{4,1,3}$ qui est 26861. Tandis que pour

$P_{3,1,2}$, le premier entier trouvé dépasse 600 milliard. Pour le graphe de $S(E_N)$ (figure 1) qui montre un fort penchant vers les facteurs de la forme $4k+3$, à partir de 50, je n'ai trouvé aucun retour vers l'axe horizontal parmi les 500 millions premiers entiers suivants.

La figure 2 permet de comparer la courbe de $W(E_N)$ et celle de la fonction $N^{0,68}$. On voit que $W(E_N)$ semble dévier de plus en plus de $\sqrt{N}$ et on pourrait se demander si $W(E_N)$ croît plus vite que N^α pour tout $\alpha < 1$. La figure 3 compare la courbe $C_2(E_N)$ et celle de la fonction $N^{0,82}$. Cette fois-ci, la croissance de $C_2(E_N)$ est beaucoup plus rapide et il y a moins d'intervalle stationnaire.

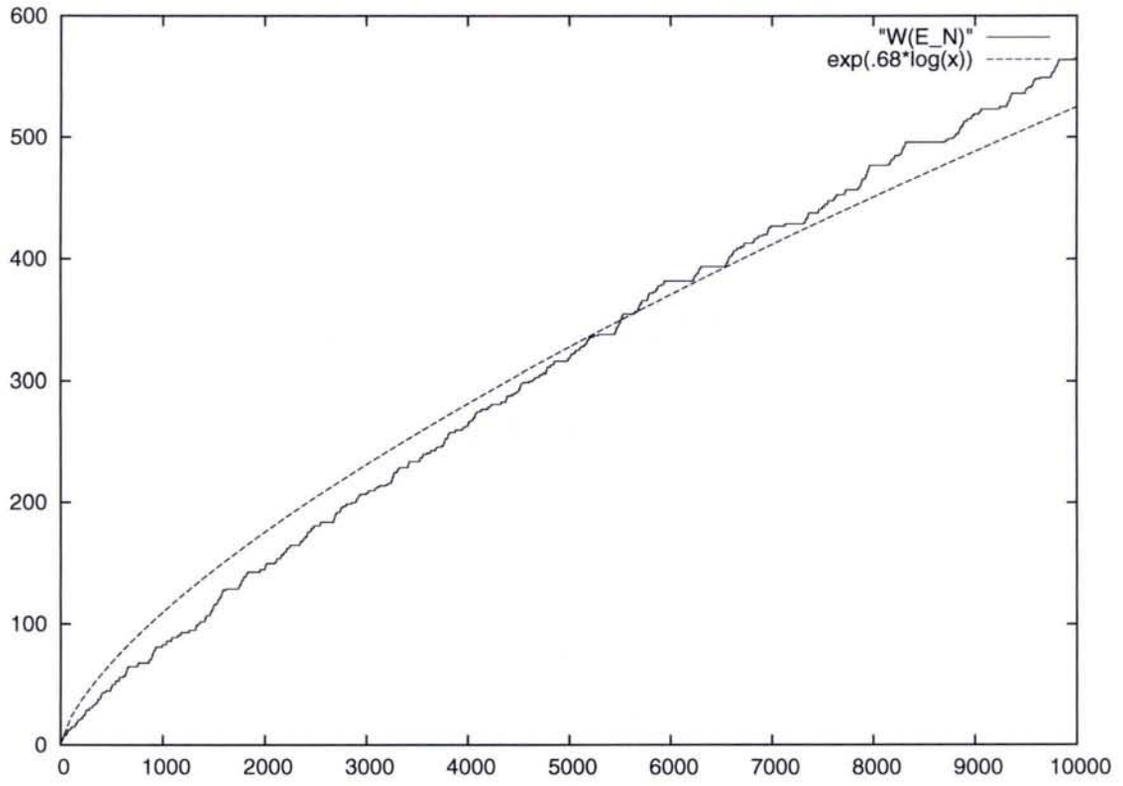

FIG. 2. Graphe de $W(E_N)$ et de $N^{0.68}$.

Pour les valeurs plus grandes que 10000, il devient laborieux de constituer un tableau de chiffres afin d'utiliser *gnuplot* pour dessiner une courbe. En effet, le calcul requiert plus de temps, surtout pour celui de $C_3(E_N)$ qui nécessite trois boucles. C'est la raison pour laquelle on se limite à la corrélation de termes consécutifs, restriction qui est signalée par

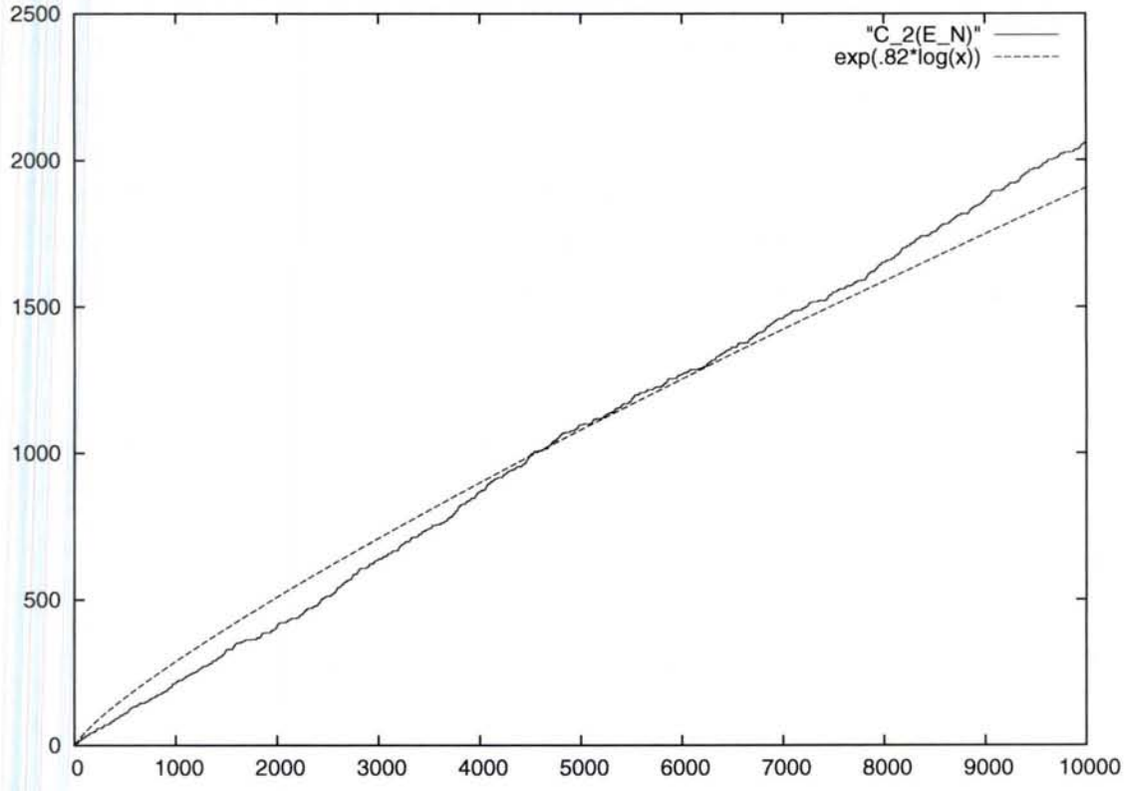

FIG. 3. Graphe de $C_2(E_N)$ et de $N^{0.82}$.

un tilde :

$$\tilde{C}_2(E_N) = \max_M \left| \sum_{n=1}^{M} e_n e_{n+1} \right|,$$

$$\tilde{C}_3(E_N) = \max_M \left| \sum_{n=1}^{M} e_n e_{n+1} e_{n+2} \right|,$$

$$\tilde{C}_4(E_N) = \max_M \left| \sum_{n=1}^{M} e_n e_{n+1} e_{n+2} e_{n+3} \right|,$$

Cependant, en lisant le tableau 1, on constate en comparant notamment $C_3(10000)$ et $\tilde{C}_3(10000)$ que ces corrélations ont un comportement très différents de celui des corrélations originales.

Pour les constructions suggérées par Erdős définies par (43) et (44), des méthodes de programmation similaires nous permettent d'obtenir le tableau 2. Les valeurs des mesures W et C_2 ont tendances de croître plus rapidement que pour la construction (45). De plus,

N	$W(E_N)$	$C_2(E_N)$	$\tilde{C}_2(E_N)$	$C_3(E_N)$	$\tilde{C}_3(E_N)$	$\tilde{C}_4(E_N)$
10	3	4	4	6	4	3
100	14	27	13	34	14	10
1000	83	215	46	165	49	37
10000	565	2058	126	1034	178	184
100000	4252	17887	416	−−	516	528
1000000	33719	149360	2303	−−	930	1601

TAB. 1. Valeurs numériques pour la construction (45).

N	$W(E_N^+)$	$C_2(E_N^+)$	$W(E_N^-)$	$C_2(E_N^-)$
10	3	4	7	5
100	20	24	47	49
1000	142	162	453	547
10000	1062	1871	4371	4951
100000	9658	20339	41709	44001
1000000	94334	214432	403333	409696

TAB. 2. Valeurs numériques pour les constructions (43) et (44).

les valeurs de $W(E_N^-)$ semblent croître quatre fois plus rapidement que celles de $W(E_N^+)$, tandis que les valeurs de $C_2(E_N^-)$ augmentent deux fois plus vite que celles de $C_2(E_N^+)$. Il y a aussi un biais sur le nombre des facteurs premiers de la forme $4k+3$.

2. Sur la coïncidence des suites basées sur les caractères

Regardons maintenant numériquement la performance du Théorème 11. Soient p un nombre premier impair et f, g deux polynômes à coefficients dans $\mathbb{Z}/p\mathbb{Z}$ de degrés k et l respectivement. Combien de coïncidences consécutives doit-on observer sur les deux suites $\left(\frac{f(n)}{p}\right)$ et $\left(\frac{g(n)}{p}\right)$ en pratique pour conclure à la coïncidence totale ? On aurait envie de faire une comparaison sur une famille de polynômes définies par (11).

Représentons une telle suite par un tableau de bits de longueur p. Si on considère tous les polynômes de degré $\leqslant d$, alors le nombre de suites à comparer est p^{d+1}. Comme on doit comparer des couples de suites, il nous faudra procéder à p^{2d+2} comparaisons. Or, pour entrer dans le cadre non trivial du Théorème 11, on est ramené à considérer p tel que $3(k+l)p^{1/2}\log p < p$. Pour $k = l = 2$, on a $p \geqslant 12907$, ce qui nécessite alors au minimum $12907^6 \simeq 10^{25}$ opérations. Supposons qu'on puisse remplacer la constante 3 (dans

l'expression $3(k+l)p^{1/2}\log p)$ par 1, il nous faudrait quand même au moins $701^6 \simeq 10^{17}$ opérations, ce qui est encore irréalisable.

Essayons d'affiner nos investigations. On remarque que si l'on ne se restreint pas à considérer des polynômes unitaires, on obtiendra trop de listes identiques. En effet si $g = cf$ avec c un résidu quadratique modulo p, alors on obtient toujours une coïncidence totale et cela nous empêche de trouver facilement les polynômes non triviaux dans notre comparaison. Pour réduire ces redondances et augmenter la visibilité dans notre test, il suffit donc de considérer deux classes de polynômes : les polynômes unitaires et les polynômes qui sont le produit d'un résidu non quadratique et d'un polynôme unitaire. Cela nous permet aussi de diviser le nombre des polynômes par $p/2$. Pour simplifier encore la tâche, on va considérer les polynômes unitaires et obtenir une autre liste par une opération `XOR` (« ou exclusif » sur les bits), quitte à modifier la construction proposée dans (11). En effet, on obtiendra des valeurs différentes sur les zéros éventuels dans $\mathbb{Z}/p\mathbb{Z}$ de ces polynômes, mais cela ne constitue pas un grand changement dans la preuve du Théorème 11. Les polynômes qui sont le produit de polynômes irréductibles de degré > 1 donneront toutefois la même suite.

C'est dans cette optique que nous avons écrit un programme en *C++* pour le test. Pour des raisons techniques d'efficacité et à cause de la complexité de l'algorithme, on se limite à tester pour $p < 64$. On représente la suite par un type d'entiers très longs de 64 bits. Comme on a toujours le problème de complexité et la limitation du Théorème 11, on ne considère que les polynômes de degré $\leqslant 3$. D'ailleurs, on doit aussi exclure les polynômes divisibles par un carré d'un autre polynôme pour éviter la redondance dans la comparaison. (Par exemple $X^2(X+2)$ et $(X+1)^2(X+2)$.) Cette exclusion n'est pas gênante en tenant compte de la modification de la construction mentionnée dans le paragraphe précédent. (Pour le même exemple, on peut tout simplement considérer $X+2$, un polynôme de degré < 3 qui est déjà représenté dans la liste.) Comme on ne considère que les polynômes de degré $\leqslant 3$, ces polynômes ne sont pas difficile à reconnaître puisqu'il s'agit des polynômes admettant une racine multiple dans $\mathbb{Z}/p\mathbb{Z}$.

Donnons la nouvelle définition de cette construction. Soient p un nombre premier et f un polynôme de degré $d \leqslant 3$, à coefficient dans $\mathbb{Z}/p\mathbb{Z}$ et ne possède pas de racine multiple. On considère la suite associée $E_N(f) = (e_n(f))_{n\in[1,p]}$ définies de la manière suivante : si n

n'est pas une racine de f dans $\mathbb{Z}/p\mathbb{Z}$, on pose

$$e_n(f) = \left(\frac{f(n)}{p}\right) \tag{65}$$

sinon, $e_n(f)$ prend la valeur du symbole de Legendre du coefficient dominant de f.

Après la modélisation d'une telle suite par un type d'entiers de 64 bits, le schéma de programmation est classique, on construit un tableau de pointeur sur ce type et on initialise le tableau. On remarque que pour les polynômes qui ne diffèrent que par une constante, les listes correspondantes ne diffèrent que par une permutation cyclique, ce qui est réalisé immédiatement par quelques opérations de « shift » des bits sans calcul supplémentaire. Le principal temps d'exécution réside dans le parcours pour la comparaison. Lorsqu'il y a une coïncidence des sous-suites qui dépasse un seuil de longueur donné par l'utilisateur, le programme affiche alors les deux polynômes et la liste correspondante. Le lecteur intéressé pourra consulter ma page web `http://www.iecn.u-nancy.fr/~oon`.

Si la valeur de p est trop petite, on peut trouver des polynômes qui ne sont carré d'aucun polynôme mais donnent une liste des résidus quadratiques. Par exemple, pour $p = 5$, le polynôme X^3+3X donne toujours des valeurs de résidus quadratiques. En fait, on est dans le cadre trivial de la proposition 2. Dans notre cas, nous obtenons des résultats non triviaux à partir de $p = 19$. Pour $p = 19$, les deux polynômes $X^3+4X+7 = (X+13)(X+14)(X+11)$ et $c(X^3+18X^2+17X+13) = c(X+9)(X^2+9X+12)$ (où c est un résidu non quadratique) donnent deux listes qui ne diffèrent que par la dernière valeur.

Le tableau 3 suivant est obtenu par le programme précédent. L'entier d est le plus grand degré des polynômes considérés et s désigne le seuil de sous-suites coïncidentes qu'on doit observer pour conclure que le polynôme associé à cette liste est l'unique représentant (à une constante multiplicative près qui est un résidu quadratique modulo p). On remarque que s ne donne pas une suite croissante et le seuil proposé dans le Théorème 11 pourrait être amélioré puisqu'on constate des coïncidences non triviales.

p premier	d degré	s seuil	p premier	d degré	s seuil
17	2	14	41	2	21
	3	–		3	28
19	2	15	43	2	21
	3	19		3	29
23	2	15	47	2	21
	3	22		3	31
29	2	18	53	2	20
	3	24		3	33
31	2	19	59	2	23
	3	25		3	34
37	2	19	61	2	21
	3	29		3	35

TAB. 3. Quelques valeurs de seuils sur les petites valeurs de p

ANNEXE A

Cryptologie

La cryptologie est la science qui étudie la confection et l'analyse de messages secrets. Ces deux aspects s'appellent respectivement la *cryptographie* et la *cryptanalyse*. La cryptographie s'est peu à peu orientée vers l'étude des techniques mathématiques ainsi que des méthodes pratiques qui permettent d'assurer la sécurité de la transmission d'informations. La cryptanalyse s'est développée parallèlement dans le but de déceler les failles éventuelles des systèmes cryptographiques envisagés.

La cryptologie possède une longue histoire passionnante. Dès le développement de la communication orale, le besoin de confidentialité est apparu. Avec le développement de l'écriture et la capacité qui en résulte de transmettre un message à distance, la mise en œuvre de la confidentialité a nécessité des techniques nouvelles. Longtemps liée au milieu diplomatique et militaire, la cryptologie était considérée comme une « science du secret », voire « science secrète ».

Un changement radical est survenu en 1976 lorsque Diffie et Hellman [**14**] ont publié un article intitulé « New Directions In Cryptography », dans lequel ils ont introduit la notion de *clé publique* et ont basé la sécurité sur la difficulté de résoudre rapidement un problème mathématique. Cette même année, l'ancêtre du National Institut of Standards and Technology (NIST) a adopté le protocole **DES** (Data Encryption Standard) comme le système standard de cryptage des données, à l'issue d'un appel d'offre lancé trois ans auparavant. Ce système a été très utilisé pendant 25 ans et le reste encore aujourd'hui, bien qu'il soit en voie de remplacement actuellement.

En 1978, Rivest , Shamir et Adleman ont proposé une méthode pratique de chiffrement à clé publique, connue sous le nom de **RSA** , et aujourd'hui très largement utilisée. Sa sécurité est basée sur la difficulté de la factorisation des nombres entiers.

Un principe fondamental de la cryptographie moderne a été énoncé par *Kerckhoffs* au XIXème siècle : la sécurité du système ne doit reposer que sur la confidentialité de la clé, le reste étant supposé connu de tous. L'idée est qu'il est impossible d'empêcher un

adversaire potentiel (espionnage ou conquête) de se procurer le mécanisme permettant de chiffrer ou déchiffrer, et qu'il ne sert donc à rien de le tenir secret. De plus, la publication du mécanisme permet son étude par un nombre important d'experts, et pas seulement par des adversaires mal intentionnés, ce qui permet de déceler beaucoup plus rapidement des failles éventuelles.

1. Quelques notions fondamentales de cryptographie

L'utilisation de la cryptographie est en général motivée par les objectifs suivants (qui ne sont pas totalement indépendants les uns des autres) :

- **La Confidentialité.** On demande de n'autoriser l'accès à l'information qu'à ceux qui possèdent un permis. La réalisation peut se faire soit par une protection physique, soit par un algorithme informatique.
- **L'intégrité des données.** Il s'agit de la préservation de l'information et de la capacité de déceler la manipulation non autorisée sur l'information transmise, comme insertion, suppression ou substitution.
- **L'authentification.** Elle se présente sous deux aspects. L'authentification de l'individu, appelée identification, consiste à vérifier une identité au moyen par exemple d'une empreinte, d'un mot de passe ou d'une carte d'accès. L'authentification de l'information elle-même consiste à s'assurer de sa provenance.
- **La non-répudiation.** C'est un service qui confirme l'identité impliquée et témoigne de l'action effectuée. La *signature digitale* en fait partie. Elle peut impliquer une tierce partie pour résoudre une dispute éventuelle.

1.1. Chiffrement, Déchiffrement, Protocole. Nous avons besoin de clarifier certaines notions pour la suite de l'exposé, mais nous n'entrerons pas dans un formalisme rigoureux. Un Schéma classique dans la cryptographie est : Une *entité* (par exemple une personne, une entreprise, etc...) veut envoyer un *message* à une autre par un *canal* de communication non sécurisé, et ne souhaite pas un quelconque *adversaire* puisse intercepter le contenu. Nous allons préciser le vocabulaire usuel. On fixe un ensemble fini $\mathcal{A}$ appelé l'*alphabet*.

- L'ensemble des messages $\mathcal{M}$ est une partie de l'ensemble des mots (finis) sur $\mathcal{A}$.
- L'ensemble des textes chiffrés $\mathcal{C}$ est aussi une partie des mots (finis) sur $\mathcal{A}$. L'ensemble $\mathcal{C}$ peut être différent de $\mathcal{M}$.

- L'ensemble des clés $\mathcal{K}$ est un ensemble tel que chaque élément $e \in \mathcal{K}$ correspond à une injection E_e de $\mathcal{M}$ dans $\mathcal{C}$. Nous identifions souvent E_e à une bijection sur son image, et nous pouvons parler de son inverse. Nous considérons habituellement un ensemble $\mathcal{C}$ tel que tous les $e \in \mathcal{K}$ correspondent à une bijection E_e de $\mathcal{M}$ dans $\mathcal{C}$. Dans ce cas, à chaque $d \in \mathcal{K}$, nous désignons D_d une bijection de $\mathcal{C}$ dans $\mathcal{M}$.

Chiffrer un message $m \in \mathcal{M}$ consiste à calculer $E_e(m) = c$ et déchiffrer un texte chiffré $c \in \mathcal{C}$ correspondant consiste à trouver D_d et calculer $D_d(c) = m$. Pour avoir un schéma complet de chiffrement et déchiffrement, il faut disposer d'un couple de clés (e, d). Lorsque $e = d$ ou plus généralement, lorsque d est « facilement » calculable à partir de e, on dit que c'est un chiffrement symétrique (on dit aussi chiffrement à clé secrète). Dans le cas contraire, on dit que c'est un chiffrement asymétrique (ou chiffrement à clé publique). Un protocole (cryptographique) est un énoncé qui spécifie les étapes à exécuter entre deux entités afin d'achever un objectif cryptographique.

1.2. Les fonctions à sens unique. Nous considérons une bijection quelconque f de X sur Y. Si déterminer f^{-1} est « presque » impossible, ou difficile dans un temps « raisonnable » pour presque tout $y \in Y$, on dit que f est une fonction à sens unique. Si X est ordonné, nous pouvons toujours tester certains éléments particuliers (par exemple, au début) de la liste. Ainsi, la difficulté de calculer f^{-1} repose sur l'absence d'un algorithme efficace permettant de trouver l'antécédent d'un élément choisi au hasard. En pratique, nous considérons certaines f telles que en donnant une information supplémentaire, calculer f^{-1} devient possible. On parle dans ce cas de *fonctions à sens unique à trappe*. L'existence d'une telle fonction n'a pas été prouvée. Leur existence en pratique repose sur la difficulté de résoudre certains problèmes mathématiques en temps raisonnable. Voici une liste de problèmes de ce type réputés difficiles :

- **Fct***(Factorisation des entiers)* : Étant donné un entier positif composé n, trouver sa factorisation en nombres premiers.
- **RC***(Racine carré)* : Étant donné un entier positif composé n et $a < n$ tel que a est un carré dans $\mathbb{Z}/n\mathbb{Z}$. Trouver une racine carrée de a.
- **RQ***(Résidu quadratique)* : Étant donné un entier impair positif composé n et $a < n$ tel que le symbole de Jacobi $\left(\frac{a}{n}\right) = 1$. Déterminer si a est un résidu quadratique dans $\mathbb{Z}/n\mathbb{Z}$.

- **LD** *(Logarithme discret)* : Étant donné un groupe cyclique fini G et $a, b \in G$ tels que a soit un générateur de G. Trouver $n \in \mathbb{N}$ tel que $a^n = b$.
- **SSE** *(Somme de sous-ensemble)* : Étant donné un ensemble des entiers positifs $\mathcal{S} = \{s_1, \ldots, s_n\}$ et $s \in \mathbb{N}$. Déterminer s'il existe $\mathcal{B} \subset \mathcal{S}$ tel que $s = \sum_{s_i \in \mathcal{B}} s_i$.

La complexité calculatoire de ces problèmes n'est pas connue, mais il est généralement admis que lorsque les paramètres sont bien choisi, ces problèmes sont insolubles algorithmiquement en un temps raisonnable. Au lieu de prouver leur difficulté absolue, on cherche plutôt à établir leur difficulté relativement à la puissance de calcul dont peut disposer un adversaire. Lorsqu'on envisage un nouveau problème mathématique sous cet aspect calculatoire, on essaie de comparer sa complexité à l'un des problèmes précédents.

1.3. Algorithme, Complexité. Nous considérons le terme « *algorithme* » comme un mécanisme réalisable par une machine de Turing, sans plus de précision. Un algorithme (déterministe) est dit (calculable) en *temps polynomial* s'il existe $k \in \mathbb{N}$ tel que le temps d'exécution dans le pire des cas est $O(n^k)$ où n est la taille des données. De la même manière, nous définissons l'*algorithme non déterministe polynomial* comme un algorithme exécutable en temps polynomial, quitte à prendre certains paramètres aléatoires en cours d'exécution. Lorsque le temps d'exécution est de la forme $\mathrm{e}^{o(n)}$, l'algorithme est dit *sous-exponentiel*. Souvent, les données entrantes sont des éléments d'un corps fini $\mathbb{F}_q$ et le temps d'exécution est

$$L_q[\alpha, c] = O\big(\exp\big((c + o(1))(\log q)^{\alpha}(\log\log q)^{1-\alpha}\big)\big)$$

où $c > 0$ et $0 < \alpha < 1$ sont des constantes.

Sur les problèmes rencontrés, nous pouvons distinguer différents niveaux d'exigence. Nous pouvons demander soit de prouver une affirmation, soit de décider dans un cas particulier, soit de trouver un exemple concret. Ces problèmes sont classés par l'ordre croissant de leur difficulté. Par exemple, nous pouvons démontrer que presque tous les réels sont transcendants sans pouvoir décider la transcendance d'un réel particulier (comme la constante d'Euler γ). Nous pouvons également prouver qu'un nombre n'est pas premier sans trouver un de ses facteurs premiers effectivement (par le test de Fermat).

Un problème est dans la classe **P** s'il peut être résolu par un algorithme déterministe polynomial. Un problème est dans la classe **NP** s'il peut être résolu par un algorithme non déterministe polynomial. De manière équivalente, un problème est dans la classe **NP** s'il

existe un algorithme déterministe polynomial qui, si on lui fournit un *certificat* (une information supplémentaire ne faisant pas partie du problème) permet de vérifier la correction d'une solution.

Prenons l'exemple de la question de la primalité (**Prm**). Il s'agit à répondre à la question suivante : un entier naturel donné n est-il premier ?

La non primalité est évidemment dans **NP**. En effet, si n n'est pas premier, on peut le vérifier en temps polynomial si on connaît un diviseur non trivial de n. Ici le certificat est le diviseur de n.

De manière moins évidente, on sait que $\mathbf{Prm} \in \mathbf{NP}$. D'après un résultat dû à Lucas, n est premier si et seulement si il existe $\alpha \in (\mathbb{Z}/n\mathbb{Z})^*$ tel que $\alpha^{n-1} \equiv 1 \bmod n$, et pour tout diviseur premier p de $n-1$, on a $\alpha^{\frac{n-1}{p}} \not\equiv 1 \bmod n$. On aboutit ainsi à une itération dont le nombre d'étapes et l'explosion combinatoire vont rester acceptables. Nous n'entrons pas dans les détails et ne précisons pas la forme du certificat.

En 2002, Agrawal, Kayal et Saxena ont prouvé que $\mathbf{Prm} \in \mathbf{P}$: il existe un algorithme déterministe polynomial qui permet de décider si un entier naturel n est premier.

On dit qu'un problème P_1 se *réduit* à un problème P_2 polynomialement et on note $P_1 \propto_P P_2$ si l'on possède un algorithme qui résout P_1 lorsqu'on peut trouver un algorithme résolvant P_2. Cela signifie que P_1 n'est pas essentiellement plus difficile que P_2. Lorsque l'on a aussi $P_2 \propto_P P_1$, on note $P_1 \equiv_P P_2$. Les algorithmes considérés sont généralement non-déterministes et ce sont eux qui sont le plus souvent efficaces en pratique. Un *oracle* est un « algorithme hypothétique », notion souvent employée dans la réduction de problèmes. Un problème P est dit NP-complet (appartenant à la classe **NPC**) si $\forall Q \in \mathbf{NP}, Q \propto_P P$.

Un résultat fondamental dans l'étude de la complexité est le théorème de Cook en 1971 qui décrit un problème **NPC**. Le problème de somme de sous-ensemble (**SSE**) est aussi NP-complet. Il est clair que $\mathbf{P} \subset \mathbf{NP}$. Une question centrale dans la théorie de complexité est de savoir si $\mathbf{P} = \mathbf{NP}$. L'abondance de questions appartenant à la classe **NP** nous font conjecturer que $\mathbf{P} \neq \mathbf{NP}$.

Bien qu'il reste décevant de ne pas pouvoir démontrer $\mathbf{P} \neq \mathbf{NP}$, une influence de cette théorie est qu'on ne cherche plus vraiment à établir l'existence absolue d'une fonction à sens unique. En fait, s'il existe une telle fonction, nous démontrons aussitôt que $\mathbf{P} \neq \mathbf{NP}$.

1.4. Fonction de hachage, Générateur de bits pseudo-aléatoires. Les fonctions de hachage, appelées parfois de condensation servent à créer un représentant compact

de taille fixe depuis un message de taille arbitraire. Ainsi, une fonction de hachage est $h : \mathcal{M} \to \mathcal{R}$ où $\mathcal{R}$ est un ensemble des mots sur $\mathcal{A}$ de longueur exactement r, par exemple. Certaines de ces fonctions sont utilisables en cryptographie, lorsqu'elles sont à sens unique, c'est-à-dire telles que h^{-1} est impossible à calculer en pratique. On parle alors de *résistance à la préimage*. Ce type de fonction est couramment utilisé par exemple pour stocker les mots de passe sur les systèmes informatique. Des propriétés supplémentaires sont nécessaires. Supposons que $\sharp\mathcal{M} = 2^t$ et $\sharp\mathcal{R} = 2^r \quad (t > r)$. Si h est bien conçue, deux messages x et x' choisis de manière aléatoire sur $\mathcal{M}$ ont une probabilité 2^{-r} (indépendante de t) d'entrer en collision, c'est-à-dire de vérifier $h(x) = h(x')$. Observons que beaucoup de collisions existent car chaque valeur de h possède environ 2^{t-r} antécédents. Pour une application cryptographique, il est nécessaire que la fonction h *résiste à la collision* : il est impossible en pratique de trouver $x \neq x'$ tels que $h(x) = h(x')$. Cette condition est indispensable pour se prémunir contre l'attaque dite des « anniversaires ». Ces fonctions sont aussi utilisées pour la vérification de l'intégrité des données (par l'envoi en même temps d'un représentant compact du message, la fonction de hachage utilisée étant publique).

Voici la réalisation d'une fonction de hachage due à *Damgård*. Supposons que $\mathcal{A} = \{0, 1\}$. Soient $f_0, f_1 : \mathcal{R} \to \mathcal{R}$ deux fonctions à sens unique qui ne se rencontrent pas, c'est-à-dire qu'il est algorithmiquement difficile de trouver $(a, b) \in \mathcal{R}^2$ tel que $f_0(a) = f_1(b)$. Fixons un $r \in \mathcal{R}$, nous pouvons définir la fonction F de $\mathcal{M}$ dans $\mathcal{R}$ qui à $x = x_1 \cdots x_k$ associe

$$F(x_1 \cdots x_k) = f_{x_1}(f_{x_2} \cdots (f_{x_k}(r))).$$

On peut montrer par récurrence sur k que cette fonction F est une fonction de hachage qui résiste à la collision.

En cryptographie, il est souvent nécessaire de choisir certains paramètres au hasard, principalement lors de la fabrication de clés. Dans ce but, il est utile de pouvoir obtenir une suite très longue de bits qui semble « aléatoire » à partir de quelques valeurs initiales appélées « graine ». C'est la notion de *générateur de bits pseudo-aléatoires* qui a été étudié dans le chapitre 1.

2. Chiffrement symétrique (ou à clé secrète)

Nous pouvons décomposer ce type de chiffrement en deux classes. Le chiffrement par bloc est un chiffrement dans lequel nous morcelons d'abord le message en bloc de même taille et l'on chiffre un bloc par unité de temps. Le chiffrement par chaîne est un cas

particulier du chiffrement précédent dans lequel le bloc est de taille minimum un (un élément de l'alphabet).

2.1. Chiffrement par bloc.

2.1.1. *Chiffrement par substitution.* Ce sont des substitutions par d'autres symboles. Si les mots remplacés utilisent les mêmes alphabets, ce sont des permutations sur $\mathcal{A}$. Pour fixer l'idée, choisissons $\mathcal{A}$ comme l'ensemble des caractères latins et nous l'identifions à $\mathbb{Z}/26\mathbb{Z}$ par l'ordre usuel.

- *Substitution mono-alphabétique :* Appelé aussi substitution simple. Ici l'ensemble des clés $\mathcal{K}$ est le groupe symétrique $\mathfrak{S}_{26}$. Pour $e \in \mathcal{K}$ et $m = m_1 \cdots m_t \in \mathcal{M}$, on pose

$$E_e(m) = e(m_1) \cdots e(m_t).$$

  Pour déchiffrer, nous appliquons $d = e^{-1}$. Lorsque $e(m) = m + 3 \bmod 26$, c'est un décalage de trois positions (à droite), il s'agit du chiffrement de *César*. Le chiffrement par décalage est très facile à déchiffrer puisqu'il n'y a que 26 possibilités. Dans le cas d'une substitution quelconque de $\mathcal{K} = \mathfrak{S}_{26}$, le déchiffrement n'est pas beaucoup plus difficile. Par exemple, en français la lettre "e" apparaît plus fréquemment que les autres lettres, ce qui permet d'identifier la lettre remplaçante. De proche en proche, nous trouvons les autres lettres. Pour mettre en œuvre cette méthode il faut disposer d'une analyse de l'entropie de la langue française.
- *Substitution poly-alphabétique :*

  Nous procédons ici par blocs de même taille t.

  Considérons par exemple le *chiffrement de Vigenère.* L'espace des clés est $\mathcal{K} = \mathfrak{S}_{26}^t$ et pour $e = (e_1, \ldots, e_t) \in \mathcal{K}$, et $m = m_1 \cdots m_t \in \mathcal{M}$, on pose

$$E_e(m) = e_1(m_1) \cdots e_t(m_t).$$

  Par exemple, prenons $t = 3$, $e = (e_1, e_2, e_3)$ avec

$$e_1 : x \mapsto x + 2 \bmod 26, \quad e_2 : x \mapsto x + 9 \bmod 26, \quad e_3 : x \mapsto x + 13 \bmod 26.$$

  Le texte

  CEC HIF FRE MEN TNE STP AST RES SUR

  devient

  ENP JRS HAR ONA VWR UCC CBG TNF UDE.

  Ce chiffrement ne conserve pas la fréquence d'apparition des lettres. Cependant, il n'est pas beaucoup plus difficile à déchiffrer. En effet, il suffit de déterminer la taille des blocs t puis procéder à la cryptanalyse comme dans le chiffrement de substitution

mono-alphabétique. Pour cela, nous remarquons que deux segments identiques du message décalés de d positions sont chiffrés de la même manière si t divise d. Nous cherchons alors les couples de segments identiques (de longueur $\geqslant 3$) dans le texte chiffré et nous notons les distances entre ces couples d_i. Il y a une bonne chance que t divise le pgcd des d_i. Ce procédé s'appelle *le test de Kasiski.*

On peut mentionner aussi la *substitution homophonique* qui consiste à remplacer une lettre par des symboles différents. Le but est de minimiser l'attaque sur la fréquence d'apparition des lettres, quitte à accroître la taille de texte chiffré.

2.1.2. *Chiffrement par transposition.* On effectue des permutations sur chaque bloc. L'espace des clés est $\mathcal{K} = \mathfrak{S}_t$ et pour $e \in \mathcal{K}$, $m = m_1 \cdots m_t \in \mathcal{M}$, on pose

$$E_e(m) = m_{e(1)} \cdots m_{e(t)}.$$

La cryptanalyse paraît facile car ce sont des permutations de t éléments. Toutefois, ce chiffrement acquiert une importance significative par composition avec d'autres chiffrements par blocs.

2.1.3. *La composition des chiffrements.* Lorsque $\mathcal{M} = \mathcal{C}$, les chiffrements sont des permutations sur $\mathcal{M}$ et nous pouvons envisager la composition des chiffrements. Nous disons aussi appliquer différents tours de chiffrements. Si la cryptanalyse de chacun de ces chiffrements considéré séparément est facile à réaliser, il n'en est plus de même après plusieurs tours de chiffrements. Une substitution a pour but de rendre plus complexe le texte chiffré. Nous disons qu'elle ajoute de la confusion. Tandis qu'une transposition a pour but de distribuer l'ordre des bits dans le message et permet d'harmoniser la fréquence d'apparition des lettres. Nous disons qu'elle ajoute de la diffusion.

Considérons par exemple le *chiffrement de Hill.* C'est une forme améliorée du chiffrement affine. L'espace des clés $\mathcal{K}$ est le groupe des matrices carrées inversibles de taille t sur $\mathbb{Z}/26\mathbb{Z}$ et pour $e \in \mathcal{K}$, $m = (m_1, \ldots, m_t) \in \mathcal{M}$, on pose

$$E_e(m) = m \cdot e \qquad \text{(produit matriciel)}.$$

La clé de déchiffrement est $d = e^{-1}$, l'inverse de la matrice du chiffrement.

Le chiffrement de Hill est plus élaboré que les chiffrements précédents car il combine substitution et transposition. En effet, le chiffrement par transposition est un cas particulier du chiffrement de Hill en choisissant la matrice de permutation correspondante.

Le système du chiffrement « Lucifer » développé par IBM en 1973 a suivi une approche similaire qui consiste à appliquer plusieurs fois des produits alternés de substitutions et

transpositions. L'innovation essentielle de **DES** en 1976 qui remplace « Lucifer » est l'introduction du chiffrement de *Feister*. Dans ce chiffrement, on effectue 16 tours. Une partie non linéaire est réalisée par une fonction F. Il présente l'avantage de ne pas nécessiter le calcul de F^{-1} pour le déchiffrement. Aujourd'hui, l'espace des clés de **DES** est considéré d'une taille insuffisante et **DES** a été remplacé par **AES** en 2001.

2.2. Chiffrement par chaîne. Contrairement au chiffrement par bloc qui fixe une taille de bloc et chiffre par une clé choisie, le chiffrement par chaîne utilise le bloc minimal mais fait varier la clé au cours du chiffrement. Le schéma est pour $e = e_1, \ldots, e_i \in \mathcal{K}$, $m = m_1 \cdots m_i \in \mathcal{M}$,

$$E_e(m) = E_{e_1}(m_1) \cdots E_{e_i}(m_i).$$

Par convention, nous appelons la clé du chiffrement cette suite de clés. La clé e peut être soit choisie au hasard, soit générée par un algorithme. Il se peut que le *texte chiffré* soit dépendant à la fois de la clé et du *texte clair* (les paramètres de la génération de clé dépendent aussi du message). Dans ce cas, le chiffrement est dit *asynchrone*.

Si la clé est périodique (i.e. $\exists k \in \mathbb{N}, \forall i \in \mathbb{N}^*, e_{i+k} = e_i$), nous retrouvons le schéma du chiffrement par bloc.

Le chiffrement par chaîne présente l'avantage de ne pas propager l'erreur lors de transmission. Il devient indispensable lorsque le canal n'est pas fiable (les erreurs de transmission sont considérables) ou si le système du chiffrement n'a pas beaucoup de mémoire (la mémoire tampon pour le traitement de données est inexistante ou très limitée).

Voici un exemple fondamental pour comprendre l'importance de cette idée. On y voit également l'intérêt de l'étude sur les suites pseudo-aléatoires.

– *le chiffrement de Vernam :* On prend $\mathcal{A} = \{0; 1\}$. Nous chiffrons à l'aide de l'addition modulo 2 notée $\oplus$ (XOR, opérateur logique « ou » exclusif) appliqué à chaque bit de message et de la clé :

$$c_i = E_{e_i}(m_i) = m_i \oplus e_i.$$

Si la clé est une suite véritablement aléatoire et n'est pas réutilisée, ce chiffrement est appelé « *one-time pad* ». Shannon a montré théoriquement que ce chiffrement est incassable en calculant l'entropie du texte chiffré. Cependant, pour réaliser cette sécurité inconditionnelle, la clé doit être d'une même longueur que le message. Le transport de clé par un canal sûr demande un coût non négligeable. Il est ouïe dire

que Moscou et Washington communiquaient par « one-time pad » à l'époque de la guerre froide.

Il est important de ne pas réutiliser la clé. Si l'on connaît

$$c_i = m_i \oplus e_i, \qquad c'_i = m'_i \oplus e_i,$$

alors $c_i \oplus c'_i = m_i \oplus m'_i$ et une cryptanalyse peut se faire sur la fréquence de redondance de ce dernier.

3. Chiffrement asymétrique (ou à clé publique)

La cryptographie a connu une révolution depuis l'apparition de l'article de Diffie et Hellman en 1976. Rappelons que dans le chiffrement symétrique, la connaissance de la clé du chiffrement e signifie pour un adversaire la connaissance de la clé du déchiffrement d. L'idée importante dans **[14]** est de séparer ces deux clés telles que d est difficile à déterminer sachant e (grâce aux fonctions à sens unique). Cela rend possible la communication sécurisée entre deux entités non préalablement mises en contact, à la différence du chiffrement symétrique qui nécessite un canal sécurisé pour transporter la clé secrète.

Le schéma est le suivant : Supposons que Alice et Bob sont deux individus qui n'avaient pas de secret partagé auparavant, mais disposent de clés (e_A, d_A) et (e_B, d_B) respectivement suivant une méthode de chiffrement asymétrique publique. e_A et e_B sont diffusées publiquement en tant que clé d'identité, tandis que d_A et d_B sont des secrets absolus. Si Alice veut envoyer à Bob un message chiffré m, elle peut tout simplement envoyer $e_B(m)$ dans un canal quelconque puisque seul Bob possède d_B pour déchiffrer dès que le message est ainsi chiffré.

Cela soulève un autre problème. Comment Bob arrive-t-il à identifier l'expéditeur du message ? À la différence du chiffrement symétrique, ici tout le monde peut envoyer le même message à Bob et il déchiffre toujours avec la même clé d_B. Là encore le schéma typique du chiffrement asymétrique offre une élégante solution. Il suffit que Alice chiffre son message avec sa clé de déchiffrement d_A, puisqu'elle est la seule à connaître d_A et tout le monde peut vérifier avec la clé publique e_A. Nous disons qu'elle a fait une *signature digitale* au message. Ainsi, Alice envoie $d_A(e_B(m))$ et à la réception, Bob applique d_B puis e_A pour retrouver m.

3.1. Chiffrement de RSA. Rivest, Shamir et Adleman ont inventé en 1978 un chiffrement asymétrique qui peut être implanté de manière pratique. C'est aussi un cryptosystème très utilisé.

- *Génération des clés :*

  On choisit de manière aléatoire deux grands nombres premiers distincts p et q, de taille comparable l'un à l'autre. Il est facile de calculer $n = pq$ et $\phi(n) = (p-1)(q-1)$. On choisit ensuite $e \in [2, \phi(n)]$ tel que e soit premier avec $\phi(n)$ et on calcule l'inverse d de e dans $\mathbb{Z}/\phi(n)\mathbb{Z}$ (par l'algorithme d'Euclide).

  La clé publique est (n, e). La clé privée est d.

- *Schéma :*
  - But : Alice veut envoyer un message $m \in \mathbb{Z}/\phi(n)\mathbb{Z}$ à Bob qui possède la clé publique (n, e) et la clé privé d.
  - Chiffrement : Alice obtient (n,e) en publique et envoie $c \equiv m^e \bmod n$.
  - Déchiffrement : Bob reçoit c et retrouve $m \equiv c^d \bmod n$ à l'aide de d.
  - Justification : $c^d \equiv m^{ed} \equiv m^{1+k\phi(n)} \equiv m \bmod n$. (On a utilisé Euler sur p et q.)

La sécurité de ce chiffrement est basée sur la difficulté du problème suivant (problème de **RSA**) : Étant donné n, e, c comme précédemment, trouver m (sans connaître p, q ou d).

Il est clair que **RSA** $\propto_P$ **Fct**. La connaissance de d joue un rôle important pour le déchiffrement. Si nous savons factoriser n, nous trouverons d et aussi le message. Si un adversaire connaît d, il peut aussi factoriser n par un algorithme non-déterministe (avec la probabilité de succès $\geqslant 1 - 2^{-n}$ après n essais). En effet, écrire $ed - 1 = 2^s t$ avec t impair, nous pouvons démontrer qu'il existe $i \in [1, s]$ tel que pour (plus de) la moitié de valeur $a \in \mathbb{Z}/n\mathbb{Z}^*$, nous avons $a^{2^{i-1}t} \not\equiv \pm 1 \bmod n$ et $a^{2^i t} \equiv 1 \bmod n$.(cf. [**59**] p.195) Dans ce cas, $(a^{2^{i-1}t} - 1, n)$ est un facteur non trivial de n. Nous avons ainsi établi :

Dans le cas général, trouver d à partir du couple (n, e) est aussi difficile à factoriser n.

Cependant, il existe d'autres attaques possibles liés au choix de e ou d ou à d'autres informations supplémentaires.

3.2. Chiffrement d'El Gamal. C'est un chiffrement basé sur le problème de **LD**.

- *Génération des clés :*

  On choisit un nombre premier p tel que le problème du logarithme discret soit difficile dans $(\mathbb{Z}/p\mathbb{Z})^*$. (En pratique, on choisit p tel qu'il possède au moins 300 chiffres et que $p - 1$ possède un grand facteur premier.) On choisit $\alpha \in (\mathbb{Z}/p\mathbb{Z})^*$ un élément

primitif et $a \in \mathbb{N}$ et on calcule $\beta \equiv \alpha^a \bmod p$. La clé publique est (p, α, β). La clé privée est a.

- *Schéma :*
  - But : Alice veut envoyer un message $m \in \mathbb{Z}/p\mathbb{Z}$ à Bob qui possède la clé publique (p, α, β) et la clé privée a.
  - Chiffrement : Alice choisit au hasard $k \in \mathbb{Z}/(p-1)\mathbb{Z}$ (k secret) et envoie (y_1, y_2) à Bob où $y_1 \equiv \alpha^k \bmod p$ et $y_2 \equiv \beta^k \bmod p$.
  - Déchiffrement : Bob retrouve $m \equiv y_2(y_1^a)^{-1} \bmod p$ à l'aide de a.

On remarque que le chiffrement possèdent un aspect probabiliste (choix de k) qui permet de donner différents messages chiffrés à un même message initial, quitte à doubler la taille du texte chiffré.

3.3. Protocole de Diffie Hellman pour établissement de clé. L'avantage du chiffrement asymétrique est qu'il ne requiert pas de canal sûr pour transmettre la clé. Ce protocole simple décrit comment créer une clé commune sur canal public. Il a été décrit dans l'article [**14**] en 1976.

- *But :*

  Alice et Bob se communique sur un canal non sécurisé et souhaitent créer une clé K qui est secret partagé seulement pour eux.

- *Schéma :*

  Alice et Bob se mettent d'accord pour les choix de p un nombre premier convenable et α un générateur de $(\mathbb{Z}/p\mathbb{Z})^*$. Alice et Bob choisissent ensuite au hasard x et y (secrets) respectivement (avec $1 \leqslant x, y \leqslant p-2$). Alice envoie $\alpha^x \bmod p$ à Bob et Bob envoie $\alpha^y \bmod p$ à Alice. Alice calcule $K = (\alpha^y)^x \bmod p$ et Bob calcule $K = (\alpha^x)^y \bmod p$. K devient le clé secret partagée.

La sécurité de ce protocole est basée sur la difficulté du problème de Diffie-Hellman (**DH**) : Étant donne p et α comme précédemment. Sachant $\alpha^x \bmod p$ et $\alpha^y \bmod p$. Calculer $\alpha^{xy} \bmod p$.

Nous avons évidemment $\mathbf{DH} \propto_P \mathbf{LD}$.

Nous remarquons que ce protocole ne fait pas l'authentification dans l'échange. D'ailleurs, il ne permet pas le transport de la clé.

4. Comparaison des chiffrements symétrique et asymétrique

Le chiffrement symétrique a une longue histoire. Le chiffrement symétrique est beaucoup plus rapide que n'importe quel chiffrement asymétrique et la clé utilisée est relativement petite en taille. Cependant la clé doit toujours rester secrète pour les deux entités. Si le réseau est grand, la gestion de clés devient difficile. Le chiffrement asymétrique possède des avantages et des désavantages en rôle inverse par rapport au chiffrement symétrique. L'absence de nécessité de canal sûr permet générer des clés éphémères. Toutefois, la sécurité est basée sur des hypothèses mathématiques non prouvées, donc ne peut pas offrir une *sécurité inconditionnelle* comme le fait « one-time pad ». En fait, tout système de chiffrement qui ne résiste pas à la recherche exhaustive sur ses clés ne peut pas avoir la sécurité inconditionnelle. En pratique, les deux techniques sont utilisées pour bénéficier de leurs avantages complémentaires. Le chiffrement asymétrique sert dans la gestion de clés et la non-répudiation des deux entités. Ensuite le chiffrement symétrique utilise la clé ainsi générée pour accélérer le chiffrement et assure d'autres services comme l'intégrité des données.

5. Quelques remarques sur les problèmes calculatoires

Nous donnons un aperçu sur les problèmes évoqués dans la section 1.2. En général, nous ne connaissons pas l'éventuelle équivalence de leur difficulté. Par exemple, le lien entre **Fct** et **LD** n'est pas connu. Rabin a été le premier à remarquer que **Fct** $\equiv_P$ **RC**. En effet, nous connaissons des algorithmes (déterministe ou non) (voir par exemple [**11**] p.33 algorithm 1.5.1 et p.44) pour trouver une racine carré de $a \in (\mathbb{Z}/p^s\mathbb{Z})^*$ lorsque $\left(\frac{a}{p}\right) = 1$. Nous étendons pour le cas général $n = \prod p_i^{s_i}$ par l'algorithme Euclidien. Donc **RC** $\propto_P$ **Fct**.

Pour voir l'inverse, supposons qu'il existe un oracle qui résout **RC** (en donnant au hasard une des racines carrés, disons). Nous choisissons alors au hasard $x \in (\mathbb{Z}/n\mathbb{Z})$ et nous demandons à l'oracle de trouver une racine carré de $x^2 \bmod (\mathbb{Z}/n\mathbb{Z})$. Si $y^2 \equiv x^2 \bmod n$ et $x \not\equiv \pm y \bmod n$, alors $(x - y, n)$ est un facteur non trivial de n. Si cet oracle (non déterministe) donne toujours une réponse équiprobablement répartie, la probabilité de succès pour chaque choix de x est $1/2$. Ainsi, **Fct** $\propto_P$ **RC**.

Nous avons également du mal à exhiber le lien entre deux instances du même problème. Par exemple, un problème **LD** modulo p est-il aussi difficile pour un autre nombre premier

q ? Sur le problème **SSE**, Impaglizzo et Naor ont démontré que le cas le plus difficile est lorsque le cardinal de $\mathcal{S}$ est égal à la taille des s_i (supposés égaux). (cf. [**30**])

Bien que le problème **SSE** soit un problème NP-complet, ce n'est pas nécessairement une bonne idée de l'utiliser pour créer une fonction à sens unique. Merkle et Hellman ont conçu un cryptosystème basé sur ce problème (ou plus généralement nous devons dire le problème de sac-à-dos) presque en même temps que le système **RSA**, mais Shamir a trouvé une faille en 1982 et Brickell a rendu praticable l'attaque. Une des raisons pour expliquer cet échec est que la complétude NP mesure la difficulté dans le pire cas, mais nous souhaitons de proposer un problème difficile en moyenne pour essentiellement tous les instances.

5.1. Problème de factorisation. Nous devons distinguer trois types de problèmes : Donner une estimation probabiliste de la primalité d'un entier n ; Déterminer (et prouver) sa primalité ; Factoriser (lorsque n est composé). Les deux premières demandes sont essentiellement résolues, surtout lorsque n possède une forme spéciale. Tandis que la dernière, durant une longue période, consistait à tester par les divisions successives des nombres premiers. S'il est possible d'aller jusqu'au $\lfloor\sqrt{n}\rfloor$, c'est aussi une preuve de sa primalité. En pratique, on commence toujours par tester jusqu'à certaine borne.

5.1.1. *Algorithme ρ de Pollard.* L'idée est de construire une suite (pseudo) aléatoire dans $\mathbb{Z}/p\mathbb{Z}$ (pour un $p|n$). Choisissons d'abord $x_0 \in \mathbb{Z}/n\mathbb{Z}$ et $f \in (\mathbb{Z}/n\mathbb{Z})[X]$ et posons $x_{k+1} \equiv f(x_k) \bmod n$. Alors la suite $y_k \equiv x_k \bmod p$ satisfait la même récurrence. Si f est bien choisi, (y_k) se comporte comme une suite (périodique à partir de certain rang) aléatoire dans $\mathbb{Z}/p\mathbb{Z}$. Nous pouvons démontrer que la période espérée est $\sqrt{p}$. Si $y_{k+t} \equiv y_k \bmod p$ alors $x_{k+t} \equiv x_k \bmod p$ et on a $(x_{k+t} - x_k, n) > 1$. Si ce dernier est différent de n, nous avons trouvé un facteur non trivial de n après $O(n^{1/4})$ étapes d'itération. Sinon, nous choisissons un autre f. Le plus simple est de considérer $f(x) = x^2 + c$ en commençant par $c = 1$.

En pratique, pour résoudre le problème de stockage pour les comparaisons, nous pouvons utiliser une astuce de Brent (cf. [**11**] p.420) dans l'implantation.

5.1.2. *Algorithme de crible quadratique.* Nous devons à *Dixon* l'idée (bien que l'idée puisse remonter à Fermat) de chercher un couple (x, y) tel que $x \not\equiv y \bmod n$ mais $x^2 \equiv y^2 \bmod n$ afin de trouver un facteur non trivial de n ($(x - y, n)$ en sera un). Pour ne pas échouer dans la recherche, nous sommes ramenés à factoriser $x_k^2 \equiv (-1)^{s_{k,0}} p_1^{s_{k,1}} \cdots p_m^{s_{k,m}} \bmod n$ pour des x_k tels que les carrés soient factorisables par des nombres premiers plus petits

à une certaine borne, puis nous cherchons des possibles combinaisons linéaires en $s_{i,j}$ qui s'annulent dans $\mathbb{Z}/2\mathbb{Z}$ pour trouver y.

Pomerance introduit une nouvelle idée en considérant le polynôme $Q(a) = (\lfloor n \rfloor + a)^2 - n$. Si $a = O(n^\varepsilon)$, on a $Q(a) = O(n^{1/2+\varepsilon})$ et $Q(a) \equiv x^2 \bmod n$ pour $x = \lfloor n \rfloor + a$. L'intérêt de considérer $Q \in \mathbb{Z}[X]$ est qu'on n'a plus besoin de factoriser mais cribler : Si $m \mid Q(a)$ alors $\forall k \in \mathbb{N}, m \mid Q(a + km)$. Nous sommes ramenés à résoudre $x^2 \equiv n \bmod m$ et nous connaissons des algorithmes pour trouver des racines carrés x et donc a. On se donne alors un long intervalle pour laisser varier a et on crible les $Q(a)$ par $m = p^k$.

Nous disposons d'autres algorithmes de factorisation. La méthode de « *crible algébrique* » a été proposé par Pollard puis raffinée par Buhler, Lenstra et Pomerance. C'est un algorithme sous-exponentiel qui permet de traiter $n > 10^{100}$. C'est avec cette méthode que le neuvième *nombre de Fermat* a été entièrement factorisé.

5.2. Problème de logarithme discret. Le sous groupe souvent donné comme exemple est $(\mathbb{Z}/p\mathbb{Z})^*$ où p est un nombre premier tel que $p - 1$ possède un grand facteur premier également. Nous remarquons que la difficulté ne dépend pas de générateur choisi. Plus généralement, nous pouvons même ne pas choisir un générateur mais demander de trouver x tel que $\alpha^x = \beta$ pourvu que β a le même ordre que α. (Cela suggère une autre généralisation à un groupe fini quelconque.) Cependant, limitons-nous dans le cas de groupe cyclique avec un générateur donné α. Nous disposons de plusieurs algorithmes de différente efficacité pour résoudre le problème. L'algorithme de Shanks (*Baby-step-Giant-step*) s'adapte à tous les groupes cycliques (finis). L'algorithme de Pohlig-Hellman est plus avantageux lorsque l'ordre du groupe ne possède que des petits facteurs premiers (et sa factorisation est connue). Quant à la méthode du calcul d'indice, elle ne s'applique qu'à $(\mathbb{Z}/p\mathbb{Z})^*$ mais c'est un algorithme sous-exponentiel en pratique.

5.2.1. *Algorithme de Shanks.* C'est un compromis entre espace et temps d'exécution dans la recherche exhaustive. Notons n l'ordre du groupe et $m = \lfloor \sqrt{n} \rfloor$. Nous remarquons si $\alpha^x = \beta$, nous pouvons écrire $x = im + j$ avec $0 \leqslant i, j \leqslant m$ et $\beta(\alpha^{-m})^i = \alpha^j$. Cela suggère la création d'un tableau de m éléments puis faire des comparaisons. La complexité est alors $O(\sqrt{n})$ (sous l'hypothèse que la loi du groupe est l'opération la plus longue).

A la différence du problème de **Fct**, nous connaissons l'effort minimal nécessaire pour résoudre le problème de **LD**. Shoup [**58**] a établit que les algorithmes génériques qui permettent de traiter n'importe quel groupe cyclique nécessite une complexité $O(\sqrt{n})$. Ainsi, l'algorithme de Shanks est optimal pour l'essentiel.

6. Cryptanalyse

La cryptanalyse est l'étude de la sécurité des cryptosystèmes et les attaques possibles des adversaires. Nous avons vu certaines analyses sur le chiffrement symétrique par bloc. Certes, la sécurité est intimement liée à chaque implantation propre du système. Nous pouvons néanmoins évaluer son niveau de manière générale.

6.1. Évaluation de sécurité.

- **Sécurité inconditionnelle :** c'est la mesure de sécurité la plus stricte. En premier lieu, nous pourrions souhaiter que le système est incassable même si l'adversaire dispose de ressource illimitée. C'est le secret parfait du point de vue de la théorie d'information. L'adversaire n'avance rien dans le déchiffrement après l'observation du texte chiffré.
- **Sécurité calculatoire :** elle mesure l'effort calculatoire nécessaire pour casser le système. Nous employons souvent la théorie de la complexité pour comparer de différents systèmes ou de donner un aperçu intuitif sur la sûreté du système. Les problèmes calculatoires sont ici la base de sécurité.
- **Sécurité prouvée :** nous pouvons utiliser la théorie d'information pour montrer la sûreté du cryptosystème dans l'aspect probabiliste. Nous pouvons également procéder comme dans la théorie de complexité, montrer que le cassage du système est aussi difficile que la solubilité d'un problème réputé difficile. Notons que sur ce dernier point, la sécurité obtenue n'est que relative.
- **Sécurité ad hoc :** ceci résume la sécurité que nous pouvons espérer de manière heuristique. Nous évaluons les méthodes actuelles connues et la ressource disposée de l'adversaire pour chercher un confort de sécurité.

6.2. Les modèles d'attaque. Nous pouvons distinguer différents niveaux d'attaque sur le cryptosystème selon l'information disponible de l'adversaire. La clé est supposée inchangée durant l'attaque.

- **Attaque du texte chiffré connu :** l'adversaire ne dispose que du texte chiffré. Tout chiffrement doit au moins résister à cette attaque.
- **Attaque du texte clair connu :** l'adversaire dispose certains messages ainsi que des textes chiffrés correspondants.
- **Attaque du texte clair choisi :** on suppose que l'adversaire a accès à une machine chiffrante. Il peut choisir son message et récupérer le texte chiffré correspondant. Le chiffrement de Hill succombe à cette attaque puisqu'il s'agit de résoudre un système d'équations linéaires.
- **Attaque du texte chiffré choisi :** on suppose que l'adversaire a accès temporairement à une machine déchiffrante. Il peut choisir un texte chiffré et récupérer le message correspondant (sans pour autant pouvoir soumettre le texte chiffré voulu).

6.3. Objectif de l'adversaire. Nous distinguons les adversaires passifs qui ne cherchent qu'à retrouver le message initial d'une part, et les adversaires qui cherchent également à modifier le contenu d'autre part. Il peut arriver que l'objectif de l'adversaire soit moins ambitieux : il ne cherche qu'à obtenir une information partielle.

- **le cassage total :** C'est la clé qui a été découverte, l'adversaire peut alors faire ce qu'il veut.
- **le cassage partiel :** Sans connaître la clé, l'adversaire réussit à obtenir une certaine information sur le texte chiffré ou à déchiffrer avec une probabilité non négligeable.

Sur ce dernier point, on trouve la notion de *sécurité sémantique* : le chiffrement est dit *sémantiquement sûr* si de tout information que l'adversaire peut obtenir (par un quelconque algorithme en temps polynomial) à partir du texte chiffré, il peut aussi le faire sans le texte chiffré. En fait, cette définition est équivalent à la *sécurité polynomiale*, qui demande que l'adversaire ne peut pas distinguer deux textes chiffrés construits à partir de deux messages différents par un algorithme en temps polynomial. (cf. **[67]**)

Dans le chiffrement RSA , comme la clé publique e est impaire et $c \equiv m^e \bmod n$, on a $\left(\frac{c}{n}\right) = \left(\frac{m}{n}\right)$. Donc ce chiffrement laisse fuir l'information sur la valeur du symbole de Jacobi. Cependant, on a pu démontrer que déterminer la parité de m est aussi difficile de trouver m lui-même. (cf. par exemple **[59]** p.207)

6.4. Exemples d'attaque spécifique sur le chiffrement RSA. Nous savons que si nous obtenons la factorisation de n, nous avons le cassage total. Nous savons qu'obtenir la clé privée d est aussi difficile que factoriser n. Toutefois, lorsque certaines informations

sont révélées, ou certains paramètres sont mal choisis, ou tout simplement l'implantation est mauvaise, le cryptosystème sera cassé sans la factorisation de n. Remarquons d'abord que le chiffrement RSA ne résiste pas à l'attaque de texte chiffré choisi : si un adversaire veut déchiffrer $c \equiv m^e \bmod n$ et s'il peut choisir $x \in (\mathbb{Z}/n\mathbb{Z})^*$ au hasard et demander à la machine déchiffrante de déchiffrer $\tilde{c} \equiv cx^e \bmod n$ pour obtenir $\tilde{m} \equiv \tilde{c}^d \bmod n$, alors puisque $\tilde{c}^d \equiv c^d x^{ed} \equiv mx \bmod n$, il obtient $m \equiv \tilde{m}x^{-1} \bmod n$.

•La connaissance de $\phi(n)$

Si $\phi(n)$ est connu, les deux égalités

$$n = pq, \quad \phi(n) = (p-1)(q-1),$$

entraînent

$$p^2 - (n - \phi(n) + 1)p + n = 0.$$

Donc, il suffit de résoudre une équation de second degré pour factoriser n.

•Lorsque p et q sont « proches »

Si $q - p = 2d$ alors $n + d^2 = (p+d)^2$. Puisque d est supposé petit, ce qui permet une recherche par test sur le carré parfait.

•Attaque de module commun

Supposons que deux entités utilise les clés (n, e_1), (n, e_2) respectives comme clé publiques avec $(e_1, e_2) = 1$. Si un adversaire sait que les textes chiffrés interceptés sont provenus du même message :

$$c_1 \equiv m^{e_1} \bmod n, \quad c_2 \equiv m^{e_2} \bmod n,$$

alors il peut retrouver m. En effet, il calcule d'abord $k_1 \equiv e_1^{-1} \bmod e_2$. Comme $e_1 k_1 - 1$ est divisible par e_2, notons $k_2 = (e_1 k_1 - 1)/e_2$, alors $e_2 k_2 - e_1 k_1 = 1$. Finalement,

$$m \equiv c_1^{k_1} c_2^{-k_2} \bmod n.$$

Ceci constitue un exemple de l'échec de protocole.

•Attaque de Wienner

Pour accélérer le déchiffrement, nous aimerions choisir d petit, mais cela peut conduire à une catastrophe : Supposons que

$$3d < n^{1/4}, \quad p < q < 2p.$$

Comme $ed \equiv 1 \bmod \phi(n)$, il existe $k < d$ tel que $ed - k\phi(n) = 1$. On a aussi $n - \phi(n) = p + q - 1 < 3p < 3\sqrt{n}$. Donc

$$\left|\frac{e}{n} - \frac{k}{d}\right| = \left|\frac{ed - kn}{dn}\right| = \left|\frac{1 + k(\phi(n) - n)}{dn}\right| < \frac{3k}{d\sqrt{n}}.$$

Comme $3k < 3d < n^{1/4}$, on en déduit

$$\left|\frac{e}{n} - \frac{k}{d}\right| < \frac{1}{3d^2}.$$

Donc k/d est une réduite du développement de e/n en fraction continue. Nous pouvons facilement mener une recherche exhaustive pour retrouver d.

6.5. Cryptanalyse linéaire et différentielle. Le travail général d'un cryptanalyste est étudier un système de chiffrement itéré que fait partie le chiffrement par bloc. Les cryptanalyses linéaire et différentielle sont deux méthodes génériques adaptés à cette situation. La *cryptanalyse linéaire* est une attaque à texte clair connu qui cherche les « relations linéaires » entre le message et le texte chiffré. Cette technique a été mise au point par Matsui et elle s'avère la plus efficace contre le **DES** . La *cryptanalyse différentielle* est une attaque à texte clair choisi qui s'intéresse aux bits entrants qui possèdent une valeur spécifique de ou exclusif. Cette technique a été « découverte » par Biham et Shamir et ils ont remarqué que le **DES** appliqué avec moins de 16 tours succombe à cette attaque. Ceci laisse penser que les chercheurs de IBM auraient connu cette technique environ quinze ans auparavant, sans la publier.

Bibliographie

[1] R. Ahlswede, L. Khachatrian, C. Mauduit et András Sárközy, On the complexity of families of binary sequences, *Period. Math. Hungar.* **46**,2 (2003), 107–118.

[2] N. Alon, Y. Kohayakawa, C. Mauduit, C.-G. Moreira et V. Rödl, Measures of pseudorandomness for finite sequences : minimal values, preprint.

[3] N. Alon, Y. Kohayakawa, C. Mauduit, C.-G. Moreira et V. Rödl, Measures of pseudorandomness for finite sequences : typical values, *Combinatorics, Probability and Computation*, à paraître.

[4] L. Blum, M. Blum et M. Shub, A simple unpredictable pseudorandom number generator, *SIAM J. Comput.* **15**,2 (1986), 364–383.

[5] J. Boyar, Inferring sequences produced by pseudo-random number generators, *J. Assoc. Comput. Mach.* **36**,1 (1989), 129–141.

[6] N.G.De Bruijn, On the number of positive integers $\leqslant x$, and free prime factors $> y$.ii, *Nederl. Akad. Wetensch. Proc. Ser. A 69=Indag. Math. 28*, 1966, 239–247.

[7] Julien Cassaigne, Sébastien Ferenczi, Christian Mauduit, Joël Rivat et András Sárközy, On finite pseudorandom binary sequences. III. The Liouville function. I., *Acta Arith.* **87**,4 (1999), 367–390.

[8] Julien Cassaigne, Sébastien Ferenczi, Christian Mauduit, Joël Rivat et András Sárközy, On finite pseudorandom binary sequences. IV. The Liouville function. II., *Acta Arith.* **95**,4 (2000), 343–359.

[9] Julien Cassaigne, Christian Mauduit et András Sárközy, On finite pseudorandom binary sequences. VII. The measures of pseudorandomness., *Acta Arith.* **103**,2 (2002), 97–118.

[10] G. J. Chaitin, On the length of programs for computing finite binary sequences, *J. Assoc. Comput. Mach.* **13** (1966), 547–569.

[11] H. Cohen, *A course in computational algebraic number theory, Graduate texts in mathematics* volume 138, Springer-Verlag, Berlin Heidelberg, 1993.

[12] Hedi Daboussi et András Sárközy, On pseudorandom properties of multiplicative functions., *Acta Math. Hungar.* **98**,4 (2003), 273–300.

[13] H. Davenport, *Multiplicative Number Theory (revised by H.L. Montgomery)*, Springer Verlag New York, édition seconde, 1980.

[14] Whitfield Diffie et Martin E. Hellman, New Directions in Cryptography, *IEEE Transactions on Information Theory* **22**,6 (1976), 644–654.

[15] P. D. T. A. Elliott, On the correlation of multiplicative and the sum of additive arithmetic functions., *Mem. Amer. Math. Soc.* **112**,538 (1994).

[16] P. Erdős, C. Pomerance et A. Sárközy, On locally repeated values of certain arithmetic functions. ii., *Acta Math. Hungar.* **49**,1-2 (1987), 251–259.

[17] H. Faure, Discrépances de suites associées à un système de numération (en dimension un), *Bull. Soc. Math. France* **109**,2 (1981), 143–182.

[18] William Feller, *Introduction to probability theory I*, John Wiley & sons, édition troisième, 1968.

[19] Etienne Fouvry et Gérald Tenenbaum, Entiers sans grand facteur premier en progressions arithmétiques, *Proc. London Math. Soc.* **63**,3 (1991), 449–494.

[20] J. N. Franklin, Deterministic simulation of random processes, *Math. Comp.* **17** (1963), 28–59.

[21] Dominique Fuata et Aimé Fuchs, *Calcul des probabilités*, Dunod, édition deuxième, 2003.

[22] I. J. Good, The serial test for sampling numbers and other tests for randomness, *Proc. Cambridge Philos. Soc.* **49** (1953), 276–284.

[23] Louis Goubin, Christian Mauduit et András Sárkőzy, Construction of large families of pseudorandom binary sequences., *J. Number Theory* **106**,1 (2004), 56–69.

[24] Katalin Gyarmati, On a family of pseudorandom binary sequences, *Periodica Math. Hungar.* **49**,2 (2004), 1–19.

[25] Katalin Gyarmati, An inequality between the measures of pseudorandomness, *Annales Univ. Sci. Budapest. Eötvös*, à paraître.

[26] Katalin Gyarmati, On a fast version of a pseudorandom generator, *General Theory of Information Transfer and Combinatorics, Conference Proceedings*, à paraître.

[27] Katalin Gyarmati, On the correlation of binary sequences, *Studia Sci. Math. Hungar.*, à paraître.

[28] D. R. Heath-Brown, Artin's conjecture for primitive roots., *Quart. J. Math. Oxford Ser.(2)* **37**,145 (1986), 27–38.

[29] T. E. Hull et A. R. Dobell, Random number generators, *SIAM Rev.* **4** (1962), 230–254.

[30] Russell Impagliazzo et Moni Naor, Efficient cryptographic schemes provably as secure as subset sum, *J. Cryptology* **9**,4 (1996), 199–216.

[31] D. E. Knuth, Construction of a random sequence, *Nordisk Tidskr. Informations-Behandling* **5** (1965), 246–250.

[32] D. E. Knuth, *The art of computer programming, Vol. 2 : Seminumerical algorithms, Addison-Wesley series in computer science and information processing*, Addison-Wesley, Massachusetts/Menlo Park, édition seconde, 1981.

[33] A. N. Kolmogorov, Three approaches to the definition of the concept "quantity of information" (Russian) , *Problemy Peredači Informacii* **1** (1965), 3–11.

[34] L. Kuipers et H. H. Niederreiter, *Uniform distribution of sequences*, John Wiley & Sons, New York, 1974.

[35] E. L. Lehmann, *Testing statistical hypotheses*, John Wiley & sons, 1959.

[36] D. H. Lehmer, Mathematical methods in large-scale computing units, *Proceedings of a Second Symposium on Large-Scale Digital Calculating Machinery*, 1949, 141–146.

[37] M. B. Levin, The uniform distribution of the sequence $\{\alpha\lambda^x\}$. (Russian), *Mat. Sb. (N.S.)* **98** (1975), 207–222.

[38] M. Luby, *Pseudorandomness and cryptographic applications*, Princeton university press, New Jersey, 1996.

[39] G. Marsaglia, Random numbers fall mainly in the planes, *Proc. Nat. Acad. Sci. U.S.A.* **61** (1968), 25–28.

[40] G. Marsaglia, A current view of random number generation, *Proceedings of the Sixteenth Syposium on the Interface*, 1985, 3–10.

[41] P. Martin-Löf, The definition of random sequences, *Information and Control* **9** (1966), 602–619.

[42] Christian Mauduit, Joël Rivat et András Sárkőzy, On the pseudo-random properties of n^c., *Illinois J. Math.* **46**,1 (2002), 185–197.

[43] CHRISTIAN MAUDUIT ET ANDRÁS SÁRKÖZY, On finite pseudorandom binary sequences I : Measure of pseudorandomness, the Legendre symbol, *Acta Arithmetica* **82** (1997), 365–377.

[44] CHRISTIAN MAUDUIT ET ANDRÁS SÁRKÖZY, On finite pseudorandom binary sequences. II. The Champernowne, Rudin-Shapiro, and Thue-Morse sequences, a further construction., *J. Number Theory* **73**,2 (1998), 256–276.

[45] CHRISTIAN MAUDUIT ET ANDRÁS SÁRKÖZY, On finite pseudorandom binary sequences. V. On $(n\alpha)$ and $(n^2\alpha)$ sequences., *Monatsh. Math.* **129**,3 (2000), 197–216.

[46] CHRISTIAN MAUDUIT ET ANDRÁS SÁRKÖZY, On finite pseudorandom binary sequences. VI. On $(n^k\alpha)$ sequences., *Monatsh. Math.* **130**,4 (2000), 281–298.

[47] A. J. MENEZES, P. C. VAN OORSCHOT ET S. A. VANSTONE, *Handbook of Applied Cryptography*, CRC Press, édition cinquième, 2001. voir aussi http ://www.cacr.math.uwaterloo.ca/hac/.

[48] H. NIEDERREITER, Quasi-Monte Carlo methods and pseudo-random numbers, *Bull. Amer. Math. Soc.* **84**,6 (1978), 957–1041.

[49] H. NIEDERREITER, *Random number generation and quasi-Monte Carlo methods*, SIAM, Philadelphia, 1992.

[50] I. NIVEN ET H. S. ZUCKERMAN, On the definition of normal numbers, *Pacific J. Math.* **1** (1951), 103–109.

[51] SHEA-MING OON, Pseudorandom properties of prime factors, *Periodica Mathematica Hungarica* **49**,2 (2004), 107 – 118.

[52] J. B. PLUMSTEAD, Inferring a sequence generated by a linear congruence, *23rd annual symposium on foundations of computer science IEEE*, 1982, 153–159.

[53] JOËL RIVAT, On pseudo-random properties of $P(n)$ and $P(n+1)$., *Period. Math. Hungar.* **43**,1-2 (2001), 121–136.

[54] JOËL RIVAT ET ANDRÁS SÁRKÖZY, On pseudorandom binary sequences and their applications., *preprint*, à paraître.

[55] MICHAEL RUBINSTEIN ET PETER SARNAK, Chebyshev's bias, *Experimental Mathematics* **3**,3 (1994), 173–197.

[56] W. SCHMIDT, *Equations over finite fields : an elemantary approach*, *Lecture notes in Mathematics* volume 536, Springer-Verlag, New-York, 1976.

[57] W. M. SCHMIDT, Irregularities of distribution. VII, *Acta Arith.* **2** (1972), 45–50.

[58] VICTOR SHOUP, Lower bounds for discrete logarithms and related problems, *Lecture Notes in Comput. Sci.* **1233** (1997), 256–266. (Advances in cryptology—EUROCRYPT '97).

[59] DOUGLAS STINSON, *Cryptographie : Theorie et Pratique*, Vuibert, édition seconde, 2003. version fraçaise.

[60] ANDRÁS SÁRKÖZY, A finite pseudorandom binary sequence., *Studia Sci. Math. Hungar.* **38** (2001), 377–384.

[61] R. C. TAUSWORTHE, Random numbers generated by linear recurrence modulo two, *Math. Comp.* **19** (1965), 201–209.

[62] GÉRALD TENENBAUM, *Introduction à la théorie analytique et probabiliste des nombres*, Société mathématique de France, 1995.

[63] GÉRALD TENENBAUM ET MICHEL MENDÈS-FRANCE, *Les nombres premiers*, collection Que sais-je ?, 571, Presses universitaires de France, 1997.

[64] J.D. VAALER, Some extremal functions in Fourier analysis, *Bull. Amer. Math. Soc.* **12**,2 (1985), 183–216.

[65] U. V. VAZIRANI ET V. V. VAZIRANI, Efficient and secure pseudorandom number generation, *Advances in cryptology*, 1984, 193–202.

[66] A. WEIL, *Sur les courbes algébriques et les variétés qui s'en déduisent*, *Act. Sci. Ind.* volume 1041, Hermann, Paris, 1948.

[67] A. C. YAO, Theory and applications of trapdoor functions, *23rd annual symposium on foundations of computer science, IEEE*, 1982, 80–91.

Index

Monsieur **OON Shea Ming**

DOCTORAT de l'UNIVERSITE HENRI POINCARE, NANCY 1

en Mathématiques

VU, APPROUVÉ ET PERMIS D'IMPRIMER N° 1077

Nancy, le 6 septembre 2005

Le Président de l'Université

UNIVERSITÉ HENRI POINCARÉ, NANCY I
LE PRÉSIDENT

J.P. FINANCE

Université Henri Poincaré, Nancy I
24-30 rue Lionnois - B.P. 3069 - 54013 NANCY Cédex
Tél. : 03 83 68 20 00 - Fax : 03 83 68 21 00

Résumé

Cette thèse porte sur la construction de certaines suites pseudo-aléatoires inspirées par les questions naturelles en théorie des nombres. Nous utilisons les deux principales mesures introduites par A. Sárközy et C. Mauduit, à savoir la mesure de bonne distribution et la mesure de corrélation de l'ordre ℓ pour étudier quelques aspects des tests a priori de ces suites. Grâce à des résultats dus à A. Weil, certains caractères de Dirichlet fournissent une large famille d'exemples de constructions intéressantes. En revanche, l'étude de la distribution des plus grands facteurs ne nous donne pas une estimation suffisamment exploitable. Cependant, on constate numériquement qu'il y a un biais sur certaines classes de facteurs premiers. On discute aussi quelques aspects probabilistes de ces mesures. On présente également une brève histoire sur le thème du hasard. Certains sujets relatifs à la cryptologie sont aussi rappelés dans une annexe.

Mot-clés : suites pseudo-aléatoires - bonne distribution - corrélation - sommes de caractères - approximation trigonométrique - entiers friables.

Abstract

This thesis presents some constructions of pseudo-random sequences inspired by natural questions in number theory. We use two measures introduced by A. Sárközy et C. Mauduit to discuss some aspects of a priori testing of these sequences. They are the well-distribution measure and correlation measure of order ℓ. On the one hand, thanks to a work of A. Weil, some Dirichlet characters give a large family of interesting examples of constructions. On the other hand, our study on a construction based on the distribution of the greatest prime factors do not supply any sufficiently exploitable estimate. However, we observe the bias on some congruence classes of prime factors. We also discuss some probability aspects of both measures. A brief history on the randomness is presented to help better comprehension, as well as some subjects in cryptology which are given in an appendix.

Key words : pseudo-random sequences - well distribution - correlation - character sums - trigonometric approximation - friable numbers.