Trois approches de la définition quantitative de l'information

A. N. Kolmogorov

Il existe deux approches courantes pour la définition quantitative de l'"information" : combinatoire et probabiliste. L'auteur décrit brièvement les principales caractéristiques de ces approches et introduit une nouvelle approche algorithmique qui utilise la théorie des fonctions récursives.

1. L'approche combinatoire

Supposons qu'une variable x puisse prendre des valeurs dans un ensemble fini X contenant N éléments. On dit que l'"entropie" de la variable x est

$$H(x) = \log_2 N.$$

En donnant à x une valeur définie

$$x = a$$

on "supprime" cette entropie et on communique de l'"information"

$$I = \log_2 N.$$

Si les variables $x_1, x_2, \ldots, x_k$ peuvent prendre indépendamment des valeurs dans des ensembles contenant respectivement $N_1, N_2, \ldots, N_k$ éléments, alors

$$H(x_1, x_2, \ldots, x_k) = H(x_1) + H(x_2) + \ldots + H(x_k). \tag{1}$$

La transmission d'une quantité d'information I nécessite

$$I' = \begin{cases} I \text{ pour } I \text{ entier} \\ \\ [I] + 1 \text{ pour } I \text{ non entier} \end{cases}$$

chiffres binaires. Par exemple, le nombre de "mots" différents composés de k zéros et uns et un seul deux est

$$2^k(k+1).$$

Par conséquent, le contenu informationnel d'un tel message est

$$I = k + \log_2 k + 1$$

c'est-à-dire que le "codage" de tels mots dans un système purement binaire nécessite[1]

$$I' \approx k + \log_2 k$$

zéros et uns.

Référence : Problemy Peredachi Informatsii, Vol. 1, No. 1, pp. 3-11, 1965.
Transcription et traduction : Denise Vella-Chemla, octobre 2025.

[1] Ici et dans ce qui suit, $f \approx g$ indique que la différence $f - g$ est bornée, tandis que $f \sim g$ indique que le rapport $f : g$ tend vers l'unité.

Les discussions sur la théorie de l'information n'abordent généralement pas en détail cette approche combinatoire, mais il me semble important de souligner son indépendance logique par rapport aux hypothèses probabilistes. Supposons, par exemple, que nous soyons confrontés au problème du codage d'un message écrit dans un alphabet composé de s lettres, sachant que les fréquences

$$p_r = \frac{s_r}{s} \tag{2}$$

d'occurrence de lettres individuelles dans un message de longueur n satisfont l'inégalité

$$\chi = -\sum_{r=1}^{s} p_r \log_2 p_r \leq h. \tag{3}$$

Il est facile de constater que pour n grand, le logarithme binaire du nombre de messages satisfaisant à l'exigence (2) a pour estimation asymptotique

$$H = \log_2 N \sim nh.$$

Pour transmettre de tels messages, il suffit donc d'utiliser environ nh chiffres binaires.

Une méthode de codage universelle permettant la transmission de tout message suffisamment long dans un alphabet de s lettres ne comportant pas plus de nh chiffres binaires n'est pas nécessairement excessivement complexe ; en particulier, il n'est pas indispensable de commencer par déterminer les fréquences p_r pour l'ensemble du message. Pour bien comprendre cela, il suffit de noter qu'en décomposant le message S en m segments $S_1, S_2, \ldots, S_m$, on obtient l'inégalité

$$\chi \geq \frac{1}{n}[n_1\chi_1 + n_2\chi_2 + \ldots + n_m\chi_m]. \tag{4}$$

Cependant, je n'entrerai pas ici dans les détails de ce problème particulier. Il me suffit de montrer que les problèmes mathématiques associés à une approche purement combinatoire de la mesure de l'information ne se limitent pas à des trivialités.

Il est parfaitement naturel d'adopter une approche purement combinatoire de la notion d'"entropie du langage" si l'on considère une estimation de sa "flexibilité", un indice de la diversité des possibilités de développement d'une langue avec un dictionnaire et des règles de construction de phrases donnés. M. Ratner et N. Svetlova ont obtenu l'estimation suivante pour le logarithme binaire du nombre N de textes russes de longueur n, exprimé comme le "nombre de symboles, espaces compris", composé de mots du dictionnaire russe S. I. Ozhegov, sous réserve uniquement de l'exigence de "correction grammaticale".

$$h = \frac{\log_2 N}{n} = 1,9 \pm 0,1.$$

Cette valeur est considérablement supérieure à l'estimation supérieure de l'"entropie des textes littéraires" pouvant être obtenue par diverses méthodes de "deviner les suites". Cet écart est tout à fait naturel, car les textes littéraires doivent satisfaire à de nombreuses exigences allant au-delà de la simple "correction grammaticale".

Il est plus difficile d'estimer l'entropie combinatoire de textes soumis à des contraintes définies et plus élaborées. Il serait, par exemple, intéressant d'estimer l'entropie de textes russes pouvant être

considérés comme des traductions suffisamment précises (en termes de contenu) d'un texte étranger donné. Seule l'"entropie résiduelle" permet de traduire de la poésie, où le "coût entropique" du respect d'une mesure et d'un système de rimes donnés peut être calculé avec une certaine précision. On peut montrer que le tétramètre iambic rimé classique, avec certaines contraintes naturelles sur la fréquence des syllabes, etc., requiert une liberté de traitement du matériel verbal caractérisée par une "entropie résiduelle" de l'ordre de 0,4 (cette estimation est basée sur la méthode ci-dessus de mesure de la longueur d'un texte en termes de "nombre de symboles, espaces compris"). D'un autre côté, si l'on tient compte du fait que les limitations stylistiques d'un genre particulier réduisent probablement l'estimation ci-dessus de l'entropie "totale" de 1,9 à seulement $1,1-1,2$, la situation devient remarquable tant dans le cas de la traduction que dans celui de la poésie originale.

J'espère que le lecteur à tendance utilitaire me pardonnera cet exemple, mais il convient de noter que le problème plus large de la mesure de l'information liée à l'effort créatif humain est de la plus haute importance.

À ce stade, passons à une discussion sur le degré selon lequel une approche purement combinatoire permet d'estimer l'information transmise par une variable x par rapport à une variable y associée. La relation entre les variables x et y, qui prennent respectivement des valeurs dans les ensembles X et Y, consiste en ce que tous les couples (x, y) appartenant au produit cartésien $X \times Y$ ne sont pas "possibles". L'ensemble U des couples possibles détermine l'ensemble Y_a de y tel que, pour un $a \in X$ donné,

$$(a, y) \in U.$$

Il est naturel de définir l'entropie conditionnelle par l'équation

$$H(y/a) = \log_2 N(Y_a) \tag{5}$$

(où $N(Y_X)$ est le nombre d'éléments de Y_X) et l'information véhiculée par x par rapport à y par la formule

$$I(x : y) = H(y) - H(y/x). \tag{6}$$

Pour le cas présenté dans le tableau, par exemple, nous avons

	1	2	3	4
1	+	+	+	+
2	+	−	+	−
3	−	+	−	−

$$I(x = 1 : y) = 0,$$
$$I(x = 2 : y) = 1,$$
$$I(x = 3 : y) = 2,$$

Clairement, $H(y/x)$ et $I(x : y)$ sont des fonctions de x (tandis que y prend la forme d'une "variable liée").

Il n'est pas difficile d'introduire dans une conception purement combinatoire la notion de "quantité d'information nécessaire pour désigner un objet x, compte tenu des exigences données imposées à la précision de la désignation". (Voir à ce sujet la littérature abondante sur l'"ε-entropie" des ensembles dans les espaces métriques.)

Il est évident que

$$H(x/x) = 0, \qquad I(x : x) = H(x). \tag{7}$$

2. L'approche probabiliste

Les avantages potentiels d'un développement plus poussé de la théorie de l'information sur la base des définitions (5) et (6) ont été éclipsés par le fait que, si l'on considère les variables x et y comme des "variables aléatoires" avec des distributions de probabilités conjointes données, on obtient un système de concepts et de relations considérablement plus riche. En parallèle avec les quantités introduites au § 1, on obtient ici

$$H_W(x) = -\sum_x p(x) \log_2 p(x) \tag{8}$$

$$H_W(y/x) = -\sum_y p(y/x) \log_2 p)(y/x) \tag{9}$$

$$I_W(x : y) = H_W(y) - H_W(y/x) \tag{10}$$

Comme précédemment, $H_W(y/x)$ et $I_W(x : y)$ sont des fonctions de x, et nous avons les inégalités

$$H_W(x) \leq H(x), \quad H_W(y/x) \leq H(y/x) \tag{11}$$

où l'égalité est vérifiée lorsque les distributions correspondantes (sur X et Y_x) sont uniformes. Les quantités $I_W(x : y)$ et $I(x : y)$ ne sont pas liées par une inégalité de sens particulier. Comme dans le § 1,

$$H_L(x/x) = 0, \quad I_L(x : x) = H_L(x) \tag{12}$$

tandis que la quantité

$$I_L(x, y) = MI_L(x : y) = MI_L(y : x) \tag{13}$$

caractérise symétriquement la "proximité de la relation" entre x et y.

Cependant, il convient de noter que l'approche probabiliste engendre un paradoxe : dans l'approche combinatoire, $I(x : y)$ est toujours positif, ce qui est naturel dans une conception naïve du contenu informationnel, mais $I_W(x : y)$ peut être négatif. Or, seule la quantité moyenne $I_W(x, y)$ est une mesure fidèle du contenu informationnel.

L'approche probabiliste est naturelle dans la théorie de la transmission d'informations sur des canaux de communication transportant des informations "en vrac" constituées d'un grand nombre de messages sans rapport ou faiblement liés, obéissant à des lois probabilistes définies. Dans ce type de problème, il existe une tendance bénigne et (dans les travaux appliqués) profondément ancrée à mélanger probabilités et fréquences au sein d'une séquence temporelle suffisamment longue (ce qui est rigoureusement justifié si l'on suppose que le mélange est suffisamment rapide). En pratique, par exemple, on peut supposer que le problème de la détermination de l'"entropie" d'un flux de télégrammes de félicitations et de la "capacité" du canal nécessaire à une transmission ponctuelle

et sans distorsion est valablement représenté par un traitement probabiliste, même en substituant habituellement des fréquences empiriques aux probabilités. Si un problème se pose ici, le problème réside dans le flou de nos conceptions de la relation entre la théorie mathématique des probabilités et les événements aléatoires réels en général.

Mais quel sens y a-t-il, par exemple, à se demander quelle quantité d'information est contenue dans "*Guerre et Paix*" ? Est-il raisonnable d'inclure ce roman dans l'ensemble des "romans possibles", voire de postuler une distribution de probabilité pour cet ensemble ? Ou, au contraire, devons-nous supposer que les scènes individuelles de ce livre forment une séquence aléatoire avec des "relations stochastiques" qui s'atténuent assez rapidement sur plusieurs pages ?

En réalité, nous sommes tout aussi peu renseignés sur la question à la mode de la "quantité d'information héréditaire" nécessaire, par exemple, à la reproduction d'une forme particulière de *cafard*. Cependant, dans les limites de l'approche probabiliste, deux variantes sont possibles. Dans la première variante, nous devons considérer l'ensemble des "formes possibles" avec une distribution de probabilité d'origine incertaine sur cet ensemble. Dans la seconde variante, les caractéristiques de la forme sont supposées être un ensemble de variables aléatoires faiblement dépendantes. La nature réelle du mécanisme de mutation fournit des arguments en faveur de la deuxième variante, mais ces arguments sont remis en cause si nous supposons que la sélection naturelle provoque l'apparition d'un système de caractéristiques cohérentes.

3. Une approche algorithmique

En réalité, il est plus fructueux d'étudier la quantité d'information "transmise par un objet" (x) "à propos d'un objet" (y). Ce n'est pas un hasard si, dans l'approche probabiliste, cela a conduit à une généralisation au cas des variables continues, pour lesquelles l'entropie est infinie, mais, dans un grand nombre de cas,

$$I_W(x,y) = \iint P_{xy}(dx\ dy) \log_2 \frac{P_{xy}(dx\ dy)}{P_x(dx)P_y(dy)}$$

est fini. Les objets réels que nous étudions sont très (infiniment) complexes, mais les relations entre deux objets distincts diminuent à mesure que les schémas utilisés pour les décrire se simplifient. Alors qu'une carte fournit une quantité considérable d'informations sur une région de la surface terrestre, la microstructure du papier et l'encre sur le papier n'ont aucun rapport avec la microstructure de la zone représentée sur la carte.

En pratique, nous nous intéressons le plus souvent à la quantité d'informations "*transmise par un objet individuel x sur un objet individuel y*". Il est vrai, comme nous l'avons déjà noté, qu'une telle estimation quantitative individuelle de l'information n'a de sens que lorsque la quantité d'informations est suffisamment importante. Il est, par exemple, vain de s'interroger sur la quantité d'informations transmises par la séquence

0 1 1 0

à propos de la séquence

1 1 0 0.

Mais si l'on prend un tableau parfaitement précis de nombres aléatoires, du type couramment utilisé en statistique, et que l'on écrive pour chacun de ses chiffres le chiffre de l'unité de son carré selon le schéma

0 1 2 3 4 5 6 7 8 9
0 1 4 9 6 5 6 9 4 1,

le nouveau tableau contiendra environ

$$\left(\log_2(10) - \frac{8}{10}\right) n$$

bits d'information sur la séquence initiale (où n est le nombre de chiffres du tableau). Par conséquent, nous proposons de définir ci-dessous

$$I_A(x\ :\ y)$$

de sorte qu'une certaine indétermination subsiste. Différentes variantes équivalentes de cette définition conduiront à des valeurs équivalentes uniquement au sens où $I_{A_1} \approx I_{A_2}$, c'est-à-dire

$$|I_{A_1} - I_{A_2}| \leq C_{A_1\,A_2},$$

où la constante $C_{A_1\,A_2}$ dépend des deux manières fondamentales de définir les méthodes universelles de programmation A_1 et A_2. Considérons un "domaine indexé d'objets", c'est-à-dire un ensemble dénombrable

$$X = \{x\}$$

avec une suite finie $n(x)$ de zéros et de uns, commençant par un un, associée à chaque élément comme indice. Notons la longueur de la suite $n(x)$ par $l(x)$ et supposons que :

1) La correspondance entre X et l'ensemble D des suites binaires de la forme décrite ci-dessus est bijective ;

2) $D \subset X$, la fonction $n(x)$ sur D est généralement récursive [1], et pour $x \in D$,

$$l(n(x)) \leq l(x) + C$$

où C est une constante ;

3) Avec x et y, l'ensemble X contient le couple ordonné (x, y), dont l'indice est une fonction généralement récursive des indices de x et y, et

$$l(x, y) \leq C_x + l(y)$$

où C_X ne dépend que de X.

Ces exigences ne sont pas toutes essentielles, mais elles simplifient la discussion. Le résultat final de la construction est invariant lors de la transition vers un nouvel index $n'(x)$ possédant les mêmes propriétés que l'ancien système et pouvant généralement être exprimé récursivement en fonction de celui-ci ; de plus, il conserve ses propriétés lorsqu'il est intégré dans un système plus grand X' (à condition que, pour les éléments du système initial, l'indice n' dans le système étendu puisse

généralement être exprimé récursivement en fonction de l'indice initial n). La nouvelle "complexité" K et la quantité d'information restent équivalentes lors de ces transformations, au sens de $\approx$.

Comme "complexité relative d'un objet y de x donné, nous prendrons la longueur minimale $l(p)$ du "programme" p permettant d'obtenir y à partir de x. La définition ainsi formulée dépend de la "méthode de programmation", qui n'est autre que la fonction

$$\varphi(p, x) = y$$

qui associe un objet y à un programme p et un objet x.

Conformément aux conceptions désormais universellement acceptées en logique mathématique moderne, nous devons supposer que la fonction φ est partiellement récursive. Pour toute fonction de ce type, nous avons

$$K_\varphi(y/x) = \begin{cases} \min\limits_{\varphi(p,x)=y} l(p) \\ \\ \infty, \text{ s'il n'existe aucun } p \text{ tel que } \varphi(p, x) = y. \end{cases}$$

Dans ce cas, une fonction

$$v = \varphi(u)$$

de $u \in X$ de portée $v \in X$ est dite partiellement récursive si elle génère une fonction partiellement récursive de la transformation d'indice.

$$n(v) = \Psi[n(u)]$$

Pour comprendre la définition, il est important de noter qu'en général, les fonctions partiellement récursives ne sont pas définies partout et qu'il n'existe pas de méthode fixe permettant de déterminer si l'application du programme p à un objet k produit un résultat. Par conséquent, la fonction $K_\varphi(y/x)$ (généralement récursive) ne peut être calculée efficacement, même si elle est connue comme finie pour tout x et y.

Théorème fondamental. *Il existe une fonction partiellement récursive $A(p, x)$ telle que, pour toute autre fonction partiellement récursive $\varphi(p, x)$, on ait l'inégalité*

$$K_A(y/x) \leq K_\varphi(y/x) + C_\varphi,$$

où la constante C_φ ne dépend ni de x ni de y.

La démonstration repose sur l'existence d'une fonction partiellement récursive *universelle*

$$\Phi(n, u)$$

qui possède la propriété qu'en fixant un indice n approprié, on peut utiliser la formule

$$\varphi(u) = \Phi(n, u)$$

pour obtenir toute autre fonction partiellement récursive. La fonction $A(p,x)$ dont nous avons besoin est donnée par la formule[2]

$$A((n,q),x) = \Phi(n,(q,x))$$

En effet, si $y = \varphi(p,x) = \Phi(n(p,x))$ alors

$$A((n,p),x) = y$$

$$l(n,p) \leq l(p) + C_n$$

Nous appellerons les fonctions $A(p,x)$ qui satisfont aux exigences du théorème fondamental (et aux méthodes de programmation qu'il définit) *asymptotiquement optimales*. Il est clair que la "complexité" correspondante $K_A(y/x)$ est finie pour tout x et y. Pour deux de ces fonctions A_1 et A_2,

$$|K_{A_1}(y/x) - K_{A_2}(y/x)| \leq C_{A_1\,A_2},$$

où $C_{A_1\,A_2}$ ne dépend pas de x et y, c'est-à-dire $K_{A_1}(y/x) \approx K_{A_2}(y/x)$. Finalement,

$$K_A(y) = K_A(y/1)$$

peut être considéré comme la "complexité de y" et nous pouvons définir la "quantité d'information véhiculée par x sur y" par la formule

$$I_A(x\ :\ y) = K_A(y) - K_A(y/x)$$

Il est facile de montrer[3] que cette quantité est toujours essentiellement positive,

$$I_A(x\ :\ y) \gtrsim 0$$

ce qui signifie que $I_A(x\ :\ y)$ n'est pas inférieur à une constante négative C qui ne dépend que des caractéristiques de la méthode de programmation choisie. Comme nous l'avons déjà noté, le théorème a été conçu pour une application à une quantité d'information si grande que, comparativement, $|C|$ est négligeable.

Notons enfin que $K_A(x/x) \approx 0,\ \ I_A(x\ :\ x) = K_A(x)$.

Bien sûr, on peut éviter les indéterminations associées à la constante c_φ, etc., en considérant des domaines particuliers des objets X, l'indexation et la fonction A, mais il est douteux que cela soit fait sans arbitraire explicite. On doit, cependant, supposer que les différentes variantes "raisonnables" présentées ici conduiront à des "estimations de complexité" qui convergeront sur des centaines de bits au lieu de dizaines de milliers. Par conséquent, des quantités telles que la "complexité" du texte de "*Guerre et Paix*" peuvent être supposées définies avec ce qui équivaut à l'unicité. Des expériences sur la devinette des suites de textes littéraires permettent d'obtenir une estimation supérieure de la

[2] $\Phi(n,u)$ est définie uniquement lorsque $n \in D$, et $A(p,x)$ est définie uniquement lorsque p est de la forme $(n,q), n \in D$.

[3] En choisissant une "fonction de comparaison de $\varphi(p,x) = A(p,1)$, nous obtenons $K_A(y/x) \leq K_\varphi(y/x) + C_\varphi = K_A(y) + C_\varphi$.

complexité conditionnelle en présence d'une consommation donnée d'"informations a priori" (sur la langue, le style, le contenu textuel) à disposition du devineur. Lors de tests menés au Département de théorie des probabilités de l'Université d'État de Moscou, ces estimations supérieures ont fluctué entre 0,9 et 1,4. Les estimations de l'ordre de 0,9 à 1,1 obtenues par N. G. Rychkov ont conduit des devineurs moins savants à suggérer qu'il communiquait par télépathie avec les auteurs des textes.

Je pense que l'approche proposée ici donne, en principe, une définition correcte de la "quantité d'information héréditaire", bien qu'il soit difficile d'obtenir une estimation fiable de cette quantité.

4. Conclusion

Les concepts abordés dans le § 3 présentent un inconvénient majeur : ils ne tiennent pas compte de la "difficulté" de préparation d'un programme p pour passer d'un objet x à un objet y. En introduisant des définitions appropriées, il est possible de prouver des propositions mathématiques rigoureusement formulées, pouvant être légitimement interprétées comme une indication de l'existence de cas où un objet permettant un programme très simple, c'est-à-dire de très faible complexité $K(x)$, ne peut être restauré par des programmes courts que grâce à des calculs d'une durée totalement irréelle. J'ai l'intention d'étudier ultérieurement la relation entre la complexité nécessaire d'un programme

$$K^t(x)$$

et sa difficulté admissible t. La complexité $K(x)$ obtenue dans le § 3 est, dans ce cas, le minimum de $K^t(x)$ après suppression des contraintes sur t.

L'utilisation des constructions du § 3 pour fournir une nouvelle base à la théorie des probabilités dépasse le cadre de cet article. En gros, la situation est la suivante : si un ensemble fini M contenant un très grand nombre d'éléments N peut être déterminé par un programme de longueur négligeable devant $\log_2 N$, alors presque tous les éléments de M ont une complexité $K(x)$ proche de $\log_2 N$. Les éléments $x \in M$ de cette complexité sont également traités comme des éléments "aléatoires" de l'ensemble M. Une discussion incomplète de cette idée peut être trouvée dans [2].

Références

1. V. A. Uspenskii, *Lectures on Computable Functions* [in Russian], Fizmatgiz, Moscou, 1960.

2. A. N. Kolmogorov, "*On tables of random numbers*", Sankhya. The Indian Journal of Statistics, Séries A, 25, 4, 369-376, 1963.

9 JANVIER 1965.

Traduction d'un extrait du livre *Probability and measure* (section Convergence de distributions) de Patrick Billingsley

Application à la théorie des nombres

Soit $g(m)$ le nombre de facteurs premiers distincts de l'entier m ; par exemple, $g(3^4 \times 5^2) = 2$. Puisqu'il existe une infinité de nombres premiers, $g(m)$ est illimité au-dessus ; de même, il retombe à 1 pour une infinité de m (pour les nombres premiers et leurs puissances). Puisque g fluctue de façon irrégulière, il est naturel de s'intéresser à son comportement moyen.

Sur l'espace Ω des entiers positifs, soit P_n la mesure de probabilité qui place la masse $1/n$ à chacun des $1, 2, \ldots, n$, de sorte que parmi les n premiers entiers positifs, la proportion contenue dans un ensemble A donné soit juste $P_n(A)$. Le problème consiste à étudier $P_n[m : g(m) \leq x]$ pour un grand n.

Si $\delta_p(m)$ vaut 1 ou 0 selon que le nombre premier p divise m ou non, alors

$$g(m) = \sum_p \delta_p(m). \tag{30.12}$$

La théorie des probabilités peut être utilisée pour étudier cette somme car sous P_n les $\delta_p(m)$ se comportent un peu comme des variables aléatoires indépendantes. Si $p_1, \ldots, p_u$ sont des nombres premiers distincts, alors par le théorème fondamental de l'arithmétique, $\delta_{p_1}(m) = \ldots = \delta_{p_u}(m) = 1$, c'est-à-dire que chaque p_i divise m si et seulement si le produit $p_1 \ldots p_u$ divise m. La probabilité sous P_n de ceci est juste n^{-1} fois le nombre de m dans l'intervalle $1 \leq m \leq n$ qui sont des multiples de $p_1 \ldots p_u$, et ce nombre est la partie entière de $n/p_1 \ldots p_u$. Ainsi

$$P_n[m : \delta_{p_i}(m) = 1, \quad i = 1, \ldots, u] = \frac{1}{n} \left\lfloor \frac{n}{p_1 \ldots p_u} \right\rfloor. \tag{30.13}$$

pour p_i distinct.

Soit maintenant X_p des variables aléatoires indépendantes (sur un espace de probabilités, une variable pour chaque nombre premier p) vérifiant

$$P[X_p = 1] = \frac{1}{p}, \quad p[X_p = 0] = 1 - \frac{1}{p}.$$

Si $p_1, \ldots, p_u$ sont distincts, alors

$$P[X_{p_i} = 1, \quad i = 1, \ldots, u] = \frac{1}{p_1 \ldots p_u}. \tag{30.14}$$

Pour $p_1, \ldots, p_u$ fixé, (30.13) converge vers (30.14) lorsque $n \to \infty$. Ainsi, le comportement de X_p peut servir de guide pour celui de $\delta_p(m)$. Si $m \leq n$, (30.12) est $\sum_{p \leq n} \delta_p(m)$, car aucun nombre premier supérieur à m ne peut le diviser. L'idée est de comparer cette somme à la somme correspondante

Transcription et traduction : Denise Vella-Chemla, octobre 2025.

$\sum_{p\leq n} X_p$.

Ceci nécessitera, de la théorie des nombres, l'estimation élémentaire[4]

$$\sum_{p\leq x} \frac{1}{p} = \log\log x + \mathrm{O}(1) \tag{30.15}$$

La moyenne et la variance de $\sum_{p\leq n} X_p$ sont $\sum_{p\leq n} p^{-1}$ et $\sum_{p\leq n} p^{-1}(1-p^{-1})$; puisque $\sum_p p^{-2}$ converge, chacune de ces deux sommes est asymptotiquement $\log\log n$. Comparer $\sum_{p\leq n} \delta_p(m)$ avec $\sum_{p\leq n} X_p$ conduit alors à conjecturer le *théorème central limite d'Erdös-Kac pour la fonction diviseur premier* :

Théorème 30.3

Pour tout x,

$$P_n\left[m : \frac{g(m) - \log\log n}{\sqrt{\log\log n}} \leq x\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}. \tag{30.16}$$

Preuve. L'argumentation utilise la méthode des moments. La première étape consiste à montrer que (30.16) n'est pas affectée si le domaine de p dans (30.12) est encore restreint. Soit $\{\alpha_n\}$ une suite tendant vers l'infini suffisamment lentement pour que

$$\frac{\log \alpha_n}{\log n} \to 0 \tag{30.17}$$

mais suffisamment rapide pour que

$$\sum_{\alpha_n < p \leq n} \frac{1}{p} = \mathrm{o}(\log\log n)^{1/2} \tag{30.18}$$

En raison de (30.15), ces deux exigences sont satisfaites si, par exemple, $\log \alpha_n = (\log n)/\log\log n$.

Définissons maintenant

$$g_n(m) = \sum_{p\leq \alpha_n} \delta_p(m). \tag{30.19}$$

Pour une fonction f d'entiers positifs, soit

$$E_n[f] = n^{-1} \sum_{m=1}^{n} f(m)$$

désigne son espérance mathématique calculée par rapport à P_n. Par (30.13) pour $u = 1$;

$$E_n\left[\sum_{p>\alpha_n} \delta\right] = \sum_{\alpha_n < p \leq n} P_n[m \,:\, \delta_p(m) = 1] \leq \sum_{\alpha_n < p \leq n} \frac{1}{p}.$$

[4]Voir, par exemple, Hardy et Wright, chapitre XXII.

D'après (30.18) et l'inégalité de Markov,

$$P_n[m \,:\, |g(m) - g_n(m)| \geq \epsilon(\log\log n)^{1/2}] \to 0.$$

Par conséquent (théorème 25.4), (30.16) n'est pas affecté si $g_n(m)$ remplace $g(m)$.

Comparons maintenant (30.19) à la somme correspondante $S_n = \sum_{p\leq\alpha_n} X_p$. La moyenne et la variance de S_n sont :

$$c_n = \sum_{p\leq\alpha_n} \frac{1}{p}, \quad s_n^2 = \sum_{p\leq\alpha_n} \frac{1}{p}\left(1 - \frac{1}{p}\right),$$

et chacune est égale à $\log\log n + o(\log\log n)^{1/2}$ par (30.18). Ainsi (voir l'exemple 25.8), (30.16) avec $g(m)$ remplacé comme ci-dessus équivaut à

$$P_n\left[m \,:\, \frac{g_n(m) - c_n}{s_n} \leq x\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du. \tag{30.20}$$

Il suffit donc de prouver (30.20).

Puisque les X_p sont bornés, l'analyse des moments (30.7) s'applique ici. La seule différence est que les termes de S_n sont indexés non pas par les entiers k compris dans l'intervalle $k \leq k_n$ mais par les nombres premiers p compris dans l'intervalle $p \leq \alpha_n$; de plus, X_p doit être remplacé par $X_p - p^{-1}$ pour le centrer. Ainsi, le $r^{ième}$ moment de $(S_n - c_n)/s_n$ converge vers celui de la distribution normale, et donc (30.20) et (30.16) suivront, par la méthode des moments, si l'on montre que lorsque $n \to \infty$.

$$E\left[\left(\frac{S_n - c_n}{s_n}\right)^r\right] - E_n\left[\left(\frac{g_n - c_n}{s_n}\right)^r\right] \to 0 \tag{30.21}$$

pour chaque r.

Maintenant, $E[S_n^r]$ est la somme

$$\sum_{u=1}^{r} {\sum}' \frac{r!}{r_1! \ldots r_u!} \frac{1}{u!} {\sum}'' E[X_{p_1}^{r_1} \ldots X_{p_u}^{ru]}, \tag{30.22}$$

où le domaine de Σ' est comme dans (30.6) et (30.7), et Σ'' est calculée sur les u-uplets $(p_1, \ldots, p_u)$ de nombres premiers distincts ne dépassant pas α_n. Puisque X_p ne prend que les valeurs 0 et 1, de l'indépendance des X_p et du fait que les p_i sont distincts, il s'ensuit que la somme dans (30.22) est

$$E[X_{p_1} \ldots X_{p_u}] = \frac{1}{p_1 \ldots p_u}. \tag{30.23}$$

Selon la définition (30.19), $E_n[g_n^r]$ est simplement (30.22) avec le terme de la somme est remplacé par $E_n[\delta_{p_1}^{r_1} \ldots \delta_{p_u}^{r_u}]$. Puisque $\delta_p(m)$ ne prend que les valeurs 0 et 1, d'après (30.13) et le fait que les p_i sont distincts, ce terme de la somme est

$$E_n[\delta_{p_1} \ldots \delta_{p_u}] = \frac{1}{n}\left\lfloor \frac{n}{p_1 \ldots p_u} \right\rfloor. \tag{30.24}$$

Mais (30.23) et (30.24) diffèrent au plus de $1/n$, et donc $E[S_n^r]$ et $E_n[g_n^r]$ diffèrent au plus de la somme (30.22), la somme étant remplacée par $1/n$. Par conséquent,

$$|E[S_n^r] - E_n[g_n^r]| \le \frac{1}{n}\left(\sum_{p\le\alpha_n} 1\right)^r \le \frac{\alpha_n^r}{n}. \tag{30.25}$$

Maintenant

$$E[(S_n - c_n)^r] = \sum_{k=0}^{r} \binom{r}{k} E[S_n^k](-c_n)^{r-k},$$

et $E_n[(g_n - c_n)^r]$ a le développement analogue. La comparaison des deux développements terme à terme et l'application de (30.25) montrent que

$$\begin{aligned} |E[(s_n - c_n)^r] - E_n[(g_n - c_n)^r]| \\ \le \sum_{k=0}^{r} \binom{r}{k} \frac{\alpha_n^k}{n} c_n^{r-k} = \frac{1}{n}(\alpha_n + c_n)^r. \end{aligned} \tag{30.26}$$

Puisque $c_n \le \alpha_n$ et que $\alpha_n^r/n \to 0$ par (30.17), (30.21) suit comme requis. □

La méthode de preuve nécessite de passer de (30.12) à (30.19). Sans cela, α_n à droite dans (30.26) vaudrait n, et il ne s'ensuivrait pas que la différence à gauche tende vers 0 ; d'où la troncature (30.19) pour un α_n suffisamment petit pour satisfaire (30.17). En revanche, α_n doit être suffisamment grand pour satisfaire (30.18), afin que la troncature laisse (30.16) inchangée.