

2. Complexités

Certaines méthodes séduisantes utilisées pour le calcul de la complexité des courtes séquences sont insuffisantes. En effet, elles ne coïncident pas avec la complexité de Kolmogorov quand la longueur des séquences augmente, ce qui est nécessaire pour que la mesure soit satisfaisante.

L'entropie de Shannon est définie par l'expression $-\sum p_i \log p_i$ (où p_i est une fréquence). Avec une suite binaire, cela donne $[-p_1 \log p_1 - p_2 \log p_2]$, où p_1 et p_2 sont respectivement les densités de 0 et de 1 dans la séquence. Avec cette mesure, la suite 0101010101010101 est la plus complexe possible pour une suite longueur 20 (elle a autant de 0 que de 1). Or la suite 10010111010100001011, qui a la même entropie de Shannon et la même longueur, est plus complexe. L'entropie de Shannon ne prend en compte que les fréquences de 0 et de 1 ; ce n'est pas suffisant pour savoir si une séquence est simple ou complexe. C'est une mesure statistique qui ne prend même pas en compte les répétitions.

La complexité par facteurs d'une suites est déterminée par le nombre de facteurs différents qu'on trouve dans s . Un facteur est une suite de symboles contigus. Ainsi, la séquence 00000 possède cinq facteurs différents : 0, 00, 000, 0000, 00000. Elle a comme complexité 5. La séquence 01011 possède les 11 facteurs différents 0, 1, 01, 10, 11, 010, 101, 011, 0101, 1011, 01011 et elle est donc plus complexe.

Cette mesure est mauvaise, car elle attribue à certaines suites régulières une grande complexité. Ainsi, la suite de Champernowne obtenue en écrivant les entiers en binaire les uns derrière les autres (0 1 10 11 100 101 110 111 1000 etc.) met en défaut la mesure par facteurs. Par exemple, la suite 0110111001011101111000 (les entiers jusqu'à 8) n'est pas complexe, car on peut la définir brièvement et pourtant, elle contient un très grand nombre de facteurs. La mesure de complexité par facteurs ne peut pas être adoptée : une mesure générale doit être équivalente à la complexité de Kolmogorov lorsque les longueurs tendent vers l'infini, or ce n'est pas du tout le cas ici.

pratique ? L'instabilité des thermomètres les plus simples (ceux fondés sur la longueur minimale des programmes) les rend inutilisables pour le classement des faibles complexités.

La solution est apparue il y a environ cinq ans. Un thermomètre nouveau, mis au point au Laboratoire d'informatique fondamentale de Lille selon les idées et les programmes de Hector Zenil, donne des réponses satisfaisantes pour les courtes et les longues séquences. Il définit une complexité conforme à l'idée générale de Kolmogorov pour les grandes complexités et classe, comme on l'attend, 0000000, 0001000, 1001101, ou toutes les séquences pour lesquelles notre intuition n'hésite pas.

Le nouveau thermomètre que nous détaillerons possède un défaut : il est extrêmement coûteux en calcul. En pratique, il ne donne de résultats que pour les séquences très courtes et, de ce point de vue, les méthodes par compression restent incontournables. Comme en astronomie où, selon le type d'objets et leur éloignement, on utilise un procédé ou un autre pour évaluer les distances, dans les mesures de complexité, on doit adopter des méthodes hybrides avec recollement de plusieurs procédés.

Un critère fondé sur la probabilité de production

L'idée sous-jacente aux thermomètres pour les basses complexités est fondée sur une propriété remarquable de la complexité de Kolmogorov : plus un objet est simple, plus il est produit fréquemment quand on fait fonctionner un ordinateur en lui donnant des programmes choisis au hasard.

Un langage de programmation universel étant donné (*Java*, *C*, *Pascal*, *Basic* sont de tels langages), nous imaginerons que nous tirons au hasard des programmes écrits dans ce langage (en fait on peut écrire les programmes en binaire et tirer un programme au hasard revient à tirer une suite de 0 et de 1 jusqu'à avoir un programme complet). Nous tomberons souvent sur des programmes grammaticalement incorrects (donc ne fonc-

tionnant pas) ou sur des programmes qui tourneront sans produire aucun résultat. Ces programmes ne nous intéressent pas : seuls ceux qui produisent en sortie une suite finie de 0 et de 1 et s'arrêtent sont pris en compte. Les expériences montrent qu'en procédant ainsi, on obtient plus souvent 0000000 que 1001101 : la probabilité d'obtenir en sortie une suite de grande complexité est plus faible que la probabilité d'obtenir une suite de faible complexité. Cette probabilité se nomme mesure de Solomonoff-Levin, en l'honneur de Ray Solomonoff (1926-2009) et Leonid Levin (professeur à l'Université de Boston), qui ont identifié ce phénomène.

Un théorème, clef des thermomètres pour les faibles complexités, fait le lien avec la complexité de Kolmogorov. Ce théorème affirme que $K[s]$ est environ égal à $-\log_2 SL[s]$, où $\log_2$ est le logarithme en base 2 (on a $\log_2(x) = \log x / \log 2$). En termes mathématiques, $K[s]$ étant une complexité de Kolmogorov définie par un langage de programmation et $SL[s]$ la mesure de Solomonoff-Levin associée au même langage, il existe une constante c , indépendante de la séquence s , telle que :

$$|-\log_2 SL[s] - K[s]| < c.$$

À l'infini, $-\log_2 SL[s]$ et $K[s]$ ne diffèrent au plus que d'une constante fixée une fois pour toutes. La quantité $-\log_2 SL[s]$ est donc aussi intéressante qu'une complexité de Kolmogorov mesurée par les plus courts programmes et coïncide presque avec elle pour les longues séquences.

Le premier avantage à considérer le nombre $-\log_2 SL[s]$ plutôt que $K[s]$ provient de ce que c'est un nombre bien plus stable. En changeant de langage de programmation, ou seulement en modifiant légèrement la définition d'un langage de programmation choisi, nous ferons beaucoup varier $K[s]$ pour les petites séquences, ce qui empêche, nous l'avons vu, de donner un sens à ce qu'on obtient. En revanche, le nombre $-\log_2 SL[s]$ ne varie que très peu. La raison est que ce nombre résulte d'un lissage opéré sur un très grand nombre de données dû au fonctionnement d'une multitude de programmes.

Cependant, l'argument le plus puissant qui fait s'intéresser à la mesure de Solo-