KOLMOGOROV COMPLEXITY ⓪

MAISON DES MATHEMATIQUES ET DE L'INFORMATIQUE

① J'AI LANCÉ UNE PIÈCE 20 FOIS ET OBTENU :

10 10 10 10 10 10 10 10 10 10

QU'EN DITES VOUS ? VOUS NE ME CROYEZ PAS.

② IDEM :

11100101100100011101

QU'EN DITES VOUS ? PEUT-ÊTRE.

③ POURQUOI ? PROBA IDENTIQUES : 2^{-20}.

LA THÉORIE CLASSIQUE DES PROBABILITÉS NE CAPTURE PAS NOTRE NOTION INTUITIVE D'"ALÉATOIRE".

REMARQUE : UN PROGRAMME GÉNÉRANT ① : "RÉPÉTER '10' 10 FOIS" DE LONGUEUR $c + \log_2(10)$.

POUR ② CELA SEMBLE IMPOSSIBLE ... INCOMPRESSIBLE ?

KOLMOGOROV COMPLEXITY ① 01/2023

MAISON DES MATHEMATIQUES ET DE L'INFORMATIQUE

UNE DES FAÇON DE DÉFINIR L'INFORMATION, ET LA QUANTIFIER : LA PLUS PETITE DESCRIPTION ALGORITHMIQUE

DÉFINITION : SOIT $x \in \{0,1\}^*$. LA DESCRIPTION MINIMALE DE x, NOTÉE $d(x)$, EST LE PLUS PETIT MOT $(\langle M \rangle, w)$ AVEC M UNE M.T. QUI SUR L'ENTRÉE w S'ARRÊTE AVEC x SUR SON RUBAN. LA COMPLEXITÉ DE KOLMOGOROV DE x EST $K(x) = |d(x)|$.

$K(x) = \min\{|\langle M \rangle w| \mid M(\varepsilon) = x\}$ MAIS LA DÉF AVEC $(\langle M \rangle, w)$ EST PRATIQUE POUR HARD-CODER DES VALEURS AVEC M CONSTANTE

TAILLE + LEX

⚠ ENCODAGE DES COUPLES (PAR EXEMPLE : $\underbrace{|u|}_{\text{DOUBLÉ}} 01 uv$ EN $2\log|u| + |u| + |v|$).

THM : $\exists c : \forall x : K(x) \leq |x| + c$.

PREUVE : AVEC LA TM M QUI S'ARRÊTE IMMÉDIATEMENT SUR TOUTE ENTRÉE ET $w := x$. □

THM : $\exists c : \forall x : K(xx) \leq K(x) + c$.

PREUVE : SOIT M LA TM QUI, SUR L'ENTRÉE $(\langle N \rangle, w)$:

1. SIMULE N SUR L'ENTRÉE w JUSQU'À L'ARRÊT AVEC s SUR LE RUBAN,
2. ÉCRIRE ss (UN DEUXIÈME MOT s) ET S'ARRÊTER.

UNE DESCRIPTION DE xx EST $(\langle M \rangle, d(x))$. □

THM : $\exists c : \forall x, y : K(xy) \leq 2.\log(K(x)) + K(x) + K(y) + c$.

COMPLEXITÉ DE LA CONCATÉNATION ; AMG...

PREUVE : $(\langle M \rangle, \underbrace{(d(x), d(y))}_{\text{COUPLE}})$ AVEC M LA MT QUI SIMULE $d(x)$ PUIS $d(y)$ POUR CALCULER xy. □

THM : $\forall c : \exists x, y : K(xy) > K(x) + K(y) + c$

PREUVE : ON A BESOIN DU RÉSULTAT SUIVANT :

THM : $\forall k$, POUR TOUT $x \in \{0,1\}^*$ SUFFISAMMENT LONG IL EXISTE UN PRÉFIXE $y \subseteq x$ TEL QUE $K(y) \leq |y| - k$. [MARTIN-LÖF RANDOM]

PREUVE : SOIT $z \subseteq x$: SOIT n L'IMAGE DE z SELON UNE BIJECTION STANDARD (LONGUEUR-LEX) ENTRE $\{0,1\}^*$ ET $\mathbb{N}$. SOIT y L'EXTENSION DE LONGUEUR n DE z SELON x, C-À-D $y = z\sigma \subseteq x$ ET $|\sigma| = n$. IL EXISTE UNE MACHINE M TELLE QUE $M(\sigma) = z\sigma$, EN UTILISANT $|\sigma|$ POUR CALCULER z. DONC $K(y) \leq |\sigma| + c$, AVEC c INDÉPENDENTE DE y (DE z ET DE σ). EN PRENANT $|z| > k + c$ ON OBTIENT

$K(y) \leq |\sigma| + c = |y| - |z| + c < |y| - k - c + c = |y| - k$. ◇

SOIT k TEL QUE $\forall x . K(x) \leq |x| + k$. SOIT z INCOMPRESSIBLE (DÉF ET EXISTENCE PAGE SUIVANTE !) SUFFISAMMENT LONG : $K(z) \geq |z|$. SOIT $\ell = c + k$ ET $x \subseteq z$ DONNÉ PAR LE THM CI-DESSUS : $K(x) < |x| - \ell$. ALORS POUR y TEL QUE $z = xy$ ON A $K(x) + K(y) + c < \underbrace{|x| - c - k}_{} + \underbrace{|y| + k}_{} + c = |z| \leq K(z)$. □

KOLMOGOROV COMPLEXITY (2).

MAISON DES MATHEMATIQUES ET DE L'INFORMATIQUE

OPTIMALITÉ DE LA DESCRIPTION.

DEF : $K_P(x)$ LA COMPLEXITÉ DE KOLMO RELATIVE AU LANGAGE DE PROG. P. (DE DESCRIPTION)

THM : $\forall P : \exists c : \forall x : K(x) \leqslant K_P(x) + c$.

PREUVE : IL FAUT QUE LA TM COMPILE+EXÉCUTE / INTERPRÈTE LA DESCRIPTION EN P.

LES MOTS INCOMPRESSIBLES SONT ALÉATOIRES

DEF : x EST c-COMPRESSIBLE LORSQUE $K(x) \leqslant |x| - c$.
x INCOMPRESSIBLE = 1-INCOMPRESSIBLE = $K(x) \not\leqslant |x|$ = PAS DE DESCRIPTION PLUS PETITE QUE LUI-MÊME.

ILS ONT DES PROP. "ALÉATOIRES" (PILE-OU-FACE).

THM : IL EXISTE DES MOTS BINAIRES INCOMPRESSIBLES DE TOUTE TAILLE.

PREUVE : COMPTAGE POUR UNE TAILLE n : "# MOTS BINAIRES" $= 2^n$. "# DESCRIPTION DE LONGUEUR $< n$" $\leqslant \sum_{i=0}^{n-1} 2^i = 2^n - 1$, MAIS DESC. DÉCRIT AU PLUS UN MOT. □

EXO : AU MOINS $2^n - 2^{n-c+1} + 1$ MOTS BINAIRES DE TAILLE n SONT c-INCOMPRESSIBLES.

THM : x-INCOMPRESSIBLE $\Rightarrow |x|_0 \simeq |x|_1$ ET PLUS LONG FACTEUR 0^k DE TAILLE $k \simeq \log_2 n$.

THM : SOIT f UNE PROPRIÉTÉ CALCULABLE VRAIE POUR PRESQUE TOUT LES MOTS ($\lim_{n\to\infty} \frac{\text{FALSE}}{2^n} = 0$).
ALORS $\forall c > 0$: LA PROP. f EST FAUSSE SUR UN NOMBRE FINI DE MOTS c-INCOMPRESSIBLES.

IDÉE DE LA PREUVE : ON OBTIENT UNE DESCRIPTION COURTE AVEC (M, i) OÙ M ÉNUMÈRE LE i-ÈME MOT NE VÉRIFIANT PAS LA PROP. □

EXEMPLES DE MOTS INCOMPRESSIBLE ?

THM : $K : \{0,1\}^* \to \mathbb{N}$ N'EST PAS CALCULABLE.

PREUVE : PAR L'ABSURDE, SUPPOSONS QUE K SOIT CALCULABLE. ALORS ON PEUT CONSTRUIRE M QUI, SUR TOUTE ENTRÉE :

1. OBTIENT SON PROPRE CODE $\langle M \rangle$
2. ÉNUMÈRE LES MOTS BINAIRES JUSQU'À TROUVER x TEL QUE $K(x) > |\langle M \rangle|$
3. ÉCRIT x ET STOP.

ALORS M DÉCRIT x MAIS EST PLUS PETITE QUE $K(x)$. ↯ □

CODEGOLF.STACKEXCHANGE.COM.

KOLMOGOROV COMPLEXITY ③

SOIT $L_K = \{x \mid K(x) \geq |x|\}$ LES INCOMPRESSIBLES

EXO : AUCUN SOUS-ENSEMBLE INFINI DE L_K (DONT L_K LUI-MÊME) N'EST SEMI-DÉCIDABLE

PREUVE : PAR L'ABSURDE : M ÉNUMÈRE JUSQU'À TROUVER x TEL QUE $K(x) > |\langle M \rangle|$ ET ÉCRIT x. ↯ □

⚠ VERSION FAIBLE

★ UNE NOUVELLE PREUVE DU 1ER THM D'INCOMPLÉTUDE DE GÖDEL BASÉE SUR LE PARADOXE DE BERRY : « LE PLUS PETIT ENTIER POSITIF QUI N'EST PAS DÉFINISSABLE EN MOINS DE SEIZE MOTS ».

THM : POUR TOUT SYSTÈME FORMEL F PERMETTANT D'EXPRIMER LA COMPLEXITÉ DE KOLMOGOROV, SI F EST CORRECT ALORS $\exists L \in \mathbb{N} : \forall x \in \{0,1\}^* :$ L'ÉNONCÉ « $K(x) > L$ » N'EST PAS PROUVABLE. (ET PARFOIS C'EST VRAI ⇒ INCOMPLET).

PREUVE (CHAITIN) : PAR L'ABSURDE, SI POUR TOUT L IL EXISTE UNE PREUVE DE « $K(x) > L$ » ALORS ON PEUT CONSTRUIRE UN PROGRAMME M QUI PREND L EN BINAIRE EN ENTRÉE ET DONNE EN SORTIE UN TEL x. ALORS $|\langle M \rangle|$ EST CONSTANT ET $|L| = \log_2(L)$, MAIS PUISQUE $(\langle \Pi \rangle, L)$ DÉCRIT x, ON A $K(x) \leq c + \log_2(L)$. POUR L SUFFISAMMENT GRAND CELA CONTREDIT « $K(x) > L$ » : F PROUVE UN ÉNONCÉ FAUX. ↯ □

REMARQUE : PAS D'AUTO-REF ! CE PROGRAMME ÉNUMÈRE LES PREUVES DE F JUSQU'À RENCONTRER UNE PREUVE DE LA FORME « $K(x) > L$ ».

THM : $K : \{0,1\}^* \to \mathbb{N}$ EST APPROXIMABLE "PAR AU-DESSUS". $\exists (K_i : \{0,1\}^* \to \mathbb{N})_{i \in \mathbb{N}}$ TELLES QUE $\forall x : \lim_{i \to +\infty} K_i(x) = K(x)$ ET $\forall i : K_i(x) \geq K(x)$.

PREUVE : SOIT c TELLE QUE $\forall x : K(x) \leq |x| + c$.
K_i EXÉCUTE TOUS LES PROGRAMMES DE TAILLE $< |x| + c$ POUR i ÉTAPES DE TEMPS, ET DONNE EN SORTIE LA TAILLE DU PLUS COURT QUI A DONNÉ x EN SORTIE (SI AUCUN, $|x| + c$). □

THM : $\exists c : \forall x \in \{0,1\}^* : K(d(x)) \geq |d(x)| - c$. LES DESC. MIN. SONT c-INCOMPRESSIBLES.

PREUVE : SOIT LA M.T. Π QUI, SUR L'ENTRÉE $(\langle R \rangle, y)$, CALCULE $R(y) = (\langle S \rangle, z)$ (SINON REJETTE) PUIS $S(z)$ ET LAISSE CE DERNIER RÉSULTAT SUR LE RUBAN.
SOIT $c = |\langle \Pi \rangle| + 1$. PAR L'ABSURDE, SI $d(x)$ EST c-COMPRESSIBLE ALORS $|d(d(x))| \leq |d(x)| - c$
MAIS $\langle \Pi \rangle d(d(x))$ DÉCRIT x AVEC LONGUEUR $|\langle \Pi \rangle| + |d(d(x))| \leq (c-1) + (|d(x)| - b) = |d(x)| - 1$
CE QUI CONTREDIT LA MINIMALITÉ DE $d(x)$. ↯. □

$= \min\{|\langle \Pi \rangle| \mid M(y) = x\}$.

NOTE : COMPLEXITÉ DE KOLMOGOROV CONDITIONNELLE $K(x|y)$ POUR CALCULER x ÉTANT DONNÉ y, ET INFORMATION MUTUELLE $I(y:x) = K(x) - K(x|y) \underset{\text{À CONSTANTE PRÈS}}{=} K(x) + K(y) - K((x,y))$.
"IN y ABOUT x"

KOLMOGOROV COMPLEXITY (4)

MAISON DES MATHEMATIQUES ET DE L'INFORMATIQUE

APPLICATION À LA DENSITÉ DES NOMBRES PREMIERS

IDÉE : LA DÉCOMPOSITION EN FACTEURS PREMIERS PERMET DE REPRÉSENTER SUCCINCTEMENT CERTAINS ENTIERS, MAIS LES INCOMPRESSIBLES RESTENT INCOMPRESSIBLES.
(= MOTS BINAIRES)

DEF : SOIT $\pi(n)$ LE NOMBRE D'ENTIERS PREMIERS $\leq n$. $\pi(11) = 5$.

THM : $\pi(n) \geq \frac{\log n}{\log\log n} - o(1)$ AVEC $o(1)$ UNE FONCTION QUI TEND VERS 0 QUAND $n \to +\infty$.

SOIT $n \in \mathbb{N}$ AVEC $n = p_1^{e_1} \cdot p_2^{e_2} \cdots p_m^{e_m}$ SA DÉCOMPOSITION EN FACTEURS PREMIERS.

ALORS ON PEUT REPRÉSENTER n AVEC $(\langle M \rangle, (e_1, \dots, e_m))$ où :

- $e_i \leq \log n$ DONC $\log\log n$ BITS DONC $(e_1, \dots, e_m)$ PEUT ÊTRE ENCODÉ PAR $0^{\log\log\log n}1 e_1 e_2 \dots e_m$ ← POUR DÉCODER LA SUITE.
- M EST UNE MACHINE CONSTANTE QUI DÉCODE $e_1, \dots, e_m$, CALCULE $p_1, \dots, p_m$ PUIS n.

DONC $K(n) \leq c + m \cdot \log\log n + \log\log\log n$

POUR n INCOMPRESSIBLE ON A $K(n) \geq \log n$, D'OÙ $\log n \leq c + m \cdot \log\log n + \log\log\log n$

$$\Rightarrow \frac{\log n - c - \log\log\log n}{\log\log n} \leq m$$

$$\Rightarrow \frac{\log n}{\log\log n} - o(1) \leq m \leq \pi(n). \quad \square$$

NB : HADAMARD ET DE LA VALLÉE POUSSIN : $\pi(n) \sim \frac{n}{\ln n}$ (1896).

NB : CONJECTURE DE GOLDBACH : TOUT $n > 2$ PAIR EST LA SOMME DE DEUX PREMIERS. (HALT)

NB : CONJECTURE DES PREMIERS JUMEAUX : $\exists$ UNE ∞ DE PREMIER p TEL QUE $p+2$ AUSSI PREMIER. (HALT-HALT).

NB : COLLATZ × HALT-HALT.

APPLICATION À LA COMPRESSION DE FICHIER

ON NE PEUT PAS FAIRE MIEUX QUE $K(x)$ POUR UN FICHIER $x \in \{0,1\}^*$.

APPLICATION À L'ALÉATOIRE

$\nexists$ ALGO POUR DÉCIDER SI UN MOT EST ALÉATOIRE ($K(x) \not\sim |x|$).