Short lecture notes[1] on Kolmogorov complexity and its applications for the course *Théorie des langages et pavages*, 1st semester of 2015–2016 (joint course of Victor Poupet and Andrei Romashchenko)

November 16th.

1 Kolmogorov complexity: the definition and basic properties

Definition 1.1. *Let $f : \{0,1\}^* \to \{0,1\}^*$ be a (partial) computable function. Then*

$$K_f(x) := \begin{cases} \min\{|p| \ : \ f(p) = x\}, \\ \infty, \quad \text{if there is no such } p. \end{cases}$$

Theorem 1.1. *There exists a partial computable function f (an* optimal decompressor*) such that for any other g there exists a constant C such that for all strings x*

$$K_f(x) \le K_g(x) + C.$$

Remark: There exist infinitely many optimal decompressors. If f_1 and f_2 are both optimal, then there exists a constant C such that for all binary strings x

$$|K_{f_1}(x) - K_{f_2}(x)| \le C.$$

We fix once and for all some optimal decompressor f_0. In what follows we omit the subscript and denote $K(x) := K_{f_0}(x)$. The value of $K(x)$ is called *Kolmogorov compelxity* of x.

Some simple properties of Kolmogorov complexity:

- $K(x) \le |x| + O(1)$;
- $K(xx) \le |x| + O(1)$;
- $K(f(x)) \le K(x) + O(1)$ for every computable f;
- for every integer $n > 0$ there exists a binary string x of length n such that $K(x) \ge n$ (an *incompressible* string);
- there exists a constant C such that for every integer $n > 0$ at least 99% of strings x of length n

$$n - C \le K(x) \le n + C;$$

- if a set $A \subset \{0,1\}^*$ is enumerable and $x \in A \cap \{0,1\}^n$, then $K(x) \le \log |A \cap \{0,1\}^n| + O(\log n)$.

Theorem 1.2. *There exists no algorithm $\mathfrak{A} : n \mapsto x_n$ that returns for every input n a string x_n such that $K(x_n) > n$.*

Corollary 1. The function $x \mapsto K(x)$ is not computable.

Corollary 2. If f_0 is a universal decompressor, then f_0 is not a total function.

Theorem 1.3 (a version of the Gödel incompleteness theorem, optional)**.** *There exists a constant C such that* ZFC *cannot prove any statement of the form "$K(x) > C$."*

Exercises 1.1. *Does there exist a universal decompressor f such that for every x the value of $K_f(x)$ is*
(a) an even number*?*
(b) a power of 2*?*
(c) a prime number?

Exercises 1.2. *If x a palindrome, then $K(x) \le |x|/2 + O(1)$.*

[1] Last update: December 16, 2015

November 23th.

Theorem 1.4. *(a) Let x be a string of length n, with pn ones and $(1-p)n$ zeros. Then*

$$K(x) \leq \left(p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} \right) n + O(\log n).$$

(b) Let n, m_0, m_1 be positive integers such that $n = m_0 + m_1$. Then there exists a binary string x of length n with m_0 zeros and m_1 ones such that

$$K(x) = \left(p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} \right) n + O(\log n),$$

where $p := m_1/n$.

Exercises 1.3. *(a) Prove that $N! = \tau(N) \cdot \left(\frac{N}{e}\right)^N$, where $\tau(N)$ is a function such that*

$$1 \leq \tau(N) \leq \mathrm{poly}(N)$$

for some polynomial $\mathrm{poly}(N)$.

(b) Prove that $N! = \lambda(N) \cdot \sqrt{N} \left(\frac{N}{e}\right)^N$, where $\lambda(N)$ is a function such that

$$c_1 \leq \lambda(n) \leq c_2$$

for some constants c_1, c_2.

Hint: *Estimate the difference between* $\ln(N!) = \sum\limits_{k=1}^{N} \ln k$ *and* $\int\limits_1^N \ln x \, dx = (x \ln x - x)\Big|_1^N$.

2 Conditional Kolmogorov complexity

Definition 2.1. *Let $f : \{0,1\}^* \times \{0,1\}^* \to \{0,1\}^*$ be a (partial) computable function. Then*

$$K_f(x \mid y) := \begin{cases} \min\{|p| \ : \ f(p,y) = x\}, \\ \infty, \quad \textit{if there is no such } p. \end{cases}$$

Theorem 2.1. *There exists a partial computable function of two arguments f_0 (an* optimal decompressor*) such that for any other computable f there exists a constant C such that for all strings x and y*

$$K_{f_0}(x \mid y) \leq K_f(x \mid y) + C.$$

We fix once and for all some optimal decompressor f_0. In what follows we omit the subscript and denote $K(x \mid y) := K_{f_0}(x \mid y)$. The value of $K(x \mid y)$ is called *conditional Kolmogorov compelxity* of x given y.

Some properties of conditional Kolmogorov complexity:

- $K(x \mid y) \leq K(x) + O(1)$;
- $K(x \mid \Lambda) = K(x) + O(1)$;
- $K(x \mid x = O(1)$;
- $K(f(x) \mid x) = O(1)$ for every computable function f;
- for every n and every y there exists a binary string x of length n such that $K(x) \geq n$;
- there exists a constant C such that for all y and for all $n > 0$, for at least 99% of x of length n

$$n - C \leq K(x \mid y) \leq n + C.$$

We fix some *computable encoding* of pairs of binary strings $\langle x, y\rangle$ and define $K(x, y)$ as $K(\langle x, y\rangle)$.

Proposition 2.1. $K(x, y) \le K(x) + K(y \mid x) + O(\log K(x))$.

Proposition 2.2. *For every number $C > 0$ there exist binary strings x and y such that*

$$K(x, y) > K(x) + K(y \mid) + C.$$

Exercises 2.1. *Prove that*
(a) $K(x, y) \le K(x) + K(y \mid x) + \log K(x) + \log\log K(x) + \log\log\log K(x) + 2\log\log\log\log K(x) + O(1)$,
(b) $K(x, y) \le K(x) + K(y \mid x) + \log K(x, y) + O(1)$.

Remark: *observe that (a) is* **not** *a corollary of (b).*

Theorem 2.2 (Kolmogorov–Levin)**.** $K(x, y) = K(x) + K(y \mid x) + O(\log K(x, y))$.

Definition 2.2. *The quantity of* information *in x on y is defined as* $I(x : y) := K(y) - K(y \mid x)$.

Exercises 2.2. *Prove that*
(a) $|I(x : y) - I(y : x)| = O(\log K(x, y))$,
(b) $I(x : y) = K(x) + K(y) - K(x, y) + O(K(x, y))$.

Exercises 2.3. *Prove that*
(a) $2K(x, y, z) \le K(x, y) + K(x, z) + K(y, z) + O(\log K(x, y, z))$,
(b) $K(z) \le K(z \mid x) + K(z \mid y) + I(x : y) + O(\log K(x, y, z))$,
(c) $K(x, y, z) + K(z) \le K(x, z) + K(y, z) + O(\log K(x, y, z))$.

November 30th.

Proposition 2.3. *There exist x and y such that*

$$I(x : y) - I(y : x) = \log K(x, y) + O(1)$$

(i.e., the mutual information is symmetric only with the logarithmic precision).

Definition 2.3. *The quantity of* information *in x on y given z is defined as $I(x : y \mid z) := K(y \mid z) - K(y \mid x, z)$.*

Exercises 2.4. *Prove that*
(a) $I(x : y \mid z) = I(y : x \mid z) + O(\log K(x, y, z))$,
(b) $I(x : y \mid z) = K(x, z) + K(y, z) - K(x, y, z) - K(z) + O(\log K(x, y, z))$.

3 Applications of Kolmogorov complexity

Example 3.1. *Let $A \subset \mathbb{N}^3$ be a finite set; denote $\pi_{12}(A)$, $\pi_{13}(A)$, and $\pi_{23}(A)$ its projections on the "coordinate planes" (e.g., $\pi_{12}(A) = \{(x, y) \ : \ \exists z (x, y, z) \in A\}$). Then*

$$|A|^2 \leq |\pi_{12}(A)| \cdot |\pi_{13}(A)| \cdot |\pi_{23}(A)|.$$

Idea of the proof: we take a non-compressible triple (x, y, z) in A^N and apply the inequality $2K(x, y, z) \leq K(x, y) + K(x, z) + K(y, z) + O(\log K(x, y, z))$; then tend N to infinity to make the term $O(\log K(x, y, z))$ negligible.

Exercises 3.1. *Prove the "continuous" version of Example 3.1: let $A \subset \mathbb{R}^3$ be a finite set; then*

$$\mathit{volume}(A)^2 \leq \mathit{area}(\pi_{12}(A)) \cdot \mathit{area}(\pi_{13}(A)) \cdot \mathit{area}(\pi_{23}(A)).$$

Example 3.2. *There exist infinitely many prime numbers.*

Idea of the proof (using Kolmogorov complexity): Assume that there exist only t prime numbers $p_1 = 2, p_2 = 3, \ldots, p_t$. Than every integer $N > 0$ can be represented as

$$N = p_1^{d_1} \cdot p_2^{d_2} \cdot \ldots \cdot p_t^{d_t}.$$

Hence,

$$K(N) \leq K(d_1, \ldots, d_t) + O(1) \leq O(t \log \log N).$$

The same time, if the binary representation of N is an incompressible word, then $K(N) \geq \log N$; this gets a contradiction for large enough N.

Example 3.3. *The language $\{0^n 1^n \ : \ n \in \mathbb{N}\}$ is not* rational *(no finite automata can recognize this language).*

Idea of the proof: Assume $\mathfrak{A}$ is an automaton that recognize this language. Let it comes to the state q_i after reading the input 1^n. Observe that given $\mathfrak{A}$ and the state q_i we can reconstruct n. It follows that $K(n) \leq K(\mathfrak{A}, i)$. We get a contradiction for n with high enough Kolmogorov complexity.

Exercises 3.2. *Prove that the language of* palindromes *is not rational.*

Example 3.4. *For every non-deterministic finite automaton with n states there exists an equivalent deterministic finite automaton with at most 2^n states. The same time, it is impossible to covert all non-deterministic finite automata with n states in equivalent deterministic finite automata with $o(2^n)$ states.*

Idea of the proof: Take the language $L^{(n)}$ over the binary alphabet that consists of all x whose nth bits from the right is equal to 1. This language can be recognized by a nondeterministic automaton with $(n + 1)$ states; on the other hand, every deterministic automaton recognizing $L^{(n)}$ must consit $\Omega(2^n)$ states. Indeed, we can reconstruct every n-bits word x given an automaton $\mathfrak{A}$ that recognizes $L^{(n)}$ and the state q_i of $\mathfrak{A}$ at which we arrive after reading x. We choose x so that $K(X \mid \mathfrak{A}) \geq n$; then we conclude that the index of a state in $\mathfrak{A}$ must contain at least $n - O(1)$ bits. Hence, the number of states in this automaton is $\Omega(2^n)$.

December 7th.

We defined the multi-head finite automaton (a generalization of standard deterministic finite automaton). Then we proved the following theorem.

Theorem 3.1. *For every integer $n > 0$ there exists a language L that can be recognized by a multi-head finite automaton with $n+1$ heads but cannot be recognized by any automaton with n heads.*

Idea of the proof: Consider the language over the alphabet $\{0, 1, \#\}$

$$L_k = \{w_1\#w_2\#\ldots\#w_{k-1}\#w_k\#w_k\#w_{k-1}\#\ldots\#w_2\#w_1 \mid w_i \in \{0,1\}^* \text{ for all } i\}.$$

It can be recognized by an automaton with n heads, if and only if $\frac{n(n-1)}{2} \geq k$.

4 Martin-Löf randomness

Proposition 4.1. *For every infinite binary sequence $\omega_1\omega_2\ldots\omega_n\ldots$ and for every $C > 0$ there exists n such that*

$$K(\omega_1\omega_2\ldots\omega_n) < n - C.$$

Remark. We can prove even a stronger statement: for every infinite binary sequence $\omega_1\omega_2\ldots\omega_n\ldots$ there exists a constant $D > 0$ such that for infinitely many n

$$K(\omega_1\omega_2\ldots\omega_n) < n - \log n + D.$$

Definition 4.1. *Let $f : \{0,1\}^* \to \{0,1\}^*$ be a (partial) computable function with a prefix free domain (if p is a prefix of q, then $f(p)$ and $f(q)$ cannot be defined simultaneously). Then*

$$KP_f(x) := \begin{cases} \min\{|p| \ : \ f(p) = x\}, \\ \infty, \quad \textit{if there is no such } p \end{cases}$$

is called a prefix Kolmogorov complexity *of x with respect to f.*

Theorem 4.1. *There exists a partial computable function f with a prefix free domain (an* optimal prefix-free decompressor*) such that for any other computable g with a prefix-free domain there exists a constant C such that for all strings x*

$$KP_f(x) \leq KP_g(x) + C.$$

We fix *some* optimal prefix-free decompressor f; in what follows we denote it $KP(x)$ (omitting the subscription f). The value $KP(x)$ is called *prefix (Kolmogorov) complexity* of x.

Some simple properties of prefix Kolmogorov complexity:

- $KP(x) \leq |x| + 2\log|x| + O(1)$;
- $KP(x) \leq K(x) + 2\log K(x) + O(1)$;
- $KP(x, y) \leq KP(x) + KP(y) + O(1)$;
- $\sum\limits_{x\in\{0,1\}^*} 2^{-KP(x)} \leq 1$.

Exercises 4.1. $KP(x) \leq K(x) + \log K(x) + 2\log\log K(x) + O(1)$.

Definition 4.2. *An infinite binary sequence $X = \omega_1\omega_2\ldots\omega_n\ldots$ is called* Martin-Löf random*, if there exists a constant $C > 0$ such that for all $n > 0$*

$$KP(\omega_1\omega_2\ldots\omega_n) > n - C.$$

Theorem 4.2. *Almost all (with respect to the uniform Bernoulli measure) infinite binary sequence $X = \omega_1\omega_2\ldots\omega_n\ldots$ are Martin-Löf random.*

Proposition 4.2. *A Martin-Löf random sequence cannot be computable.*

Theorem 4.3. *For all Martin-Löf random binary sequence $X = \omega_1\omega_2\ldots\omega_n\ldots$*

$$\lim_{n\to\infty}\frac{\omega_1+\omega_2+\ldots+\omega_n}{n} = \frac{1}{2}$$

(*the limit of frequencies exists and is equal to* $1/2$).

Theorem 4.4. *For all Martin-Löf random binary sequence $X = \omega_1\omega_2\ldots\omega_n\ldots$*

$$\frac{\omega_1+\omega_2+\ldots+\omega_n}{n} - \frac{1}{2} = O\left(\sqrt{\frac{\log n}{n}}\right).$$

Corollary 4.1 (the law of large numbers in the form of Hardy and Littlewood)**.** *For almost all (with respect to the uniform Bernoulli measure) infinite binary sequence $X = \omega_1\omega_2\ldots\omega_n\ldots$*

$$\frac{\omega_1+\omega_2+\ldots+\omega_n}{n} - \frac{1}{2} = O\left(\sqrt{\frac{\log n}{n}}\right).$$

Remark: It can be proven that for almost all (with respect to the uniform Bernoulli measure) infinite binary sequence $X = \omega_1\omega_2\ldots\omega_n\ldots$, for every $\varepsilon > 0$ and all large enough n

$$\left|\frac{\omega_1+\omega_2+\ldots+\omega_n}{n} - \frac{1}{2}\right| \le (1+\varepsilon)\sqrt{\frac{2\ln\ln n}{n}}$$

(the law of iterated logarithm). But we do not prove this bound in our course.

Exercises 4.2. *Prove that a sequence $X = \omega_1 0\omega_2 0\omega_3 0\ldots\omega_n 0\ldots$ cannot be Martin-Löf random.*

Exercises 4.3. *Prove that a sequence $X = \omega_1\omega_2\ldots\omega_n\ldots$ cannot be Martin-Löf random.*

December 14th.

5 Infinite sequences without simple factors

Proposition 5.1. *For every k, and for every Martin-Löf random sequence of bits $X = \omega_1\omega_2\ldots\omega_n\ldots$ there exists n such that*

$$\omega_{n+1} = \omega_{n+2} = \omega_{n+3} = \ldots = \omega_{n+k} = 0$$

(i.e., in every Martin-Löf random sequence there is a subword $\underbrace{00\ldots0}_{k}$).

Theorem 5.1. *(a) For every $\alpha < 1$ there exists a sequence of bits $X = \omega_1\omega_2\ldots\omega_n\ldots$ and a constant C such that for every n*

$$K(\omega_{n+1}\omega_{n+2}\omega_{n+3}\ldots\omega_{n+k}) > \alpha k - C.$$

(b) For every $\alpha < 1$ there exists a sequence of bits $X = \omega_1\omega_2\ldots\omega_n\ldots$ and a constant C such that for every n

$$K(\omega_{n+1}\omega_{n+2}\omega_{n+3}\ldots\omega_{n+k}\,|\,\omega_1\omega_2\ldots\omega_n) > \alpha k - C.$$

Remark: In some applications it easier to use the following corollary from Theorem 5.1: (a) For every $\alpha < 1$ there exists a sequence of bits $X = \omega_1\omega_2\ldots\omega_n\ldots$ such that for all *large enough* n

$$K(\omega_{n+1}\omega_{n+2}\omega_{n+3}\ldots\omega_{n+k}) > \alpha k.$$

(b) For every $\alpha < 1$ there exists a sequence of bits $X = \omega_1\omega_2\ldots\omega_n\ldots$ such that for all *large enough* n

$$K(\omega_{n+1}\omega_{n+2}\omega_{n+3}\ldots\omega_{n+k}\,|\,\omega_1\omega_2\ldots\omega_n) > \alpha k.$$

Definition 5.1 (rational powers of words)**.** *Let $z = \omega_1\omega_2\ldots\omega_n$ be a word of length n over some (finite) alphabet. Then we denote by z^r the word*

$$\underbrace{\underbrace{\omega_1\omega_2\ldots\omega_n}_{n}\underbrace{\omega_1\omega_2\ldots\omega_n}_{n}\ldots\underbrace{\omega_1\omega_2\ldots\omega_n}_{n}\underbrace{\omega_1\omega_2\ldots\omega_k}_{k\le n}}_{r\cdot|z|=m\cdot n+k},$$

which is a concatenation of several copies of z and (possibly) of some prefix of z, so that the total length of the result is equal to $r\cdot|z|$. In particular, the integer *powers of z are defined as $z^2 = zz$, $z^3 = zzz$, etc.*

Theorem 5.2. *For every rational $p/q > 1$ there exists a (finite) alphabet and an infinite sequence of letters from this alphabet $\omega_1\omega_2\ldots\omega_n\ldots$ that contains no factors of the form $z^{p/q}$.*

Theorem 5.3 (without proof)**.** *For every $\gamma > 1$ there exists an infinite sequence of letters over some finite alphabet $X = \omega_1\omega_2\ldots\omega_n\ldots$ such that*
(i) for every rational $p/q > \gamma$ there exists no factors of the form $z^{p/q}$ in X, and
(ii) for every positive rational $p/q < \gamma$ there exists some factors of the form $z^{p/q}$ in X.

Theorem 5.4. *There exists a 2-dimensional binary configuration*

$$\tau \;:\; \mathbb{Z}^2 \to \{0,1\}$$

where each $n\times n$ pattern has Kolmogorov complexity $\Omega(n^2)$.

Proposition 5.2. *The set of all binary configurations $\tau \;:\; \mathbb{Z}^2 \to \{0,1\}$ such that every $n\times n$ pattern in τ has Kolmogorov complexity $\Omega(n^2)$ is a (non-empty) effectively closed subshift.*

Theorem 5.5 (not necessary for the final exam)**.** *For every non-empty subshift of finite type (in particular, for every tile set that allows tilings of the plane) there exists some valid configuration*

$$\tau \ : \ \mathbb{Z}^2 \to A$$

where each $n \times n$ pattern has Kolmogorov complexity $O(n)$.

Exercises 5.1. *Let $\omega_1\omega_2\ldots\omega_n$ be a binary word of length n (where n is an even integer), and*

$$\omega_1 + \omega_2 + \ldots + \omega_n = n/2.$$

Prove that $K(\omega_1\omega_2\ldots\omega_n \,|\, n) < n - \frac{1}{2}\log n + O(1)$.

Exercises 5.2. *Let $\omega_1\omega_2\ldots\omega_n$ be a binary word of length n such that $K(\omega_1\omega_2\ldots\omega_n \,|\, n) > n$. Prove that*

$$\left|\omega_1 + \omega_2 + \ldots + \omega_n - \frac{n}{2}\right| = \Omega(\sqrt{n})$$

Exercises 5.3. *Let $X = \omega_1\omega_2\ldots\omega_n\ldots$ be a Martin-Löf random sequence of bits. Prove that for all large enough n the prefix of this sequence $\omega_1\omega_2\ldots\omega_n$ contains at least one factor $\underbrace{00\ldots0}_{\sqrt{\log n}}$ (i.e., at least $\sqrt{\log n}$ zeros in a row).*

Exercises 5.4. *Denote $i_{-1}i_0{,}i_1i_2i_3\ldots$ the binary expansion of the number π (i.e., π in base 2). Prove that all sequences of the form*

$$\omega_1 i_1 \omega_2 i_2 \ldots \omega_n i_n \ldots$$

(for arbitrary $\omega_n \in \{0,1\}$) are not Martin-Löf random.

References

[1] Ming Li and Paul Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, 3rd Edition, 2008.

[2] Bruno Durand, Alexandr Zvonkin, *Complexité de Kolmogorov.*
https://www.labri.fr/perso/zvonkin/Research/kolmogorov.pdf

[3] Laurent Bienvenu, Mathieu Hoyrup, *Une brève introduction á la théorie effective de l'aléatoire.*
http://www.liafa.jussieu.fr/~lbienven/docs/publications/Une_breve_introduction_a_la_theorie_effective_de_l_aleatoire-SMF.pdf

[4] Alexander Shen, *Algorithmic Information Theory and Kolmogorov Complexity.*
http://www.lirmm.fr/~ashen/uppsala-notes.pdf

[5] Lance Fortnow, *Kolmogorov complexity.*
http://people.cs.uchicago.edu/~fortnow/papers/kaikoura.pdf

[6] A. Shen, V.A. Uspensky, N.K. Vereshchagin, *Kolmogorov complexity and algorithmic randomness*, 2014.
http://www.lirmm.fr/~ashen/kolmbook-eng.pdf