Egzamin z Teorii informacji. Rozwiązanie zadania o złożoności Kołmogorowa.

10 lutego 2017

Zadanie

Dowieść, że dla prawie wszystkich słów binarnych losowych w sensie Kołmogorowa (tzn. $C_U(w) \geq |w|$), liczba jedynek w słowie o długości n wynosi

(a) co najmniej $\sqrt{n}$,

(b) co najmniej $\frac{1}{4}n$.

Rozwiązanie. W obu przypadkach będziemy rozumować nie wprost. Opiszemy sposoby reprezentacji słowa x, które w przypadku, gdy liczba jedynek jest „mała", prowadzą do kompresji i zaprzeczają losowości. Oczywiście punkt (a) wynika z punktu (b), ale przedstawimy osobno jego prostsze rozwiązanie. Liczbę jedynek w słowie x oznaczamy przez $\sharp_1(x)$.

Ad (a). Niech x będzie słowem o długości $n \geq 2$, $x = x_1 \dots x_n$. Dla $1 \leq i \leq n$, określamy standardową reprezentację binarną $bin(i)$ liczby i jako słowo długości $\lfloor \log n \rfloor + 1$ (w razie potrzeby dopełnione początkowymi zerami). Natomiast $bin'(i)$ oznacza „bezprefiksową" reprezentację i (znaną z ćwiczeń), np. w ciąg $bin(i)$ wkładamy co drugą cyfrę 0, a ostatnią 1. Przypuśćmy, że 1 występuje w słowie x dokładnie na pozycjach $i_1, i_2, \dots, i_k$. Niech

$$w_x \quad = \quad bin'(n)bin(i_1)\dots bin(i_k)$$

(zauważmy, że w_x jest określone również, gdy 1 w ogóle nie występuje). Oczywiście odwzorowanie $x \mapsto w_x$ jest różnowartościowe i

$$|w_x| \quad = \quad (\lfloor \log n \rfloor + 1) \cdot (2 + \sharp_1(x)). \tag{1}$$

Niech T będzie maszyną Turinga, która dla danego $w = w_x$ odtwarza słowo x; jeśli w nie jest postaci w_x dla żadnego x, maszyna T się zapętla. Gdy $\sharp_1(x) < \sqrt{n}$, z równości (1) otrzymujemy

$$C_T(x) \quad \leq \quad (\log n + 1) \cdot (2 + \sqrt{n}).$$

Z zasady niezmienniczości[1] mamy, że wówczas

$$C_U(x) \quad \leq \quad (\log n + 1) \cdot (2 + \sqrt{n}) + c, \tag{2}$$

dla pewnej stałej c. Dla słów losowych mamy więc nierówność $n \leq (\log n + 1) \cdot (2 + \sqrt{n}) + c$, która jednak może zachodzić tylko dla skończenie wielu n.

Ad (b). Metoda z poprzedniego punktu nie zadziałałaby w tym przypadku, bo $\frac{1}{4}n \cdot \log n$ jest asymptotycznie większe niż n. Zamiast tego, dla każdego $n \geq 1$, rozważmy M_n zbiór słów binarnych długości n, w których liczba jedynek jest mniejsza niż $\frac{1}{4}n$. Niech $m_n = |M_n|$. Słowa w M_n możemy uporządkować leksykograficznie w ciąg $y_1, y_2, \dots, y_{m_n}$. Słowo x długości n, w którym występuje mniej niż $\frac{1}{4}n$ jedynek jest równe pewnemu słowu y_j w powyższym ciągu. Określamy wówczas

$$v_x \quad = \quad bin'(n)bin(j)$$

gdzie $bin'(n)$ jest jak w poprzednim punkcie, a za $bin(j)$ przyjmujemy zwykłe rozwinięcie binarne liczby j (nie dodajemy początkowych zer). Jak poprzednio, istnieje maszyna Turinga, powiedzmy S, która dla danego v_x odtwarza x. Żeby oszacować $C_S(x)$ — i w konsekwencji $C_U(x)$, potrzebujemy oszacować maksymalną wartość j, czyli m_n. Zauważmy, że $m_n = \sum_{i < \frac{1}{4} \cdot n} \binom{n}{i}$, a tę właśnie wielkość oszacowaliśmy dla dowolnego $0 < \lambda \leq \frac{1}{2}$ w dowodzie Drugiego Twierdzenia Shannona, dokładniej[2]

$$\sum_{i \leq \lambda \cdot n} \binom{n}{i} \quad \leq \quad 2^{n \cdot H(\lambda)},$$

[1] Zob. np. Notatki https://www.mimuw.edu.pl/~niwinski/Info/2016-info.pdf, *Fact 8*, str. 46.

[2] Zob. Notatki *Lemma 7*.

gdzie $H(\lambda) = -\lambda \log \lambda - (1-\lambda)\log(1-\lambda)$. Jak pamiętamy, funkcja $H(\lambda)$ przyjmuje w przedziale $[0,1]$ maksymalną wartość 1 jedynie dla $\lambda = \frac{1}{2}$. Mamy więc $|bin(j)| \leq n \cdot H\left(\frac{1}{4}\right) + 1$, gdzie $H\left(\frac{1}{4}\right) < 1$. Korzystając z zasady niezmienniczości

$$C_U(x) \quad \leq \quad 2\log n + nH\left(\frac{1}{4}\right) + c, \tag{3}$$

dla pewnej stałej c. Dla słów losowych mamy więc nierówność $n \leq 2\log n + nH(\frac{1}{4}) + c$, która jednak może zachodzić tylko dla skończenie wielu n, bo oczywiście $n(1 - H(\frac{1}{4})) - 2\log n \to \infty$.

Uwaga. Z dowodu wynika, że w punkcie (b) moglibyśmy zastąpić $\frac{1}{4}$ przez dowolną stałą $< \frac{1}{2}$. Szacowanie można też otrzymać w inny sposób, używając Wzoru Stirlinga.