Zadanie 2 z listy 4 - "Kompresja Danych"

Łukasz Klasiński

19 kwietnia 2020

Zadanie 2

Podaj przykład prawdopodobieństw oraz ciągów danych, które uniemożliwiają wykonanie przeskalowań E_1, E_2 i E_3 na długich fragmentach ciągu wejściowego. Jak wpłynie to na wymaganą długość binarnych reprezentacji końców przedziału $[l, p)$? W swoim rozwiązaniu możesz przyjąć, że każdą z liczb $F(1), \ldots, F(n+1)$ można zapisać na k bitach.

Rozwiązanie

Prostym przykładem takich danych są następujące prawdopodobieństwa:

$$p_1 = \frac{1}{n}$$

$$p_2 = \frac{n-2}{n}$$

$$p_3 = \frac{1}{n}$$

Oraz wyraz składający się wyłącznie ze znaków a_2. Po pierwszym znaku nasze $[l, p)$ będzie mało wartości $[\frac{1}{n}, \frac{n-1}{n})$. Widzimy zatem, że dla $n > 4$ żadna z operacji E_1, E_2, E_3 nie nie zostanie wykonana. Kolejne przedziały liczymy już rekurencyjne:

$$[l_{old} + F(2)(p_{old} - l_{old}), l_{old} + F(3)(p_{old} - l_{old})$$

Załóżmy teraz, że $F(1) \ldots F(n+1)$ można zapisywać na k bitach. Widać, że p_{old} jest znacząco większe od l_{old}, zatem można przyjąć że $(p_{old} - l_{old}) \sim p_{old}$. Po wykonaniu tej operacji, przedziały zwiększają się o $k + m$ bitów, gdzie m to ilość bitów p_{old}.

Jako że dla odpowiednio dużego n, l_{old} jest małe (a przedziały z każdą iteracją zmieniają się $\sim$ ł$_{old}$), to l, p będą zbiegać bardzo powoli do przedziału $E_3 = [0.25, 0.75)$. Dodać zatem, że w zależności od n, przedziały mogą wymagać k^t bitów, gdzie t to ilość iteracji(znaków) po których $[l, p)$ będzie mogło być przesunięte operacją E_3. Przykładowo dla $n = 100$, kolejne wartości przedziału wyglądają następująco:

$$(0.01, 0.99)$$

$$(0.0197, 0.9802)$$

$$(0.038815, 0.961184)$$

$$(0.04803960, 0.951960)$$

$$\ldots$$

$$(0.23271268350260568, 0.7672873164973939)$$

Aż przedziały dojdą do $[0.25, 0.75)$ oraz wykona się przesunięcie E_3. Nastąpi cyklicznie takie samo wyliczanie przedziałów, ponieważ l i p znowu przyjmą wartości $\sim 0.25, 0.75$. Oczywiście można taką sytuację obejść poprzez odpowiednie zaokrąglanie l i p, ale realistycznie nie jest to potrzebne bo takie przypadki raczej nie zachodzą.