Teoria informacji i kodowania: Lista 3

Zadanie 1. Podaj algorytm sprawdzający, czy zbiór słów jest kodem jednoznacznie dekodowalnym.

Wsk.: Podejście używające automatów może być najprostsze, ale prawdopodobnie nie najwydajniejsze.

Zadanie 2. Wykaż, że nierówność Krafta–McMillana zachodzi również dla przeliczalnego ciągu długości $\ell_1, \ldots, \ell_n, \ldots$.

Zadanie 3 (* możliwe, że się nie da)**.** Spróbuj udowodnić nierówność Krafta–McMillana używając funkcji tworzących.

Zadanie 4. Pokaż, że kod Huffmana jest optymalny (wśród kodów prefiksowych), tj. ma najmniejszą oczekiwaną długość słowa kodowego. Przypadek binarny jest prostszy, niż ogólny.

Zapewne trzeba skorzystać z obserwacji podanej na wykładzie: jeśli prawdopodobieństwa słów kodowych spełniają $p_1 \geq p_2 \geq \cdots \geq p_k$ to długości słów kodowych w optymalnym kodzie spełniają $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_k$.

Zadanie 5. Konstrukcja kodów Huffmana, dla danych prawdopodobieństw $\{p(u)\}_{u \in U}$ minimalizuje średnią długość słowa kodowego

$$\sum_{u \in U} p(u)|C(u)| \ . \tag{1}$$

Jednak różne kody prefiksowe realizujące to optimum mogą mieć różne sumy długości kodu

$$\sum_{u \in U} |C(u)| \ . \tag{2}$$

Zmodyfikuj algorytm dla kodów Huffmana tak, aby zwracał on kod C, który osiąga minimalną średnią długość kodu (1) oraz (wśród takich kodów) minimalną sumę długości dłów kodowych (2). Dla uproszczenia można się skupić na kodach binarnych.

Zadanie 6 (Gra w 20 pytań)**.** Rozważmy grę, w której Alicja wybiera przedmiot $u \in U$ według pewnego rozkładu prawdopodobieństwa p, a Bob chce ustalić, który to przedmiot za pomocą pytań o podzbiory, tj. dla dowolnego $U' \subseteq U$ może zapytać „Czy $u \in U'$?"; Alicja zawsze udziela prawdziwej odpowiedzi. Bob chce znaleźć strategię, która minimalizuje średnią liczbę pytań; Bob zna rozkład prawdopodobieństwa p! Zredukuj ten problem do problemu optymalnego kodu prefiksowego.

Pytania mogą wydawać się nieco sztuczne, gdyż mogą pytać o dowolne podzbiory. Załóżmy, że elementy U są posortowane liniowo według p, tzn. mamy dane $u_1, \ldots, u_n$ takie, że $p(u_1) \geq p(u_2) \geq \cdots \geq p(u_n)$ i chcemy ustalić u_a wybrane przez Alicję, ale możemy zadawać tylko pytania postaci „Czy $a \geq j$?". Pokaż, że optymalne strategie dla obu wariantów mają tę samą oczekiwaną liczbę pytań (dla tego samego rozkładu prawdopodobieństwa p). Podaj strategię dla drugiego wariantu.

Wsk.: Oblicz optymalne długości za pomocą algorytmu Huffmana. Wykorzystaj strategię budowy kodu z dowodu nierówności Krafta–McMillana.

Zadanie 7 (Kodowanie Shannon-Elias-Fano)**.** Dla symboli $1, \ldots, n$ i prawdopodobieństw $p_1, \ldots, p_n$ zdefiniujmy $\mathrm{CDF}(i) = \sum_{j<i} p_i$. Niech też $\ell_i := \lceil -\log p_i \rceil + 1$. Kodem dla $C(i)$ jest ℓ_i pierwszych cyfr po przecinku (w rozwinięciu binarnym) liczby $\mathrm{CDF}(i) + \frac{p_i}{2}$ (czyli średnia między $\mathrm{CDF}(i), \mathrm{CDF}(i+1)$). Pokaż, że jest to kod prefiksowy.

Wsk.: Pokaż, że przedziały $[C(i), C(i) + \ell_i)$ są rozłączne.

Zadanie 8. Dla przypomnienia, kodowanie ANS wygląda następująco: każde p_i jest postaci $p_i = \frac{c_i}{C}$, gdzie $C = \sum_{i=1}^{k} c_i$; niech $C_i = \sum_{j<i} c_i$. Jeśli w zakodowaliśmy jako n, to wi kodujemy jako

$$\left\lfloor \frac{n}{c_i} \right\rfloor \cdot C + C_i + (n \bmod c_i) \ .$$

Jak należy zakodować ϵ, żeby to kodowania było jednoznacznie dekodowalne? Jakie liczby naturalne są w obrazie kodowania?

Zadanie 9 (* Nie wiem, czy istnieje dobre rozwiązanie)**.** Porównaj oczekiwany rozmiar wyjścia algorytmu ANS dla ciągu $X_1, \ldots, X_n$ zmiennych i.i.d. $\sim X$ względem entropii $nH[X]$. Możesz założyć, że ANS działa jak w poprzednim zadaniu (w szczególności, prawdopodobieństwa dane są przez ułamki o tym samym mianowniku) oraz uszeregować prawdopodobieństwa $p_1, \ldots, p_k$ (oryginalny algorytm ANS używa $p_1 \geq p_2 \geq \cdots \geq p_k$).