Кодирование и энтропия

Есть алфавит A из n букв $a_1, \ldots, a_k$. Мы хотим кодировать буквы этого алфавита двоичными словами. Формально говоря, *кодом* для алфавита A называется набор из k двоичных слов $c_1, \ldots, c_k$. Они называются *кодовыми словами* данного кода; слово c_i называется *кодом* буквы a_i. Всякое слово в алфавите A кодируется двоичным словом, получаемым соединением кодов соответствующих букв. Будем называть код *однозначно декодируемым*, если коды любых двух различных слов различны. (В частности, коды любых двух букв должны быть различны.) Код называется *префиксным*, если ни одно из кодовых слов не является началом (префиксом) другого.

Задача 1. (а) Докажите, что для любого префиксного кода выполнено *неравенство Крафта*: $\sum_i 2^{-l(c_i)} \leqslant 1$. (Здесь $l(u)$ обозначает длину слова u.)

(б) Докажите обратное: для любого набора натуральных чисел $l_1, \ldots, l_k$ с $\sum_i 2^{-l_i} \leqslant 1$ найдётся префиксный код $c_1, \ldots, c_k$ с $l(c_i) = l_i$.

(в*) Докажите, что неравенство Крафта выполнено и для любого однозначного кода (для однозначных кодов оно называется *неравенством Макмиллана*).

Задача 2. Докажите, что всякий префиксный код является однозначно декодируемым, но обратное неверно.

Код естественно выбирать так, чтобы использовать меньше битов: наиболее частые буквы разумно кодировать коротко. Пусть фиксированы неотрицательные числа $p_1, \ldots, p_k$, в сумме равные единице (вероятности букв $a_1, \ldots, a_k$). *Средней длиной кода* называется величина $\sum_i p_i l(c_i)$. Средняя длина, естественно, зависит не только от кода, но и от вероятностей букв.

Задача 3. Покажите, что задачу построения префиксного кода минимальной средней длины (для данных вероятностей букв $p_1, \ldots, p_k$) можно переформулировать так: найти числа $q_1, \ldots, q_k$, являющиеся степенями двойки, для которых $q_1 + \ldots + q_k \leqslant 1$ и сумма $\sum_i p_i \log(1/q_i)$ минимальна.

Сначала изучим "релаксацию" этой задачи: пусть q_i — произвольные неотрицательные числа, в сумме не большие 1, не обязательно степени двойки.

Задача 4. Покажите, что минимум указанной выше суммы достигается при $q_i = p_i$. [Можно считать, что $\sum q_i = 1$; условный экстремум можно искать дифференцированием. Можно воспользоваться также выпуклостью логарифма.]

Этот минимум, то есть величину $\sum p_i \log(1/p_i)$, называют *энтропией Шеннона* распределения вероятностей $p_1, \ldots, p_n$. Другими словами, энтропия Шеннона

$$H(p_1, \ldots, p_n) = -\sum_i p_i \log p_i,$$

при этом полагают $0 \ln 0 = 0$ по непрерывности.

Для произвольного распределения $q_1, \ldots q_k$ разницу между $p_i \log(1/q_i)$ и минимумом $p_i \log(1/p_i)$ называют *дивергенцией Кульбака–Лейблера* между распределениями вероятностей p_i и q_i. Её можно записать как

$$\sum_i p_i \log \frac{p_i}{q_i}.$$

Задача 5. Докажите, что дивергенция неотрицательна и обращается в нуль, лишь если $p_i = q_i$ при всех i.

Задача 6. Докажите, что энтропия Шеннона не меньше *минимальной энтропии*, определяемой как $H_{\min} = \min_i(-\log p_i)$.

Задача 7. (а) Найдите энтропию Шеннона *равномерного распределения*: $p_1 = \ldots = p_k = 1/k$. (б) Докажите, что энтропия любого неравномерного распределения на k исходах строго меньше энтропии равномерного распределения.

Задача 8. (а) Докажите, что средняя длина любого префиксного кода алфавита с вероятностями букв $p_1, \ldots, p_k$ не меньше энтропии распределения $p_1, \ldots, p_k$.
(б) Докажите, что обратное верно с точностью до добавления единицы к энтропии: для любого распределения вероятностей $p_1, \ldots, p_k$ найдётся префиксный код со средней длиной меньше $H(p_1, \ldots, p_k) + 1$.

Задача 9. Рассмотрим следующий вариант игры в 10 вопросов. Алиса задумала число от 1 до k, выбрав его случайно с некоторым распределением вероятностей $p_1, \ldots, p_k$. Боб, задавая вопросы с ответами "да/нет", должен узнать задуманное число. Докажите, что любой алгоритм Боба в среднем задаёт не менее $H(p_1, \ldots, p_k)$ вопросов.

Кодом Хаффмана для данного набора вероятностей $p_1, \ldots, p_k$ называется код, построенный следующим рекурсивным алгоритмом:

- Если $k = 1$, то выдать пустое слово в качестве единственного кодового слова.
- При $k > 1$ заменить две самые маленькие вероятности на их сумму, затем запустить алгоритм на полученном распределении (на $(k-1)$-буквенном алфавите), а в полученном коде заменить кодовое слово c, соответствующее сумме двух самых малых вероятностей, на слова $c0$ и $c1$ (которые и будут кодами соответствующих букв).

Задача 10. Докажите, что код Хаффмана префиксный и его средняя длина минимальна среди всех возможных длин префиксных кодов.