Энтропия Шеннона и кодирование

Пусть фиксированы неотрицательные числа $p_1, \ldots, p_k$, в сумме равные единице (то есть, распределение вероятностей на множестве из k элементнов, называемых в дальнейшем *буквами*). Энтропия Шеннона этого распределения определяется как

$$H = p_1(-\log p_1) + p_2(-\log p_2) + \ldots + p_k(-\log p_k)$$

(при $p = 0$ мы полагаем $p \log p = 0$, доопределяя тем самым функцию $p \mapsto p \log p$ по непрерывности).

Задача 1. Докажите, что энтропия Шеннона неотрицательна. В каком случае она равна нулю?

Задача 2. Докажите, что энтропия Шеннона не меньше *минимальной энтропии*, определяемой как $H_{\min} = \min_i(-\log p_i)$.

Величина $\sum_i p_i \log \frac{p_i}{q_i}$ называется *дивергенцией Кульбака–Лейблера* между распределениями вероятностей p_i и q_i.

Задача 3. Докажите, что дивергенция неотрицательна и обращается в нуль, лишь если $p_i = q_i$ при всех i.

Задача 4. (а) Найдите энтропию Шеннона *равномерного распределения*: $p_1 = \cdots = p_k = 1/k$. (б) Докажите, что энтропия любого неравномерного распределения на k исходах строго меньше энтропии равномерного распределения.

Кодом для алфавита, состоящего из букв $a_1, \ldots, a_k$, 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 *инъективным*, если коды различных букв различны, и *однозначно декодируемым*, если коды любых двух различных слов различны. Код называется *префиксным*, если ни одно из кодовых слов не является началом (префиксом) другого. *Средней длиной кода* называется величина $\sum_i p_i l(c_i)$ (через $l(x)$ обозначаем длину слова x). Средняя длина зависит не только от самого кода, но и от вероятностей букв.

Задача 5. Докажите, что всякий префиксный код является однозначно декодируемым, но обратное неверно.

Задача 6. (а) Докажите, что для любого префиксного кода выполнено *неравенство Крафта*: $\sum_i 2^{-l(c_i)} \leq 1$.
(б) Докажите обратное: для любого набора натуральных чисел $l_1, \ldots, l_k$ с $\sum_i 2^{-l_i} \leq 1$ найдётся префиксный код $c_1, \ldots, c_k$ с $l(c_i) = l_i$.
(в) Докажите, что неравенство Крафта выполнено и для любого однозначаного кода (для однозначных кодов оно называется *неравенством Макмиллана*).

Задача 7. (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 менее чем на 1 превышающей энтропию распределения.

Задача 8. 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 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;
Иначе заменить две самые маленькие вероятности на их сумму, затем запустить алгоритм на полученном распределении (на $k-1$-буквенном алфавите), и наконец 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 9. Докажите, что код Хаффмана префиксный и его средняя длина минимальна среди всех возможных длин префиксных кодов.

Будем алгоритм кодирования (и построенный им код) называть *сбалансированным*, если для некоторой константы d длины кодовых слов удовлетворяют неравенству $l(c_i) < -\log p_i + d$.

Задача 10. Докажите, что код Хаффмана не является сбалансированным.

Задача 11. (Арифметический код.) Отложим на отрезке $[0, 1]$, начиная с начала, без пропусков неперекрывающиеся отрезки длин $p_1, \ldots, p_k$. Для каждого $i = 1, \ldots, k$ выберем отрезок вида $[0.c; 0.c1]$ наибольшей длины, целиком включенный в i-ый из полученных отрезков, и возьмем c в качестве кода для i-ой буквы. Докажите, что $c_1, \ldots, c_k$ есть префиксный код со средней длиной менее $H + 2$.

Задача 12. * Код Шеннона–Фано строится следующим рекурсивным алгоритмом. Сначала упорядочим вероятности по убыванию: $p_1 \geq \ldots \geq p_k$. Затем отложим на отрезке $[0, 1]$, начиная с начала, без пропусков неперекрывающиеся отрезки длин $p_1 \ldots, p_k$. Коды тех букв a_i, соответствующий которым отрезок попал на левую половину отрезка $[0, 1]$, будут начинаться на 0, а коды остальных букв будут начинаться на 1. (Может оказаться, что один из отрезков не лежит целиком ни в левой, ни в правой половинах. Если этот отрезок первый или последний, то начнём его код, соответственно, с 0 или 1. Иначе отнесём его куда угодно.) Далее рекурсивно применяем алгоритм к буквам, код которых начинается на 0, и к буквам, код которых начинается на 1. (На втором и последующих шагах после укладывании будут получаться отрезки, меньшие отрезка $[0, 1]$. Алгоритм делит полученный отрезок ровно пополам.) (а) Докажите, что построенный код является префиксным и его средняя длина не более, чем на константу превосходит энтропию Шеннона. (При этом неважно, что вероятности упорядочены.) (б) Докажите, что если центральный отрезок всегда относить к буквам, код которых начинается на 0, то построенный код сбалансирован. (А здесь это важно.)

Задача 13. Пусть фиксированы вероятности букв $p_1, \ldots, p_k$ и дано произвольное n, для которого все числа $np_1, \ldots, np_k$ — целые. Докажите, что количество слов длины n, в которых i-ая буква встречается ровно np_i раз, равно $2^{nH+O(\log n)}$. Константа, скрытая в $O(\log n)$, зависит от $p_1, \ldots, p_k$, но не зависит от n.