1. Пусть $\sum p_i = 1$ и все $p_i > 0$. Определим функцию $h(p) = \sum p_i \log \frac{1}{p_i}$. К чему стремится значение этой функции для следующих последовательностей:

a) $1/2, 1/4, 1/8, \ldots, 1/2^n, 1/2^n$;

b) $1/3, 1/9, 1/9, \ldots, 1/3^n, 1/3^n, 1/3^n$?

2. Веса 2 монеток выбираются случайно и независимо среди чисел $1, \ldots, 4$. Какова энтропия Шеннона случайной величины, равной результату сравнения на чашечных весах весов первой и второй монетки?

Определение 1

Будем называть кодом функцию $C\colon \{a_1, \ldots, a_n\} \to \{0,1\}^*$, сопоставляющую буквам некоторого алфавита кодовые слова. Если любое сообщение, которое получено применением кода C, декодируется однозначно (т.е. только единственным образом разрезается на образы C), то такой код называется однозначно декодируемым. Код называется префиксным (беспрефиксным, prefix-free), если никакое кодовое слово не является префиксом другого кодового слова.

3. Пусть слова $c_1, \ldots, c_k$ задают префиксный код, и слова $d_1, \ldots, d_\ell$ образуют префиксный код. Докажите, что слова вида $c_i d_j$ также образуют префиксный код.

4. Докажите, что для любого префиксного кода со множеством кодовых слов $a_1, \ldots, a_n$ выполняется следующее неравенство:

$$\sum_{i=1}^{n} 2^{-|a_i|} \le 1.$$

5. Докажите, что если есть набор целых чисел $\ell_1, \ldots, \ell_n$, удовлетворяющие неравенству

$$\sum_{i=1}^{n} 2^{-\ell_i} \le 1,$$

то существует префиксный код с кодовыми словами $c_1, \ldots, c_n$, где $|c_i| = \ell_i$.

Определение 2 (код Шеннона — Фано)

Отсортируем вероятности, $p_1 \ge p_2 \ge \cdots \ge p_n$. «Уложим» вероятности p_i в отрезок $[0,1]$, получая таким образом точки

$$0 \le p_1 < p_1 + p_2 < \cdots < p_1 + p_2 + \cdots + p_n \le 1.$$

Разобьём интервал пополам, и скажем, что все коды, отвечающие точкам слева от разреза, начинаются с нуля, а точкам справа — с единицы. Если отрезок пересекает разрез, и он самый левый (первый), то соответствующий код начинается с 0; если отрезок пересекает разрез и он самый правый (последний), то код начинается с 1. Иначе выбираем ноль или единицу произвольным образом. Продолжаем рекурсивно этот процесс, пока в интервале не останется ровно один отрезок.

6. Докажите, что:

a) код Шеннона — Фано является префиксным;

b) если центральный отрезок относить туда, куда попала его большая часть,то кодирование Шеннона — Фано не является сбалансированным (то есть не существует константы d, для которой выполнено $\ell(c_i) \le -\log p_i + d$ для любых k и любых исходных вероятностей $p_1, \ldots, p_k$);

c) если центральный отрезок всегда относить к правой половине, то кодирование Шеннона — Фано также не является сбалансированным.

7. Приведите пример такого распределения вероятностей, что код Шеннона–Фано не является оптимальным.

Определение 3 (Арифметическое кодирование)

Назовём стандартным интервалом интервал вида $[0.v0, 0.v1)$, где v — некоторая последовательность битов. Уложим вероятности p_i в отрезок $[0, 1]$, получатся точки

$$0 \le p_1 < p_1 + p_2 < \cdots < p_1 + p_2 + \cdots + p_n \le 1.$$

Пусть $(0.v_i0, 0.v_i1)$ — максимальный стандартный интервал в отрезке

$$[p_1 + p_2 + \cdots + p_{i-1}, p_1 + p_2 + \cdots + p_i].$$

Тогда сопоставим i-ой букве код v_i0. Легко заметить, что код получился префиксным, так как если v_i является префиксом v_j, то интервал $(0.v_j0, 0.v_j1)$ вложен в интервал $(0.v_i0, 0.v_i1)$, а такого при построении v_i не может произойти.

8. Докажите, что арифметическое кодирование сбалансировано с константой 2.