А.Г.Дьячков

Теория информации

§7. Последовательные планы поиска

Содержание

1. Асимптотика длины последовательного (M, k)-плана для модели поиска А. Реньи.
2. Префиксный код, кодовое дерево, неравенство Крафта.
3. Обратная и прямая теоремы Шеннона для префиксных кодов.
4. Теорема кодирования для алфавитного кода.
5. Границы минимально возможного числа последовательных проверок в булевой модели поиска m дефектов среди M:
$$\log_2 C_M^m \le N_{noc}(m; M) \le m \log_2 M.$$
6. Оценка среднего числа проверок для биномиальной модели поиска дефектов.

7.1 Определения и обозначения.

Мы будем рассматривать следующую математическую модель задачи, называемой *поиском*. Пусть дано конечное множество
$$A = \{a_1; a_2; \ldots; a_M\},$$
состоящее из M элементов (*факторов*), с некоторым фиксированным, но неизвестным подмножеством $S = \{a_{i_1}, a_{i_2}, \ldots, a_{i_m}\}$, $1 \le i_1 < i_2 < \ldots < i_m \le M$, которое *надо найти*. Элементы подмножества S будем называть *дефектными* элементами (или *значимыми* факторами).

Для поиска $S \subseteq A$ разрешается провести серию из N экспериментов (*групповых проверок*), в каждом из которых можно выбрать некоторое подмножество $T \subseteq A$ и выяснить: *содержит или нет* тестируемое подмножество T *хотя бы один* дефектный элемент. Обозначая символом $\emptyset$ пустое подмножество, можно написать, что двоичный результат проверки $y \in \{0; 1\}$ вычисляется по следующему правилу:
$$y = \begin{cases} 1, \text{ если } S \bigcap T \ne \emptyset, \\ 0, \text{ если } S \bigcap T = \emptyset. \end{cases}$$

Данную модель вычисления результата проверки естественно назвать *булевой* моделью. На выбор подмножества T_n, $n = \overline{1, N}$, которое проверяется в n-ом эксперименте, возможны некоторые ограничения. Этот выбор может зависеть также от результатов $y_1, y_2, \ldots, y_{n-1}$ предыдущих экспериментов.

На основании итога проверок, т.е. двоичной последовательности $y_1^N = (y_1, y_2, \ldots, y_N)$, где

$$y_n = \begin{cases} 1, \text{ если } S \bigcap T_n \neq \emptyset, \\ 0, \text{ если } S \bigcap T_n = \emptyset, \end{cases} \tag{1}$$

экспериментатор должен однозначно найти неизвестное подмножество S. Примерами реальных процедур поиска, сводящихся к описанной модели, является поиск дефектных приборов, поиск эффективных лекарств (ядов), поиск ошибок в программе для ЭВМ, поиск нужных карточек в каталоге библиотеки, радиолокационный поиск и т.п.

Все стратегии (*планы*) поиска естественно классифицировать на *статические,* т.е. такие стратегии, для которых выбор n-ой проверки T_n, $n = 1, 2, \ldots, N$, *не зависит* от результатов $y_1, y_2, \ldots, y_{n-1}$ предыдущих $n-1$ проверок, и *последовательные,* когда такая зависимость допускается. Примером статических процедур поиска одного дефектного элемента ($|S| = m = 1$) при наличии ограничений на проверки ($|T_n| \leq k < \frac{1}{2}, n = \overline{1, N}$) были изученные в §**4** разделяющие планы статического поиска в модели А. Реньи ((M, k)-планы).

Цель данного параграфа — исследование некоторых моделей планов последовательного поиска. Далее в этом разделе и в следующих разделах **7.2** и **7.3** мы рассматриваем последовательные планы поиска множества S, состоящего из одного дефектного элемента. Некоторые модели последовательного поиска множества S, состоящего из нескольких дефектных элементов, разбираются в разделе **7.4**. Способ описания последовательных планов, аналогичных статическим планам из §**4**, покажем на примере последовательного плана поиска одного дефектного элемента (значимого фактора) во множестве A, состоящем из $M = 6$ элементов, т.е. $|S| = m = 1$, $S \in \{a_1; a_2; \ldots; a_6\}$. Данный план состоит из $N = 3$ проверок и в каждой проверке число проверяемых элементов $|T_n| \leq 2$, $n = \overline{1, 3}$. Его можно назвать последовательным $(6, 2)$-планом для модели поиска А. Реньи и записать в виде таблицы:

$$X = \begin{array}{c} \begin{matrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \end{matrix} \\ \left[\begin{array}{cc|cc|cc} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ - & - & 1 & 0 & 1 & 0 \end{array}\right] \end{array}, \tag{2}$$

в которой номерам единичных позиций n-ой строки X соответствуют номера элементов, составляющих проверяемое подмножество T_n, $n = \overline{1, N}$.

Такая запись последовательного плана X означает, что тестируемое подмножество первой проверки $T_1 = \{a_1; a_2\}$. **а)** Если результат первой проверки $y_1 = 1$, то для второй проверки выбираем $T_2 = \{a_1\}$ и если результат второй проверки $y_2 = 1$ ($y_2 = 0$), то однозначно определяется $S = a_1$ ($S = a_2$). В этом случае третья проверка не нужна. **б)** Если результат первой проверки $y_1 = 0$, то во второй проверке тестируется $T_2 = \{a_3; a_4\}$ и в зависимости от результата второй проверки $y_2 = 1$ ($y_2 = 0$) подмножество, тестируемое в третьей проверке — $T_3 = \{a_3\}$ ($T_3 = \{a_5\}$). По результатам трех (y_1, y_2, y_3) (или двух (y_1, y_2)) проверок однозначно определяется дефектный элемент $S \in A$, который в таблице (2) соответствует столбцу X вида (y_1, y_2, y_3) или (y_1, y_2). Например, если $(y_1, y_2, y_3) = (0, 0, 1)$, то $S = \{a_5\}$.

Согласно теореме 8 из §**4**, длина оптимального статического $(M, \leq k)$-плана X при $M = 6$, $k = 2$ есть

$$N_{\text{ст}} = \lceil 2(M-1)/(k+1) \rceil = \lceil 10/3 \rceil = 4 > N_{\text{пос}} = 3,$$

а таблица проверок, соответствующих этому оптимальному статическому плану X, имеет вид

$$X_{\text{ст}} = \begin{array}{|cccccc|} \multicolumn{1}{c}{a_1} & a_2 & a_3 & a_4 & a_5 & \multicolumn{1}{c}{a_6} \\ \hline 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ \hline \end{array}. \qquad (3)$$

Сравнение таблиц (2) и (3) показывает возможность уменьшения числа проверок N за счет последовательного планирования.

Задача 7.1. Пусть $1 \le k \le M/2$. Укажите (M,k)-план последовательного поиска в модели А. Реньи, длина которого N является минимально возможной и вычисляется по формуле

$$N = N_{\text{пос}}(M;k) = m + \lceil \log_2(M - mk) \rceil,$$

$$m = \left\lceil \frac{M}{k} \right\rceil - 2.$$

Указание. Воспользуйтесь тем, что значение $x = m$ является наименьшим целочисленным решением неравенства $M - kx \le 2k$, т.е. $x \ge \dfrac{M}{k} - 2$.

Задача 7.2. Докажите, что при $k = \lceil Mp \rceil$, $0 < p \le 1/2$, введенное в задаче 7.1 число удовлетворяет неравенству

$$N_{\text{пос}}(M, \lceil Mp \rceil) \le \log_2 M + \frac{1}{p} + \log_2(2p+2),$$

откуда следует, что при фиксированном p, $0 < p \le 1/2$ и $M \to \infty$ величина

$$N_{\text{пос}}(M, \lceil Mp \rceil) = \log_2 M(1 + o(1)).$$

Сравните этот результат с асимптотикой длины оптимального статического $(M, \lceil Mp \rceil)$-плана, вычисленной в теореме 8 из §4.

Отметим, что если по плану $X_{\text{ст}}$ (3) проводить последовательные проверки, то таблица соответствующего последовательного плана имеет вид

$$X = \begin{array}{|c|c|ccc|c|} \multicolumn{1}{c}{a_1} & \multicolumn{1}{c}{a_2} & a_3 & a_4 & \multicolumn{1}{c}{a_5} & \multicolumn{1}{c}{a_6} \\ \hline \multicolumn{1}{|c}{0} & \multicolumn{1}{c}{1} & 0 & 0 & \multicolumn{1}{c}{0} & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & - & - & 1 & 0 & - \\ 0 & - & - & - & 1 & - \\ \hline \end{array}. \qquad (4)$$

Таблица

$$X = (\mathbf{x}(a_1), \mathbf{x}(a_2), \ldots, \mathbf{x}(a_M))$$

последовательного плана поиска одного дефектного элемента $S \in A$, аналогичная (2) и (4), в общем случае состоит из M столбцов, где двоичный столбец $\mathbf{x}(a_m)$, $m = \overline{1, M}$, представляет собой результаты проверок, если дефектный элемент $S = a_m$. Строки таблицы X задают тестируемые подмножества соответствующих проверок. Пусть

$$L(m) = l(\mathbf{x}(a_m)), \ m = \overline{1, M},$$

обозначает длину (число двоичных символов) столбца (*слова*) $\mathbf{x}(a_m)$.

Отметим, что в статическом плане длина любого слова $\mathbf{x}(a_m)$ одна и та же и равна длине плана $N = L(m),\ m = \overline{1, M}$. Для последовательного плана слова $\mathbf{x}(m)$ могут иметь разные длины $L(m)$ и, *по определению*, длина последовательного плана

$$N = \max_{m=\overline{1,M}} L(m). \tag{5}$$

Заметим также, что статический план является частным случаем последовательного плана.

Сопоставлением с примерами таблиц планов поиска (2) - (4) легко установить справедливость следующего утверждения.

Лемма 1 *Таблица X, состоящая из M двоичных столбцов $\mathbf{x}(a_m)$, $\overline{1, M}$, с длинами $L(m)$ является последовательным планом поиска (с однозначным восстановлением) одного дефектного элемента $S \in A = \{a_1; a_2; \ldots; a_M\}$ тогда и только тогда, когда выполнены следующие два условия:*

а) *если при $m \neq m'$ длина $L(m) = L(m')$, то слово $\mathbf{x}(a_m) \neq \mathbf{x}(a_{m'})$,*

б) *если при $m \neq m'$ длина $L(m') < L(m)$, то более короткое слово $\mathbf{x}(a_{m'})$ не является началом более длинного слова $\mathbf{x}(a_m)$.*

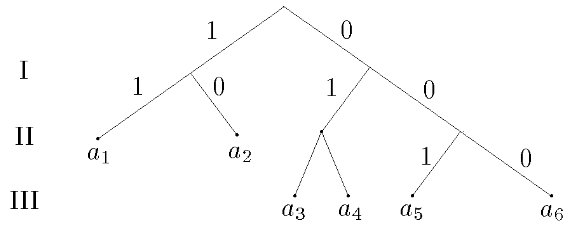
Таблица X, обладающая свойствами **а)** и **б)**, называется *кодом со свойствами префикса* (или *префиксным кодом*).

7.2 Кодовые деревья

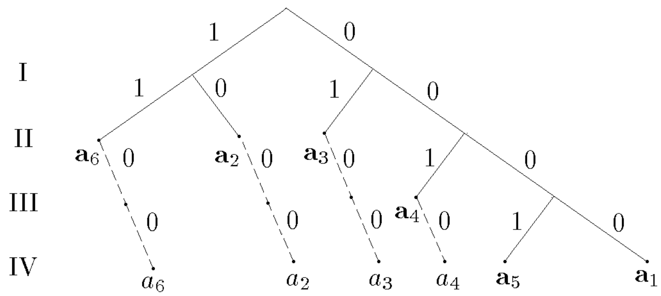
Рассмотрим полное двоичное N-уровневое *дерево,* которое имеет 2^N *ветвей* и столько же *концевых узлов* на уровне N. В этом дереве имеется один начальный узел, на нулевом уровне и 2^n узлов на уровне $n = \overline{1, N}$. Для определенности, будем считать что дерево "растет" сверху вниз, из каждого узла на уровне $n = \overline{0, N-1}$ *выходят два ребра,* входящие в узлы уровня $n+1$. Левое ребро *пометим* символом 1 и назовем *единичным* ребром, а правое ребро пометим символом 0 и назовем *нулевым* ребром. В каждый из узлов на уровне $n = \overline{1, N}$ *входит одно ребро,* помеченное либо 0 либо 1 и *выходящее из узла уровня* $n-1$.

Пусть префиксный код X имеет длину N, определяемую согласно (5). Кодовому слову $\mathbf{x}(a_m)$ длины $L(m)$, $m = \overline{1, M}$, в полном двоичном дереве взаимно-однозначно соответствует *кодовый путь* (последовательность ребер, помеченных двоичными символами слова $\mathbf{x}(a_m)$), который *выходит* (начинается) из начального узла и *входит* (оканчивается) в некоторый узел на уровне $n = L(m)$. Этот узел назовем *конечным кодовым узлом* и пометим символом a_m. Совокупность всех M кодовых путей в двоичном N-уровневом дереве, которое выходят из начального узла и заканчиваются в узлах, помеченных символами a_m, $m = 1, M$, назовем *кодовым деревом* префиксного кода X.

Приведем кодовые деревья для префиксных кодов (2) и (4).



(2′)



(4′)

На рис. (4') кодовое дерево для последовательного плана (4) пунктиром дорисовано до кодового дерева статического плана (3).

Свойство префиксности кода в терминах кодового дерева означает, что через конечный кодовый узел, помеченный символом a_m, может проходить только один кодовый путь, оканчивающийся в этом узле и соответствующий слову $\mathbf{x}(a_m)$.

Задача 7.3. Докажите, что кодовое дерево соответствует $(M, \leq k)$-плану последовательного поиска в модели А. Реньи тогда и только тогда, когда через любое его единичное ребро проходит $\leq k$ кодовых путей. Сравните это утверждение с кодовыми деревьями (2') (4').

Необходимое и достаточное условие существования префиксного кода (кодового дерева) с заданным набором длин кодовых слов (кодовых путей) $1 \leq L(1) \leq L(2) \leq \ldots \leq L(M) = N$ дает

Теорема 1 (Неравенство Крафта) 1) *Длины кодовых путей $L(m)$, $m = \overline{1, M}$, любого кодового дерева удовлетворяют неравенству*

$$\sum_{m=1}^{M} 2^{-L(m)} \leq 1. \tag{6}$$

2) *Наоборот, если некоторый набор натуральных чисел $1 \leq L(1) \leq L(2) \leq \ldots \leq L(M) = N$ удовлетворяет неравенству (6), то существует кодовое дерево с M кодовыми путями, длинами которых являются числа $L(m)$, $m = \overline{1, M}$.*

Доказательство. 1) Рассмотрим произвольный конечный узел кодового дерева, помеченный символом a_m, $m = \overline{1, M}$. Так как в исходном *полном двоичном дереве* этот узел находится на уровне $n = L(m)$, то из него, спускаясь вниз по дереву, можно попасть *ровно* в $2^{N-L(m)}$ концевых узлов (уровня N) полного двоичного дерева. В силу свойства префиксности, множества таких концевых узлов, соответствующих различным значениям $m=\overline{1,M}$ *не пересекаются.* Отсюда, учитывая что общее число концевых узлов полного двоичного дерева равно 2^N, получаем

$$\sum_{m=1}^{M} 2^{N-L(m)} \le 2^N,$$

что равносильно (6). Утверждение **1)** доказано.

2) *По индукции.* Пусть $1 \le L(1) \le L(2) \le \ldots \le L(m) \le L(m+1) \le \ldots \le L(M) = N$, где $1 \le m < M$, и в полном двоичном дереве уже построены кодовые пути с длинами $L(i),\ i=\overline{1,m}$, конечные кодовые узлы которых помечены элементами $a_i,\ i=\overline{1,m}$, соответственно. Из неравенства (6) следует, что

$$\sum_{i=1}^{m} 2^{-L(i)} < 1.$$

Поэтому всегда найдется узел полного двоичного дерева любого уровня $n,\ L(m) \le n \le N$, который при $n=L(m+1)$ можно использовать в качестве конечного кодового узла на уровне $L(m+1)$ и пометить символом a_{m+1}. Утверждение **2)** доказано.

Теорема 1 доказана.

Задача 7.4. Проверьте, что для кодового дерева (2') в неравенстве Крафта достигается равенство, а для кодового дерева (4') в неравенстве Крафта — строгое неравенство. Объясните почему?

7.3 Теоремы кодирования для префиксных кодов.

Отметим важное для приложений свойство префиксных кодов, которые используются при кодировании сообщений, подлежащих передаче по каналу связи. Будем интерпретировать множество $A=\{a_1,a_2,\ldots,a_M\}$ как конечный алфавит из M символов. Из символов A составлено *сообщение*

$$\mathbf{a}=(a_{i_1},a_{i_2},\ldots,a_{i_n},\ldots),\ a_{i_n}\in A,$$

которое необходимо передавать по двоичному каналу без шума. Например, если сообщение $\mathbf{a}$ записано в русском алфавите, то в его M символов входят 32 буквы, знаки препинания, символ пропуска между словами и т.п. Заменим буквы сообщения $\mathbf{a}$ соответствующими кодовыми словами префиксного кода и через $\mathbf{x}(\mathbf{a})$ обозначим получившуюся двоичную последовательность

$$\mathbf{a} \Rightarrow \mathbf{x}(\mathbf{a})=(\mathbf{x}(a_{i_1}),\mathbf{x}(a_{i_2}),\ldots,\mathbf{x}(a_{i_n}),\ldots).$$

Далее $\mathbf{x}(\mathbf{a})$ передается по двоичному каналу без шума, на выходе которого приемник по неискаженной двоичной последовательности $\mathbf{x}(\mathbf{a})$ должен восстановить переданное сообщение

$$\mathbf{x}(\mathbf{a}) \Rightarrow \mathbf{a}=(a_{i_1},a_{i_2},\ldots,a_{i_k},\ldots)$$

Очевидно следующее важное

Свойство префиксного кода. *Если известно начало последовательности* $\mathbf{x}(\mathbf{a})$, *то в силу свойства префиксности кода* X, *при чтении* $\mathbf{x}(\mathbf{a})$ *слева направо однозначно восстанавливается исходное сообщение* $\mathbf{a}$.

Пусть на множестве $A = \{a_1, a_2, \ldots, a_M\}$ задано распределение вероятностей $\mathbf{p} = (p_1, p_2, \ldots, p_M)$, где

$$p_m = \Pr\{a_m\},\ 0 < p_m < 1, \sum_{m=1}^{M} p_m = 1.$$

В задаче последовательного поиска число p_m интерпретируется как вероятность данному элементу $a_m,\ m = \overline{1, M}$, быть дефектным (значимым). В задаче кодирования сообщений число p_m можно рассматривать как вероятность (частоту) появления буквы $a_m,\ m = \overline{1, M}$, в письменном тексте. Следовательно, важным критерием префиксного кода X является его средняя длина

$$\overline{L} = \sum_{m=1}^{M} L(m) p_m, \tag{7}$$

где $L(m) = l(\mathbf{x}(a_m))$ — длина (число двоичных символов) слова $\mathbf{x}(a_m)$.

Пусть

$$H(\mathbf{p}) = -\sum_{m=1}^{M} p_m \log p_m \tag{8}$$

энтропия Шеннона распределения вероятностей $\mathbf{p}$, где в определении (7) и далее в этом разделе используются двоичные логарифмы и экспоненты. Имеет место следующее утверждение, называемое *теоремой кодирования для префиксных кодов.*

Теорема 2 *Пусть на множестве A задано распределение вероятностей $\mathbf{p}$. Тогда справедливы следующие два утверждения, называемые соответственно обратной и прямой теоремами Шеннона для префиксных кодов.* **1)** *Для любого префиксного кода средняя длина*

$$\overline{L} \geq H(\mathbf{p}). \tag{9}$$

2) *Существует префиксный код со средней длиной*

$$\overline{L} \leq H(\mathbf{p}) + 1. \tag{10}$$

Доказательство. 1) Если $L_m,\ m = \overline{1, M}$ — длины слов произвольного фиксированного префиксного кода, то в силу неравенства Крафта (6), число

$$Q = \sum_{m=1}^{M} 2^{-L(m)} \leq 1. \tag{11}$$

Введем распределение вероятностей $\mathbf{q} = (q_1, q_2, \ldots, q_M)$, где

$$q_m = 2^{-L(m)}/Q,\ m = \overline{1, M}. \tag{12}$$

Применяя доказанное в §5 свойство любых двух распределений вероятностей $\mathbf{p}$ и $\mathbf{q}$, а затем определение (8), имеем

$$H(\mathbf{p}) \leq -\sum_{m=1}^{M} p_m \log q_m = -\sum_{m=1}^{M} p_m \log \frac{2^{-L(m)}}{Q} =$$

$$= \sum_{m=1}^{M} L(m)p_m + \log Q = \overline{L} + \log Q \le \overline{L}$$

где воспользовались определением (7) и неравенством (11). Утверждение **1)**, т.е. неравенство (9), доказано.

2) Для заданного $\mathbf{p} = (p_1, p_2, \ldots, p_M)$ определим числа $L(m)$, $m = \overline{1, M}$, следующим образом

$$L(m) = \lceil -\log p_m \rceil = \lceil \log 1/p_m \rceil. \tag{13}$$

Имеем

$$-\log p_m \le L(m) \le -\log p_m + 1 \tag{14}$$

Из левой части (14) следует $2^{-L(m)} \le p_m$, $m = \overline{1, M}$. Поэтому для набора чисел (13) выполняется неравенство

$$\sum_{m=1}^{M} 2^{-L(m)} \le \sum_{m=1}^{M} p_m = 1.$$

В силу теоремы 1, это означает существование префиксного кода, длины слов которого определены (13). В силу правой части (14), для средней длины такого префиксного кода имеем

$$\overline{L} = \sum_{m=1}^{M} p_m L(m) \le \sum_{m=1}^{M} p_m(-\log p_m + 1) = H(\mathbf{p}) + 1.$$

Утверждение 2), т.е. неравенство (10) доказано. Теорема 2 доказана.

Рассмотрим важный частный случай префиксных кодов, обладающих свойством лексикографической упорядоченности. Это свойство означает следующее. Кодовому слову

$$\mathbf{x}(a_m) = (x_1(a_m), x_2(a_m), \ldots, x_{L(m)}(a_m)), \ m = \overline{1, M},$$

префиксного кода сопоставим действительное число

$$q(m) = \sum_{n=1}^{L(m)} x_n(a_m) 2^{-n}, \ 0 \le q(m) < 1,$$

двоичное разложение которого задается кодовым словом $\mathbf{x}(a_m)$.

Определение. Префиксный код X называется *алфавитным,* если при $m' < m$ число $q(m') < q(m)$.

Очевидно следующее важное

Свойство алфавитного кода. В последовательном плане поиска одного дефектного элемента, который проводится в соответствии с алфавитным кодом, *множество* T_n, $n = \overline{1, N}$, элементов, тестируемых в n-ой групповой проверке, обязательно *состоит из соседних элементов множества* A.

Пример. Алфавитный код при $M = 6$ длины 4.

$$X = \begin{array}{c} \begin{array}{cccccc} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \end{array} \\ \boxed{\begin{array}{cccccc} 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1\,| & 0 & 0 & 0 & 1 \\ - & - & 0 & 1 & 1 & - \\ & & - & 0 & 1 & - \end{array}} \end{array}.$$

Отметим, что код (4) не является алфавитным, а код (2) становится алфавитным путем замены его нулевых элементов на единичные и наоборот.

Имеет место следующая теорема существования алфавитных кодов.

Теорема 3 *Для произвольного распределения вероятностей* $\mathbf{p}$ *существует алфавитный код со средней длиной*

$$\overline{L} \le H(\mathbf{p}) + 2, \tag{15}$$

где $H(\mathbf{p})$ *— энтропия (8).*

Доказательство. Для заданного распределения вероятностей $\mathbf{p} = (p_1, p_2, \ldots, p_M)$ построим монотонно возрастающую последовательность чисел

$$q(m) = \begin{cases} p_1/2, \text{ если } m = 1, \\ \sum\limits_{i=1}^{m-1} p_i + \frac{1}{2}p_m, \text{ если } m = \overline{2, M}. \end{cases}$$

Заметим, что при $m' < m$ разность

$$q(m) - q(m') = \frac{1}{2}p_{m'} + \sum_{m'<i<m} p_i + \frac{1}{2}p_m,$$

а потому для любого $m' \ne m$ значение

$$|q(m) - q(m')| > \max\left\{\frac{p_m}{2}; \frac{p_{m'}}{2}\right\}. \tag{16}$$

Введем числа $L(m) = \lceil -\log p_m \rceil + 1$, для которых справедливы неравенства

$$1 - \log p_m \le L(m) \le -\log p_m + 2. \tag{17}$$

Из левого неравенства (17) следует, что

$$2^{-L(m)} \le 2^{\log p_m - 1} = \frac{1}{2}p_m, \ m = \overline{1, M}. \tag{18}$$

Построим теперь код $X = (\mathbf{x}(a_1), \mathbf{x}(a_2), \ldots, \mathbf{x}(a_M))$, выбирая в качестве кодового слова $\mathbf{x}(a_m)$, $m = \overline{1, M}$, первые $L(m) = \lceil -\log p_m \rceil + 1$ знаков в двоичном разложении числа $q(m)$. Такой код, в силу монотонности $q(m)$, $m = \overline{1, M}$, обладает свойством *лексикографической упорядоченности.*

Докажем (от противного), что X обладает свойством *префиксности.* Пусть $\mathbf{x}(a_m)$ и $\mathbf{x}(a_{m'})$ такие, что $L(m') \le L(m)$ и $\mathbf{x}(a_{m'})$ является началом $\mathbf{x}(a_m)$. По построению кодовых слов $\mathbf{x}(a_m)$, $m = \overline{1, M}$, в этом случае очевидно, что

$$|q(m) - q(m')| < 2^{-L(m')} \le \frac{1}{2}p_{m'},$$

где второе неравенство вытекает из (18). Последнее неравенство противоречит (16).

Для средней длины построенного алфавитного кода X, в силу правой части (17), имеем

$$\overline{L} = \sum_{m=1}^{M} L(m)p_m \leq \sum_{m=1}^{M} p_m(-\log p_m + 2) = H(\mathbf{p}) + 2,$$

т.е. неравенство (15).

Теорема 3 доказана.

7.4 Последовательный поиск нескольких дефектов.

а) Гипергеометрическая модель.

Рассмотрим две ситуации такого поиска. 1) Известен объем дефектного множества $S = \{a_{i_1}, a_{i_2}, \ldots, a_{i_m}\} \subseteq A$, т.е. известно число m, $1 < m < M$. 2) Для заданного числа m, $1 < m < M$ известно, что число дефектных элементов $|S| \leq m$.

Задача 7.5. а) Докажите, что для модели 1) необходимое число последовательных проверок

$$N \geq \log C_M^m,$$

а для модели 2) необходимое число последовательных проверок

$$N \geq \log \sum_{i=0}^{m} C_M^i.$$

б) Для моделей 1) и 2) постройте стратегии последовательного поиска, которые находят дефектное множество S за $N \leq m \log M = |S| \log M$ проверок.

б) Биномиальная модель.

Пусть число элементов дефектного множества $|S|$ имеет биномиальное распределение, т.е.

$$\mathbf{Pr}\{|S| = m\} = C_M^m p^m q^{M-m},\ 0 \leq m \leq M,$$

где $0 < p < q < 1$, $p + q = 1$, — фиксированные числа. Это означает, что каждый элемент множества $A = \{a_1, a_2, \ldots, a_M\}$ является независимо от других дефектным с вероятностью p и — недефектным с вероятностью $q = 1 - p$. Зафиксируем некоторое число $k = 1, 2, \ldots$ и рассмотрим следующий

Последовательный план поиска. Разбиваем все множество $A = \{a_1, a_2, \ldots, a_M\}$ на M/k групп, где в каждой группе по k элементов. Делаем групповые тестирования всех групп поочередно, т.е. $T_1 = \{a_1; \ldots; a_k\}$, $T_2 = \{a_{k+1}; \ldots; a_{2k}\}$, ... Если результат тестирования группы T_n, $n = \overline{1; M/k}$, положителен, т.е. $y_n = 1$, то проверяем все элементы группы индивидуально.

Пусть $\xi^{(k)}$ — число тестов (случайная величина), которые надо израсходовать на одну группу. Очевидно,

$$\xi^{(k)} = \begin{pmatrix} 1 & k+1 \\ (1-p)^k & 1-(1-p)^k \end{pmatrix}.$$

Пусть η_M — число тестов (случайная величина), которые затрачены на поиск дефектного множества S. Очевидно, что среднее значение

$$\overline{\eta}_M = \frac{M}{k}\overline{\xi^{(k)}} = \frac{M}{k}\{(1-p)^k + (k+1)[1-(1-p)^k]\} =$$

$$= M\left[\frac{k+1}{k} - (1-p)^k\right] = M\left[\frac{1}{k} + 1 - (1-p)^k\right] \leq$$

$$\leq M\left(\frac{1}{k} + pk\right),$$

где воспользовались неравенством Бернулли $(1+x)^k \geq 1 + xk,\ x \geq -1$.

Положим теперь $k = 1/\sqrt{p}$, т.е. выберем значение k, которое минимизирует правую часть оценки для $\overline{\eta}_M$. Имеем

$$\overline{\eta}_M \leq 2M\sqrt{p}.$$

Эта оценка показывает, что групповое тестирование заведомо выгоднее индивидуального, когда $2\sqrt{p} < 1$, т.е. при $p < 1/4$.

Пример. Если $p = 1/100$, то оптимальный об‘ем тестируемой группы $k = 10$. При этом среднее число тестов $\overline{\eta}_M$, затрачиваемых при описанном групповом тестированиии на поиск дефектного множества $S \subseteq A = \{a_1, a_2, \ldots, a_M\}$, не превышает $M/5$. Следовательно, при вероятности дефектного элемента $p \leq 1/100$ групповое тестирование при поиске дефектов по крайней мере в 5 раз выгоднее проведения поэлементных проверок.