ГЛОССАРИЙ

§ 9 Постановка задачи равномерного кодирования

Алфавитом кода называют некоторое множество $A = \{a_1, \ldots, a_D\}$; *кодовыми символами* называют элементы множества A; *кодовым словом* называют последовательность кодовых символов; *кодом над алфавитом* A называют любое семейство кодовых слов; *объемом кода* называют число элементов в коде. Другими словами, *код* — это множество последовательностей $a_{i_1}a_{i_2}\ldots a_{i_t}$ элементов из A.

Код называется *равномерным*, если все его слова имеют одинаковую длину m, это число называется *длиной кода*. Если хотя бы два кодовых слова имеют различные длины, то код называют *неравномерным*.

Кодированием сообщений ансамбля X посредством кода называется отображение (необязательно взаимно однозначное) множества сообщений во множество кодовых слов.

Общая схема равномерного кодирования

X — ансамбль, X^n — последовательность сообщений длины n,
A — алфавит, $D = |A|$ — число символов в алфавите,
A^m — все последовательности длины m,
Равномерное кодирование — это отображение

$$f \ : \ X^n \to C \subseteq A^m, \quad C = f(X^n).$$

Число $R =: \frac{\log |C|}{n \log D}$ называется *скоростью равномерного кодирования источника при разбиении последовательности сообщений на блоки длины n.*

При двоичном кодировании $A = \{0, 1\}$, $D = |A| = 2$, и скорость равномерного кодирования будет измеряться в двоичных символах на сообщение:

$$R =: \frac{\log |C|}{n} \text{ (бит/символ источника).} \tag{9.1}$$

Поэтому нас будет интересовать кодирования с наименьшей скоростью.

Теорема 9.1. *Кодирование с однозначным декодированием возможно только тогда, когда скорость кодирования не меньше энтропии ансамбля.*

Рассмотрим ситуацию, когда $|X|^n > |C|$. Тогда кодовых слов недостаточно для того, чтобы сопоставить каждой последовательности источника свое кодовое слово. Выделим в X^n подмножество T, такое, что $|T| = |C|$, и каждой последовательности из множества T сопоставим индивидуальное (единственное) кодовое слово. Множество T будем называть *множеством однозначно кодируемых последовательностей*. Остальным последовательностям из $X^n \setminus T$ сопоставим произвольные кодовые слова. Множество $X^n \setminus T$ будем называть *множеством неоднозначно кодируемых и декодируемых блоков*. Декодер при получении некоторого кодового слова $\overline{c} \in C$ будет выдавать получателю соответствующую этому слову последовательность из множества T.

Когда источник породит последовательность из дополнения к множеству T, выход декодера не будет совпадать со входом кодера. Это событие называют ошибкой кодирования и вероятность

$$\mathrm{P}_e = \Pr\,(x \notin T) = 1 - \sum_{x \in T} p(x)$$

называется *вероятностью ошибки кодирования.*

Вывод: *При кодировании с помощью равномерных кодов основная задача состоит в определении наименьшей возможной скорости кодирования, при которой вероятность ошибки может быть сделана произвольно малой. Наименьшая достижимая скорость кодирования является характеристикой источника сообщений и называется скоростью создания информации.*

§ 10 Постановка задачи неравномерного кодирования

Обозначим через m_i длину слова, кодирующего сообщение $x_i \in X$. Пусть $p(x_i)$ — вероятность этого сообщения. Число

$$\bar{m}(X) =: \sum_{x_i \in X} m_i p(x_i) \tag{10.1}$$

называется *средней длиной кодовых слов*, кодирующих ансамбль сообщений $\{X, p(x)\}$.

Предположим, что неравномерный код используется для кодирования отрезков сообщений длины n, т. е. для кодирования ансамбля $\{X^n, p(\bar{x})\}$, где $p(\bar{x})$ — распределение вероятностей на X^n, задаваемое с помощью задания источника, $\bar{x} \in X^n$. Пусть $\bar{m}(X^n)$ — средняя длина

кодовых слов.

Число

$$R =: \frac{\bar{m}(X^n)\log D}{n} \tag{10.2}$$

называется *средней скоростью неравномерного кодирования посредством D-ичного кода при разбиении последовательности сообщений на блоки длины n*. Средняя скорость неравномерного кодирования измеряется в двоичных символах на сообщение.

Средняя скорость неравномерного кодирования зависит от выбора n и множества кодовых слов. Наша цель состоит в определении наименьшей достижимой средней скорости.

Коды, в которых ни одно слово не является началом другого, называются префиксными. Коды, в которых ни одно слово не является окончанием другого, называются суффиксным. Коды, в которых любая последовательность кодовых слов допускает однозначное разбиение на кодовые слова, называются кодами со свойством однозначного декодирования.

Теорема 10.1. Если побуквенное (алфавитное) кодирование является префиксным или суффиксным, то оно однозначно декодируемое.

Теорема 10.2. (Неравенство Крафта)

1. Если существует префиксный код в алфавите объема D с длинами кодовых слов $m_1, m_2, \ldots, m_M$, то

$$\sum_{i=1}^{M} D^{-m_i} \le 1. \tag{10.3}$$

2. Если натуральные числа D, $m_1, m_2, \ldots, m_M$, удовлетворяют неравенству (10.3), то существует префиксный код в алфавите объема D с длинами кодовых слов $m_1, m_2, \ldots, m_M$. ¤

При $A = \{0, 1\}$ получаем $D = |A| = 2$ и

Следствие 10.2.1. (Неравенство Крафта) Для того чтобы существовал префиксный код в двоичном алфавите с длинами кодовых слов m_1, m_2, ..., m_M, необходимо и достаточно, чтобы $\sum_{i=1}^{M} 2^{-m_i} \le 1$.

Теорема 10.3. (Неравенство Мак–Миллана) Если алфавитное кодирование с длинами кодовых слов m_1, m_2, ..., m_M, является однозначно декодируемым, то справедливо неравенство (10.3).

Из теоремы 10.3 и второго утверждения теоремы 10.2 вытекает

Следствие 10.3.1. Если алфавитное кодирование с длинами кодовых слов $m_1, m_2, \ldots, m_M$, является однозначно декодируемым, то существует префиксный код с таким же набором длин кодовых слов.

Теорема 10.4. Средняя длина $\bar{m}$ лучших префиксных двоичных кодов лежит в границах $H(X) \le \bar{m} \le 1 + H(X)$.

Для кодирования сообщений источника пригодны только те коды, которые допускают однозначное декодирование.

Скоростью создания информации дискретным источником при неравномерном кодировании называется наименьшее число H такое, что для любого $R > H$ найдется n (длина кодируемых сообщений) и неравномерный код со средней скоростью кодирования R, который допускает однозначное декодирование.

Теорема 10.5. Для любого кода со свойством однозначного декодирования справедливо неравенство

$$\bar{m}(X) \ge \frac{H(X)}{\log D}.$$

Кроме того, если вероятности сообщений являются целыми отрицательными степенями числа D: $p(x_i) = D^{-m_i}$, $i = 1, 2, \ldots, M$, то существует D-ичный код, имеющий среднюю длину

$$\bar{m}(X) = \frac{H(X)}{\log D},$$

равную нижней границе.

Теорема 10.6. Существует D-ичный код со свойством однозначного декодирования, для которого

$$\bar{m}(X) < \frac{H(X)}{\log D} + 1.$$

Следствие 10.6.1. Средняя длина $\bar{m}$ кодовых слов лучших неравномерных кодов лежит в границах

$$\frac{H(X)}{\log D} \le \bar{m}(X) < \frac{H(X)}{\log D} + 1.$$

Теоремы 10.5 и 10.6 представляют собой прямую и обратную теоремы при побуквенном кодировании источника, выбирающего сообщения из ансамбля $\{X, p(x)\}$.

§ 11 Оптимальные неравномерные коды

Пусть сообщения в ансамбле

$$X = \{x_1, \dots x_M\}, \quad p(x_1) \geq p(x_2) \geq \dots \geq p(x_M),$$

имеют произвольные вероятности. Будет предполагаться, что сообщения в ансамбле упорядочены по убыванию.

Ограничимся рассмотрением префиксных кодов. Рассматривать будем только двоичный случай: $A = \{0, 1\}$, $D = 2$.

Лемма 11.1. *В оптимальном коде слово, соответствующее наименее вероятному сообщению, имеет наибольшую длину.*

Лемма 11.2. *В оптимальном двоичном префиксном коде два наименее вероятных сообщения кодируются словами одинаковой длины, одно из которых оканчивается нулем, а другое единицей.*

Рассмотрим методы построения двоичных префиксных кодов, которые в ряде случаев приводят к кодам с минимально возможной средней длиной. Достоинством предлагаемых методов является простота их реализации. Впервые такие коды предложили одновременно в 1948–49 годах Р. Фано и К. Шеннон.

Код Фано

Алгоритм Фано сводится к последовательному выполнению следующих шагов:

1. Сообщения ансамбля $X = \{x_1, \dots, x_M\}$ упорядочиваем по убыванию вероятностей: $p(x_1) \geq p(x_2) \geq \dots \geq p(x_M)$.

2. Разбиваем ансамбль X на два подансамбля $X^{(0)}$ и $X^{(1)}$ с помощью некоторого порогового целого числа $1 \leq k^{(1)} \leq M - 1$, так, чтобы абсолютная величина

$$K^{(1)} = \Big| \sum_{i=1}^{k^{(1)}} p(x_i) - \sum_{i=k^{(1)}+1}^{M} p(x_i) \Big| \tag{11.4}$$

достигала наименьшего возможного значения. Сообщениям подансамбля $X^{(0)}$ приписываем 0, сообщениям подансамбля $X^{(1)}$ приписываем 1.

3. Если подансамбли $X^{(0)}$, $X^{(1)}$ состоит более чем из двух сообщений, то разбиваем множество сообщений каждого из них на две части $X^{(00)}$, $X^{(01)}$ и $X^{(10)}$, $X^{(11)}$, соответственно, с помощью пороговых целых чисел

$$1 \leq k^{(11)} \leq k^{(1)} - 1, \; k^{(1)} \leq k^{(12)} \leq M - 1,$$

так, чтобы абсолютные величины

$$K^{(21)} = \Big| \sum_{i=1}^{k^{(11)}} p(x_i) - \sum_{i=k^{(11)}+1}^{k^{(1)}} p(x_i) \Big|, \tag{11.5}$$

$$K^{(22)} = \Big| \sum_{i=k^{(1)}+1}^{k^{(12)}} p(x_i) - \sum_{i=k^{(12)}+1}^{M} p(x_i) \Big| \tag{11.6}$$

достигали наименьших возможных значений. Сообщениям из $X^{(00)}$, $X^{(10)}$ с нулевыми последними индексами приписываем 0, сообщениям из подгрупп $X^{(01)}$, $X^{(11)}$ с единичными последними индексами приписываем 1.

Если подансамбль $X^{(ij)}$, $i, j \in \{0, 1\}$ состоит более чем из одного сообщения, то переходим к шагу 4. Если все подансамбли содержат по одному сообщению, то переходим к шагу 5.

4. Если есть подансамбли, состоящие более чем из одного сообщения, то разбиваем каждый из них на две подансамбля, исходя из соотношения, аналогичного (11.4). Сообщениям из подансамблей с нулевыми последними индексами приписываем нуль, сообщениям из подансамблей с единичными последними индексами приписываем единицу.

Если все образовавшиеся подансамбли состоят из одного сообщения, то переходим к шагу 5.

Если есть подансамбли, состоящие более чем из одного сообщения, то то повторяем шаг 4.

5. Если образовавшиеся подансамбли состоят из одного сообщения, то последовательно, начиная с первой метки, выписываем нули и единицы, относящиеся к каждому сообщению ансамбля X.

В итоге получается двоичный префиксный код для заданного ансамбля X.

Код Шеннона

Алгоритм Шеннона сводится к последовательному выполнению следующих шагов:

1. Сообщения ансамбля $X = \{x_1, \ldots, x_M\}$ упорядочиваем по убыванию вероятностей: $p(x_1) \geq p(x_2) \geq \ldots \geq p(x_M)$.

2. Находим числа m_i, $i = 1, 2, \ldots, M$, исходя из неравенств:

$$\frac{1}{2^{m_i}} \leq p(x_i) < \frac{1}{2^{m_i - 1}}. \tag{11.7}$$

3. Сопоставим каждому сообщению кумулятивную вероятность по правилу

$$q_1 = 0,\ q_2 = p(x_1),\ q_3 = p(x_1) + p(x_2),\ q_M = \sum_{i=1}^{M-1} p(x_i). \tag{11.8}$$

4. Находим первые после запятой m_i знаков в разложении числа q_i в двоичную дробь: $i = 1, 2, \ldots M$. Цифры этого разложения, стоящие после запятой, являются кодовым словом, соответствующим сообщению x_i.

5. Если необходимо, производим операцию усечения.

Лемма 11.3. *Построенный по алгоритму Шеннона код является префиксным.*

Лемма 11.4. *Средняя длина кодовых слов ансамбля X кода, построенного по алгоритму Шеннона не превосходит* $1 + H(X)$.

Код Хаффмена

Предложенный алгоритм был придуман в 1952 г. студентом Дэвидом Хаффменом в процессе выполнения домашнего задания.

Алгоритм Хаффмана. 1. Сообщения ансамбля упорядочиваем по убыванию вероятностей: $p(x_1) \geq p(x_2) \geq \ldots \geq p(x_M)$.

2. Два наименее вероятные сообщения склеиваем в одно и приписываем ему суммарную вероятность склеенных сообщений.

3. Если полученный ансамбль состоит из двух или более сообщений, то переходим к первому пункту, т. е. в полученном ансамбле вероятности упорядочиваем по убыванию. Затем два наименее вероятные сообщения склеиваем в одно и приписываем ему суммарную вероятность склеенных сообщений.

4. Через конечное число шагов получим ансамбль, состоящий из одного сообщения.

5. Двигаясь от заключительного ансамбля к начальному и расставляя кодовые символы 0 и 1 по принципу «вверх — 1», «вниз —0» получаем оптимальный код.

Троичный код Фано

Алгоритм Фано: 1. Сообщения ансамбля $X = \{x_1, \ldots, x_M\}$ упорядочиваем по убыванию вероятностей: $p(x_1) \geq p(x_2) \geq \ldots \geq p(x_M)$.

2. Разбиваем ансамбль X на три подансамбля $X^{(0)}$, $X^{(1)}$ и $X^{(2)}$ с помощью двух пороговых целых чисел $k^{(1)}$ и $k^{(2)}$, таких, что:

$$1 < k^{(1)} < k^{(2)},\ \ k^{(1)} \leq k^{(2)} \leq M - 1,$$

так, чтобы абсолютная величина

$$K = \Big| \sum_{i=1}^{k^{(1)}} p(x_i) - \sum_{i=k^{(1)}+1}^{k^{(2)}} p(x_i) \Big| + \Big| \sum_{i=k^{(2)}+1}^{M} p(x_i) - \sum_{i=k^{(1)}+1}^{k^{(2)}} p(x_i) \Big| + \Big| \sum_{i=1}^{k^{(1)}} p(x_i) - \sum_{i=k^{(2)}+1}^{M} p(x_i) \Big| \tag{11.9}$$

достигала наименьшего возможного значения. Сообщениям подансамбля $X^{(0)}$ приписываем 0, сообщениям подансамбля $X^{(1)}$ приписываем 1, сообщениям подансамбля $X^{(2)}$ приписываем 2.

3. Если подансамбль $X^{(0)}$ состоит более чем из двух сообщений, то разбиваем множество его сообщений на три части $X^{(00)}$, $X^{(01)}$ и $X^{(02)}$, соответственно, с помощью пороговых целых чисел

$$1 < k^{(11)} \leq k^{(12)} < k^{(1)} - 1,$$

так, чтобы абсолютная величина

$$K = \Big| \sum_{i=1}^{k^{(11)}} p(x_i) - \sum_{i=k^{(11)}+1}^{k^{(12)}} p(x_i) \Big| + \Big| \sum_{i=k^{(12)}+1}^{k^{(1)}} p(x_i) - \sum_{i=k^{(11)}+1}^{k^{(12)}} p(x_i) \Big| + \Big| \sum_{i=1}^{k^{(11)}} p(x_i) - \sum_{i=k^{(12)}+1}^{k^{(1)}} p(x_i) \Big|$$

достигала наименьших возможных значений. Сообщениям из ансамбля $X^{(00)}$ приписываем 0, сообщениям из $X^{(01)}$ приписываем 1, а сообщениям из ансамбля $X^{(02)}$ приписываем 2.

Если подансамбль $X^{(ij)}$, $i, j \in \{0, 1, 2\}$ состоит более чем из одного сообщения, то переходим к шагу 4. Если все подансамбли содержат по одному сообщению, то переходим к шагу 5.

4. Если есть подансамбли, состоящие более чем из одного сообщения, то разбиваем каждый из них на три подансамбля, исходя из соотношения, аналогичного (11.9). Сообщениям из подансамблей с нулевыми последними индексами приписываем 0, сообщениям из подансамблей с единичными последними индексами приписываем 1, сообщениям из подансамблей, которые заканчиваются на двойку приписываем 2.

Если все образовавшиеся подансамбли состоят из одного сообщения, то переходим к шагу 5.

Если есть подансамбли, состоящие более чем из одного сообщения, то то повторяем шаг 4.

5. Если образовавшиеся подансамбли состоят из одного сообщения, то последовательно, начиная с первой метки, выписываем нули, единицы и двойки, относящиеся к каждому сообщению ансамбля X.

В итоге получается троичный префиксный код для заданного ансамбля X.

Троичный код Шеннона

Алгоритм Шеннона сводится к последовательному выполнению следующих шагов:

1. Сообщения ансамбля $X = \{x_1, \ldots, x_M\}$ упорядочиваем по убыванию вероятностей: $p(x_1) \geq p(x_2) \geq \ldots \geq p(x_M)$.

2. Находим числа m_i, $i = 1, 2, \ldots M$, исходя из неравенств:

$$\frac{1}{3^{m_i}} \leq p(x_i) < \frac{1}{3^{m_i - 1}}.$$

3. Сопоставим каждому сообщению кумулятивную вероятность по правилу:

$$q_1 = 0,\ q_2 = p(x_1),\ q_3 = p(x_1) + p(x_2),\ q_M = \sum_{i=1}^{M-1} p(x_i).$$

4. Находим первые после запятой m_i, $i = 1, 2, \ldots M$, знаков в разложении числа q_i в троичную дробь. Цифры этого разложения, стоящие после запятой, являются кодовым словом, соответствующим сообщению x_i.

5. Если необходимо, производим операцию усечения.

Троичный код Хаффмена

1. Сообщения ансамбля $X = \{x_1, \ldots, x_M\}$ упорядочиваем по убыванию вероятностей: $p(x_1) \geq p(x_2) \geq \ldots \geq p(x_M)$.

2. Три наименее вероятные сообщения склеиваем в одно и приписываем ему суммарную вероятность склеенных сообщений.

3. Если полученный ансамбль состоит из двух или более сообщений, то переходим к первому пункту, т. е. в полученном ансамбле вероятности упорядочиваем по убыванию. Затем три наименее вероятные сообщения склеиваем в одно и приписываем ему суммарную вероятность склеенных сообщений.

4. Через конечное число шагов получим ансамбль, состоящий из одного сообщения.

5. Двигаясь от заключительного ансамбля к начальному и расставляя кодовые символы 0, 1 и 2 по принципу «вверх — 1», «вниз — 0» и «центр — 2» получаем оптимальный код.