А.Г. Дьячков

Теория информации

§1. Введение

Содержание

1.1 Передача информации по каналу связи

Задачи теории информации (математической теории связи), изучаемые в данном курсе лекций, можно показать на приводимой ниже блок-схеме, которая описывает передачу конечной двоичной последовательности (информационного слова) по двоичному каналу связи:

$$\boxed{\text{Источник}} \overset{U_1^k}{\Longrightarrow} \boxed{\text{Кодер}} \overset{X_1^N = E(U_1^k)}{\Longrightarrow} \boxed{\text{Канал}} \overset{Y_1^N}{\Longrightarrow} \boxed{\text{Декодер}} \overset{\tilde{U_1^k} = D(Y_1^N)}{\Longrightarrow} \boxed{\text{Адресат}},$$

$$\Pr\{U_1^k = \mathbf{u}\} = \mathsf{p}_k(\mathbf{u}), \quad \Pr\{Y_1^N = \mathbf{y} | X_1^N = \mathbf{x}\} = P_N(\mathbf{y}|\mathbf{x}), \quad \Pr\{\tilde{U_1^k} \neq U_1^k\} \le \epsilon,$$

$$\mathbf{u} = (u_1, u_2, \ldots, u_k) \in \{0,1\}^k, \ \mathbf{x} = (x_1, x_2, \ldots, x_N) \in \{0,1\}^N, \ \mathbf{y} = (y_1, y_2, \ldots, y_N) \in \{0,1\}^N.$$

Эта блок-схема означает следующее.

- *Источник* генерирует предназначенную для *адресата* двоичную (из 0 и 1) последовательность $U_1^k = (U_1, U_2, \ldots, U_k)$ длины $k \ge 1$, которая называется *сообщением источника*. На языке теории вероятностей сообщение источника интепретируется как *случайная* (с точки зрения адресата) величина с известным, или неизвестным, распределением вероятностей $\mathsf{p}_k = \mathsf{p}_k(\mathbf{u})$. Значения $\mathbf{u} \in \{0,1\}^k$, для которых вероятности $\mathsf{p}_k(\mathbf{u}) > 0$, будем называть *информационными словами*. Число информационных слов обозначим через M. В частности, сообщение источника может принимать равновероятные значения на множестве $\{0,1\}^k$ двоичных слов длины k, т.е. $M = 2^k$ и $\mathsf{p}_k(\mathbf{u}) = 2^{-k}$ для любого $\mathbf{u} \in \{0,1\}^k$. Зафиксируем некоторую нумерацию информационных слов целыми числами от 1 до M. В дальнейшем, не нарушая общности, мы будем часто рассматривать информационные слова как элементы множества целых чисел от 1 до M.

- Для передачи по *каналу связи* информационное слово поступает на вход *кодера*, который отображает (кодирует) его с помощью *функции кодирования* E в *передаваемое* по каналу связи двоичное *кодовое* слово

$$\mathbf{x} = E(\mathbf{u}), \quad \mathbf{u} \in \{0,1\}^k, \ \mathbf{x} \in \{0,1\}^N,$$

длины $N \geq k$. Совокупность всех M слов образует *код* длины N и объема M. В дальнейшем код

$$X = \|x_n(i)\|, \quad n = 1, 2, \ldots, N, \quad i = 1, 2, \ldots, M,$$

будем записывать как двоичную $(N \times M)$-матрицу, столбец которой

$$\mathbf{x}(i) = (x_1(i), x_2(i), \ldots, x_N(i)) \in \{0, 1\}^N, \quad i = 1, 2, \ldots, M,$$

является кодовым словом для информационного слова с номером i. Кодовые слова можно интерпретировать как значения случайной величины

$$X_1^N = (X_1, X_2, \ldots, X_N) = E(U_1^k),$$

называемой *входным сигналом* длины $N \geq k$. Число N называется *временем работы канала* для передачи сообщения источника, а число $R \triangleq \frac{\log_2 M}{N}$ называется *скоростью* кодирования (передачи) сообщения источника (по каналу связи). Для сообщения с равновероятными значениями скорость $R \triangleq k/N$ определяет число двоичных информационных символов, передаваемых за единицу времени работы канала. Важным частным случаем кодера является *линейный* кодер, задаваемый двоичной $(N \times k)$-матрицей G, линейно независимые столбцы которой образуют базис подпространства размерности $\leq k$ в линейном пространстве $\{0, 1\}^N$ над полем двоичных чисел. При этом функция кодирования E определяется как операция произведения в двоичном поле матрицы G на столбец (информационное слово) $\mathbf{u}$, т.е.

$$\mathbf{x} = E(\mathbf{u}) \quad \Longleftrightarrow \quad \mathbf{x} = G\mathbf{u}.$$

- Канал связи описывается условным распределением, т.е. набором условных вероятностей

  $$P_N = \|P_N(\mathbf{y}|\mathbf{x})\|, \quad \mathbf{x}, \mathbf{y} \in \{0, 1\}^N,$$

  где $P_N(\mathbf{y}|\mathbf{x})$ – условная вероятность получения на выходе канала двоичной последовательности (*выходного сигнала*) $Y_1^N = (Y_1, Y_2, \ldots, Y_N) = \mathbf{y}$, если входной сигнал $X_1^N = \mathbf{x}$. Важным примером канала связи является *двоичный симметричный канал без памяти* (ДСК), в котором вероятность ошибки (искажения) при передаче одного двоичного символа (переходная вероятность) задается параметром p, $0 < p < 1/2$, и ошибки в разные моменты времени передачи происходят независимо, т.е.

  $$P_N(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{N} P_1(y_i|x_i) = p^{\rho(\mathbf{x},\mathbf{y})}(1-p)^{N-\rho(\mathbf{x},\mathbf{y})},$$

  где $P_1(0|1) = P_1(1|0) = p$, $P_1(0|0) = P_1(1|1) = 1 - p$, а $\rho(\mathbf{x}, \mathbf{y})$ – число несовпадающих позиций (расстояние Хэмминга) в словах $\mathbf{x}$ и $\mathbf{y}$. Другим примером канала связи, который мы будем изучать в §**2**, является *канал с равновесным шумом* (КРШ). Условные вероятности этого канала определяются следующим образом

  $$P_N(\mathbf{y}|\mathbf{x}) = \begin{cases} \binom{N}{t}^{-1}, & \text{если } \rho(\mathbf{x}, \mathbf{y}) = t, \\ 0, & \text{если } \rho(\mathbf{x}, \mathbf{y}) \neq t, \end{cases}$$

  где t, $1 \leq t \leq N/2$, – фиксированный целочисленный параметр. Определение КРШ означает, что сигнал на выходе КРШ имеет равномерное распределение на сфере

Хэмминга радиуса t, центр которой является входным сигналом КРШ. Пусть $\lceil a \rceil$ обозначает наименьшее целое число $\geq a$. В дальнейшем будем задавать параметр КРШ в виде $t = \lceil pN \rceil$, где по аналогии с ДСК фиксированное значение p, $0 < p < 1/2$ не меняется при изменении N. В § **3** мы будем рассматривать *комбинаторную* модель двоичного канала связи, которая называется каналом с t ошибками и задается неравенством $\rho(\mathbf{x}, \mathbf{y}) \leq t$. Другими словами, при передаче за время N может произойти не более t ошибок.

- Выходной сигнал Y_1^N канала связи поступает на вход *декодера*, который с помощью *декодирующей функции* D

$$\mathbf{u} = D(\mathbf{y}), \quad \mathbf{u} \in \{0,1\}^k, \; \mathbf{y} \in \{0,1\}^N,$$

преобразует его в *принятую* последовательность (*сообщение адресата*)

$$\tilde{U_1^k} = (\tilde{U}_1, \tilde{U}_2, \ldots, \tilde{U}_k) = D(Y_1^N).$$

В качестве важного примера декодирующей функции введем рассматриваемое в § **2** *декодирование по максимуму правдоподобия* (МП - декодирование), которое будем обозначать символом $D(\mathbf{y})$. Это декодирование зависит от функции кодирования $\mathbf{x} = E(\mathbf{u})$ и определяется следующим образом. *При заданном* $\mathbf{y}$ *положим* $D(\mathbf{y}) = \mathbf{u}$*, где информационное слово* $\mathbf{u}$ *имеет наименьший номер среди всех информационных слов* $\mathbf{u}'$*, для которых достигается*

$$\max_{\mathbf{u}'=1,2,\ldots,M} P_N(\mathbf{y}|E(\mathbf{u}')) = P_N(\mathbf{y}|E(\mathbf{u})).$$

- Пара кодирование–декодирование (E, D) называется *методом передачи* сообщения источника по каналу P_N. В качестве критерия метода передачи (E, D) рассматривается *вероятность несовпадения* сообщения источника с сообщением адресата, т.е. число

$$Q(\mathsf{p}_k, P_N, E, D) = \Pr\{\tilde{U_1^k} \neq U_1^k\},$$

называемое *вероятностью ошибки*. Отметим доказываемое в §**2** важное свойство *оптимальности* МП-декодирования, которое состоит в следующем. *Если сообщение источника имеет равномерное распределение, то для любой функции кодирования вероятность ошибки минимизируется при МП-декодировании.* Проблема кодирования для передачи информации по каналу связи формулируется следующим образом. *Пусть заданы распределение вероятностей сообщения источника* p_k*, канал связи* P_N *и допустимое значение вероятности ошибки* $\epsilon > 0$*. Требуется построить метод передачи* (E, D)*, вероятность ошибки которого* $Q(\mathsf{p}_k, P_N, E, D) \leq \epsilon$*.*

Пусть $N \to \infty$, $M \triangleq \lceil 2^{NR} \rceil$ а скорость передачи $R > 0$ – фиксирована. Приведем полученные К.Шенноном необходимые (обратная теорема Шеннона) и достаточные (прямая теорема Шеннона) асимптотические условия существования метода передачи с вероятностью ошибки, стремящейся к нулю. Для фиксированной скорости передачи R и канала связи P_N введем оптимальную (минимаксную) вероятность ошибки:

$$\mathcal{E}(P_N, R) = \min_{(E,D)} \max_{\mathsf{p}_k} Q(\mathsf{p}_k, P_N, E, D).$$

Теорема Шеннона. *Существует определяемое каналом связи* P_N *число* C*, называемое пропускной способностью канала, для которого справедливы следующие два утверждения*:

1. Обратная теорема. *Если $R > C$, то*

$$\varlimsup_{N\to\infty} \mathcal{E}(P_N, R) \geq 1 - \frac{C}{R},$$

т.е. при $R > C$ и $N \to \infty$ не существует метода передачи с вероятностью ошибки, стремящейся к нулю.

2. Прямая теорема. *Если $R < C$, то*

$$\lim_{N\to\infty} \mathcal{E}(P_N, R) = 0,$$

т.е. при $R < C$ и $N \to \infty$ существует метод передачи с вероятностью ошибки, стремящейся к нулю.

Обратная теорема будет доказана в § **5**. В основе вывода этой теоремы лежит предложенное К.Шенноном фундаментальное понятие *информации пары случайных величин.*

В случае канала с равновесным шумом (КРШ) прямая теорема Шеннона получена в § **2**. Для дискретного канала без памяти (ДКБП), являющегося обобщением ДСК на произвольные конечные алфавиты, эта теорема доказывается в § **6**. В методе доказательства, который называется *методом случайного кодирования*, вводится *ансамбль* функций кодирования E (или кодов X) с некоторым заданным на нем распределением вероятностей. Например, ансамбль линейных функций кодирования представляет собой множество, состоящее из 2^{Nk} равновероятных двоичных $(N\times k)$-матриц G. Это означает, что все Nk элементов случайной матрицы G являются независимыми одинаково распределенными двоичными случайными величинами, принимающими значения 0 и 1 с вероятностями 1/2.

Пусть на выходе канала используется декодирование по максимуму правдоподобия. Тогда вероятность ошибки кода, как функция на ансамбле, является случайной величиной. Рассматривается усреднение (математическое ожидание) вероятности ошибки по ансамблю кодов, для которого удается построить верхнюю оценку, стремящуюся к нулю, если скорость передачи $R < C$. Отсюда следует существование функции кодирования E (или кода X), для которой при МП-декодировании вероятность ошибки стремится к нулю при $R < C$. Для примеров ДСК и КРШ, задаваемых параметром p, $0 < p < 1/2$, пропускная способность C вычисляется (см. §**5**) по формуле

$$C = 1 - h(p), \quad h(p) = -p\log_2 p - (1-p)\log_2(1-p).$$

Функция $h(p)$ называется *двоичной энтропией.*

1.2 Математическая модель задачи поиска

Рассмотрим следующую математическую модель логической 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 $S = \{a_{i_1}, a_{i_2}, \ldots, a_{i_m}\}$, $1 \leq i_1 < i_2 < \ldots < i_m \leq M$, которое *надо найти*. Элементы подмножества S называются *дефектными* элементами (или *значимыми* факторами).

Обычно предполагается, что неизвестное подмножество значимых факторов принадлежит некоторому *известному* классу допустимых подмножеств.

Для поиска $S \subseteq A$ разрешается провести серию из N экспериментов (*групповых проверок*), в каждом из которых можно выбрать некоторое подмножество (*группу*) $T \subseteq A$ и выяснить: *содержит или нет* тестируемая группа T *хотя бы один* дефектный элемент. Обозначая символом $\varnothing$ пустое множество, можно написать, что двоичный результат проверки $y \in \{0;1\}$ вычисляется по следующему правилу:

$$y = \begin{cases} 1, \text{ если } S \bigcap T \neq \varnothing, \\ 0, \text{ если } S \bigcap T = \varnothing. \end{cases}$$

Данную модель вычисления результата проверки при групповом тестировании 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 групповых проверок. На основании итога проверок $\mathbf{y} = (y_1, y_2, \ldots, y_N)$, где

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

экспериментатор должен *однозначно восстановить* (найти) неизвестное подмножество значимых факторов 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. Важный пример введенных А.Реньи статических стратегий поиска одного дефектного элемента в конечном множестве рассматривается в § **4**. Основные результаты по исследованию модели А.Реньи приведены в конце данного вводного раздела. Некоторые задачи последовательного поиска и связанные с ними теоретико-информационные проблемы изучаются в § **7**.

Набор N статических проверок удобно описывать с помощью двоичной $(N \times M)$-матрицы

$$X = \|x_n(i)\|, \quad x_n(i) \in \{0;1\}, \quad n = 1, 2, \ldots, N, \quad i = 1, 2, \ldots, M,$$

называемой статическим *планом* поиска длины N в конечном множестве объЯма M. По определению, *номера единичных позиций n-ой строки плана X являются номерами факторов, включенных в n-ую проверку*, т.е. элемент плана

$$x_n(i) \triangleq \begin{cases} 1, & \text{если } a_i \in T_n, \\ 0, & \text{если } a_i \notin T_n. \end{cases}$$

Пусть

$$\mathbf{x}_n = (x_n(1), x_n(2), \ldots, x_n(M)) \in \{0;1\}^M, \quad n = 1, 2, \ldots, N,$$

обозначает n-ую строку плана, которая идентифицирует n-ую проверку, а столбец

$$\mathbf{x}(i) = (x_1(i), x_2(i), \ldots, x_N(i)) \in \{0;1\}^N, \quad i = 1, 2, \ldots, M,$$

обозначает i-ый столбец плана X.

Рассмотрим случай, когда дефектное множество S состоит из одного значимого фактора a_i. Тогда, как легко заметить, совокупность N результатов проверок представляет собой i-ый столбец плана X, т.е. $\mathbf{y} = \mathbf{x}(i)$. Следовательно, статический план X обладает свойством однозначного восстановления тогда и только тогда, когда все столбцы $\mathbf{x}(i)$ плана X отличны друг от друга. Столбец $\mathbf{x}(i)$ можно назвать *кодовым словом*, в которое план X *кодирует фактор* a_i. Это означает, что введенное в предыдущем разделе понятие двоичного кода X для кодирования сообщения источника, принимающего M значений, эквивалентно понятию статического плана X для поиска одного значимого фактора, который отождествляется с сообщением источника. Очевидно также, что модель поиска одного значимого фактора из M-множества факторов с помощью групповых проверок эквивалентна модели передачи сообщения источника, принимающего значения в M-множестве, по каналу без шума.

1.3 Модель поиска А.Реньи

В этой модели рассматриваются статические планы поиска одного значимого фактора из M-множества факторов при наличии ограничений на объемы тестируемых групп. Введем параметр p, $0 < p \leq 1/2$, который задает ограничение на число единиц (вес) строк плана X:

$$\sum_{i=1}^{M} x_n(i) \leq k = \lceil pM \rceil, \quad n = 1, 2, \ldots, N.$$

Данное ограничение означает, что в каждой из N статических проверок может тестироваться не более $k = \lceil pM \rceil$ факторов.

Обозначим через $N(M, k) = N(M, \lceil pM \rceil)$ минимальное значение параметра N, для которого план поиска X обладает свойством однозначного восстановления, т.е. столбцы X отличны друг от друга. В § **4** будет доказана приводимая ниже теорема Шеннона, устанавливающая асимптотику минимальной длины статического плана для модели поиска А. Реньи.

Теорема Шеннона. *Если $M \to \infty$, то справедливо асимптотическое равенство*

$$N(M, \lceil pM \rceil) = \frac{\log_2 M}{h(p)}(1 + o(1)),$$

где символ $h(p)$ обозначает введенную выше двоичную энтропию.