СПбГЭТУ «ЛЭТИ»
ПЕРВЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ

А.Б. Левина

Теория информации и теория кодирования

Основы теории кодирования

СПбГЭТУ «ЛЭТИ», 2022 г.

1 ВВЕДЕНИЕ

На данной лекции будут изучены основы теории кодирования, историю развития данного направления, применимость данного направления на практике, математические основы теории кодирования, ввести понятие ошибки, понятие корректирующих кодов и кодов обнаруживающих ошибку.

2. КАНАЛЬНОЕ КАДИРОВАНИЕ

На первой лекции мы рассмотривали блок схему работы системы связи:

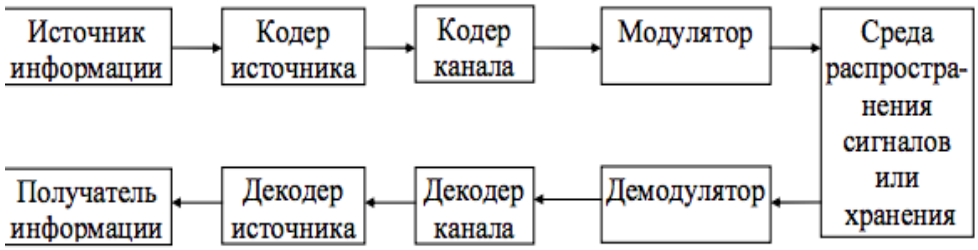


Рис. 1: Блок-схема системы связи.

Вспомним еще раз о каждом объект, представленном на схеме:

1. *Источник* – любой объект порождающий сообщение, которое необходимо передать в пространстве или во времени
2. *Кодер источника* – представление информации в более компактной форме
3. *Кодер канала* – обработка информации с целью защиты при передаче
4. *Модулятор* – преобразование в сигналы, согласованные с природой канала средой накопителя

Можно выделить несколько разделов теории информации:

- *Кодирование дискретных источников* – кодирование без потерь, кодирование для каналов без шума, сжатие
- *Кодирование информации для передачи по каналу с шумом* - защита информации от помех в каналах связи
- *Кодирование с заданным критерием качества* - методы кодирования, обеспечивающие наилучший компромисс между качеством и затратами на передачу информации
- *Кодирование информации для систем со многими пользователями* - оптимальное взаимодействие абонентов, использующих общий ресурс

– *Секретная связь, системы защиты информации от несанкционированного доступа*

Упоминалось, что в курсе будет рассматривать два направления:

1. ***Кодирование источника***

   Когда говорят о кодирование источника подразумевают сжатие данных.

   Сжатие данных (*data compression*) — алгоритмическое преобразование данных, производимое с целью уменьшения занимаемого ими объёма. Применяется для более рационального использования устройств хранения и передачи данных.

   Синонимы: *упаковка данных*, *компрессия*, *сжимающее кодирование*, *кодирование источника*.

   Обратная процедура называется восстановлением данных (распаковкой, декомпрессией).

2. ***Канальное (помехоустойчивое) кодирование***

   У него есть две цели:

   Обнаружение ошибок обеспечивается с помощью кодов, обнаруживающих ошибки

   Исправление ошибок обеспечивается корректирующими кодами (коды, исправляющие ошибки, коды с коррекцией ошибок, помехоустойчивые коды).

Кодирование источника мы подробно изучили на второй лекции, и сейчас переходим к канальному кодированию.

Обнаружение ошибок *в технике связи*— действие, направленное на контроль целостности данных при записи/воспроизведении информации или при её передаче по линиям связи.

2. Исправление ошибок (**коррекция ошибок**) — процедура восстановления информации после чтения её из устройства хранения или канала связи.

Для обнаружения ошибок используют **коды обнаружения ошибок**, для исправления — **корректирующие коды** (**коды, исправляющие ошибки**, **коды с коррекцией ошибок**, **помехоустойчивые коды**).

С ними мы сейчас и начнем наше знакомство.

В системах связи возможны несколько стратегий борьбы с ошибками:

- обнаружение ошибок в блоках данных и *автоматический запрос повторной*

передачи повреждённых блоков — этот подход применяется, в основном, на канальном и транспортном уровнях;

- обнаружение ошибок в блоках данных и отбрасывание повреждённых блоков — такой подход иногда применяется в системах потокового мультимедиа, где важна задержка передачи и нет времени на повторную передачу;
- *исправление ошибок* (*forward error correction*) применяется на физическом уровне.

3. КОДЫ ОБНАРУЖЕНИЯ И ИСПРАВЛЕНИЯ ОШИБОК

Определение: *Корректирующие коды* — коды, служащие для обнаружения или исправления ошибок, возникающих при передаче информации под влиянием помех, а также при её хранении.

Определение: Коды обнаружения ошибок - могут только установить факт наличия ошибки в переданных данных, но не исправить её.

Блочный код — в информатике тип канального кодирования, увеличивающий избыточность сообщения так, чтобы в приёмнике можно было расшифровать его с минимальной (теоретически нулевой) погрешностью, при условии, что скорость передачи информации (количество передаваемой информации в битах в секунду) не превысила бы канальную производительность.

Система блочного кодирования получает на входе *k*-значное кодовое слово *W*, и преобразовывает его в *n*-значное кодовое слово *C(W)*.

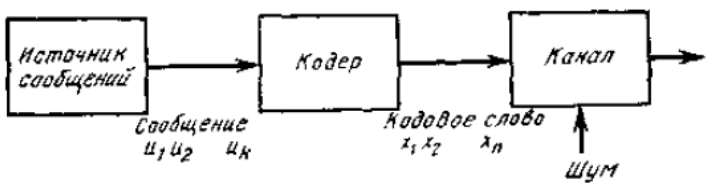
Это кодовое слово и называется **блоком**.

В области математики и теории информации **линейный код** — это важный тип блокового кода, использующийся в схемах определения и коррекции ошибок.

Линейные коды:

«+» по сравнению с другими кодами, позволяют реализовывать более эффективные алгоритмы кодирования и декодирования информации;

«-» линейность легко взламывается.



Основные понятия:

- Сообщение (k символов)
- Проверочные символы (n-k символов)
- Кодовое слово (n символов)
- Код

$$\mathbf{H}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \mathbf{H}\mathbf{x}^T = \mathbf{0}, \qquad (1.1)$$

где $((n-k)\times n)$ — матрица $\mathbf{H}$, называемая *проверочной матрицей кода*, имеет вид

$$\mathbf{H} = [\mathbf{A} \mid \mathbf{I}_{n-k}]. \qquad (1.2)$$

Здесь $\mathbf{A}$ — некоторая фиксированная $((n-k)\times k)$-матрица из 0 и 1, а

$$\mathbf{I}_{n-k} = \begin{pmatrix} 1 & & 0 \\ & 1 & \\ 0 & & \ddots\ 1 \end{pmatrix}$$

— единичная матрица размера $(n-k)\times(n-k)$. Все операции в уравнении (1.1) выполняются *по модулю 2*, т. е. $0+1=1$; $1+1=0$; $-1=+1$. Мы будем называть это *двоичной арифметикой*.

Пусть H – любая двоичная матрица, *линейный код с проверочной матрицей H* состоит из всех векторов x таких, что $Hx^T=0$.

Если есть сообщение, то как найти соответствующие кодовое слово?

Для этого используется порождающая матрица:

$$\mathbf{G} = [\mathbf{I}_k \mid -\mathbf{A}^T].$$

Параметры кода:

1. Длина: говорят что кодовое слова $x_1,\ldots x_n$ имеет длину n
2. Размерность: Если H имеет n-k линейно независимых строк, то имеется 2^k кодовых слов, это называется размерностью кода
3. Код называется [n, k] – кодом
4. Код использует n символов для передачи k символов то эффективностью/скоростью кода является R=k/n
5. Если x и y – кодовые слова данного кода, то x+y тоже кодовое слова этого кода, если c – любой элемент поля, то cx- тоже кодовое слово

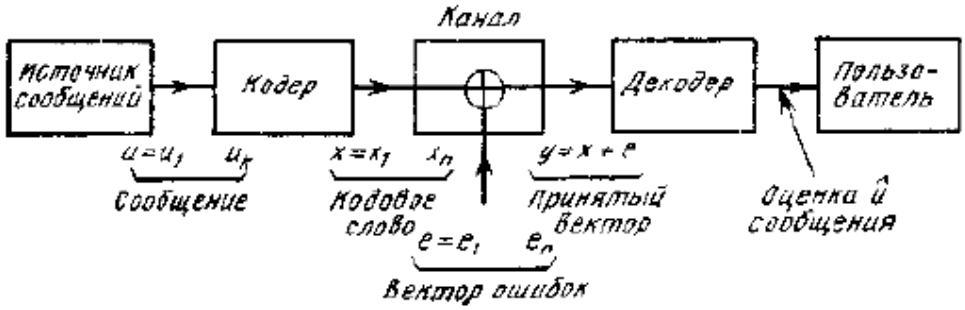
6. Минимальное расстояние кода $d = \min\{dist(u,v)\} = \min\{wt(u-v)\}, v \in C, u \in C, v \neq u$

Определение: Линейный код длины n, размерности k с минимальными расстоянием d называется [n, k, d] -кодом.

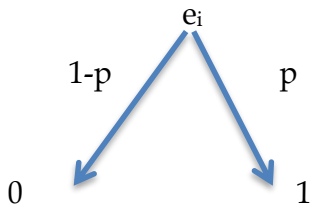
Теорема 1: Минимальное расстояние линейного кода равно минимальному весу ненулевых слов.

Теорема 2: Код с минимальным расстоянием d может исправлять наибольшее целое число от (d-1/2) ошибок.

Рассмотрим декодирование линейного кода.



Вектор ошибки: $e=y-x=e_1 \ldots e_n$



Стратегия декодера - выбор наиболее вероятного для принятого y вектора ошибки e. Что бы это обеспечить производится *Декодирование по максимуму правдоподобия*.

Prob {e=00000}=$(1-p)^5$

Prob {e=01000}=$p(1-p)^4$

Prob {e=10010}=$p^2(1-p)^3$

V- вектор ошибки веса a, то Prob {e=V}=$p^a(1-p)^{n-a}$ $p<1/2$, то $1-p>p$

Стратегия декодера – y декодируется в ближайшее кодовое слово x, т.е. выбирается вектор ошибок e с наименьшим весом. Тогда происходит *декодирование в ближайшее кодовое слово/полное декодирование*.

При декодировании возможна *ошибка декодирования* –декодер выдает неправильное слово.

Тогда происходит один из трех сценариев:

- *Полное декодирование*

  Декодер ищет ближайшее по весу кодовое слово

- *Неполное декодирование*

  Произошло не более чем l ошибок – исправляем, в противном случае нет

- *Обнаружение ошибок*

  Только проверка на наличие ошибки

Введем понятие смежного класса – это поможет нам для проведения декодирования.

Определение смежного класса. Пусть $\mathscr{C}$ является $[n, k]$ линейным кодом над полем из q элементов. Для любого вектора $\mathbf{a}$ множество

$$\mathbf{a}+\mathscr{C}=\{\mathbf{a}+\mathbf{x}\cdot\mathbf{x}\in\mathscr{C}\}$$

называется *смежным классом* (или *сдвигом*) кода $\mathscr{C}$. Любой вектор $\mathbf{b}$ находится в некотором смежном классе (например, в $\mathbf{b}+\mathscr{C}$). Два вектора $\mathbf{a}$ и $\mathbf{b}$ лежат в одном и том же смежном классе, тогда и только тогда, когда $(\mathbf{a}-\mathbf{b})\in\mathscr{C}$. Каждый смежный класс содержит q^k векторов.

Вектор из смежного класса имеющий минимальный вес, называется *лидером смежного класса*.

Стандартное расположение. Полезным способом описания работы декодера является таблица, названная *стандартным расположением* для кода. Первая строка ее состоит из самого кода с нулевым кодовым словом с левой стороны: $\mathbf{x}^{(1)}=\mathbf{0}; \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(s)}$ $(s=q^k)$; другими строками являются другие смежные классы $\mathbf{a}_i+\mathscr{C}$, расположенные в том же самом порядке и с лидерами этих классов с левой стороны:

$$\mathbf{a}_i+\mathbf{x}^{(1)},\ \mathbf{a}_i+\mathbf{x}^{(2)},\ \cdots,\ \mathbf{a}_i+\mathbf{x}^{(s)}.$$

Синдром. Имеется простой способ, как определить, в каком смежном классе находится $\mathbf{y}$: надо вычислить вектор $\mathbf{S}=\mathbf{H}\mathbf{y}^T$, который называется *синдромом* $\mathbf{y}$.

Свойства синдрома. (1). $\mathbf{S}$ представляет собой вектор-столбец длины $n-k$.

(2). Синдром вектора $\mathbf{y}$, $\mathbf{S}=\mathbf{H}\mathbf{y}^T$, равен нулю, если и только если $\mathbf{y}$ — кодовое слово (по определению кода). Поэтому, если никаких ошибок не произошло, синдром вектора $\mathbf{y}$ равен нулю (но не обратное). В общем случае, если $\mathbf{y}=\mathbf{x}+\mathbf{e}$, где $\mathbf{x}\in\mathscr{C}$, то

$$\mathbf{S}=\mathbf{H}\mathbf{y}^T=\mathbf{H}\mathbf{x}^T+\mathbf{H}\mathbf{e}^T=\mathbf{H}\mathbf{e}^T. \qquad (1.22)$$

(3). Если для двоичного кода имеются ошибки на позициях с номерами $a, b, c, \ldots$, так что $\mathbf{e}=0\ldots\underset{a}{1}\ldots\underset{b}{1}\ldots\underset{c}{1}\ldots0$, то из (1.22) получаем, что

$$\mathbf{S}=\sum_i e_i\mathbf{H}_i=\mathbf{H}_a+\mathbf{H}_b+\mathbf{H}_c+\cdots$$

где $\mathbf{H}_i$ — i-й столбец $\mathbf{H}$.

Теорема 3: Для двоичного кода синдром равен сумме тех столбцов H, где произошли ошибки.

Определение. *Вероятностью ошибки, или вероятностью ошибки на слово,* $P_{ош}$ для данной схемы декодирования называется вероятность появления неправильного кодового слова на выходе декодера.

Если имеется M кодовых слов $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}$, которые, как мы предполагаем, используются с равной вероятностью, то

$$P_{ош} = \frac{1}{M} \sum_{i=1}^{M} \text{Prob} \left\{ \text{выход декодера} \neq \mathbf{x}^{(i)} \mid \mathbf{x}^{(i)} \text{ было послано} \right\}. \quad (1.23)$$

Определение. Предположим, что код содержит M кодовых слов $\mathbf{x}^{(i)} = x_1^{(i)} \ldots x_n^{(i)}$, $i = 1, \ldots, M$, и первые k символов $x_1^{(i)} \ldots x_k^{(i)}$ в каждом кодовом слове являются информационными символами. Пусть $\hat{\mathbf{x}} = \hat{x}_1 \ldots \hat{x}_n$ — символы на выходе декодера. Тогда *вероятность ошибки на символ* $P_{симв}$ определяется как средняя вероятность того, что после декодирования информационный символ является ошибочным:

$$P_{симв} = \frac{1}{kM} \sum_{j=1}^{k} \sum_{i=1}^{M} \text{Prob} \left\{ \hat{x}_j \neq x_j^{(i)} \mid \mathbf{x}^{(i)} \text{ было послано} \right\}. \quad (1.30)$$

Пропускная способность двоичного симметричного канала с вероятностью ошибки p равна $C(p) = 1 + p \log_2 p + (1-p) \log_2 (1-p)$.

Теорема Шеннона:

Для любого ε>0 если R<C(p) и n достаточно велико, то существует [n,k] двоичный код со скоростью k/n≥R вероятность ошибки которого $P_{ош} \leq \varepsilon$

Введем также понятие нелинейного кода:

Определение: Назовем (n, M, d) кодом множества из M векторов длины n такое, что любые два вектора различаются по меньшей мере в d позициях и d является наибольшим числом.

4. ЭНТРОПИЯ

Вспомним еще раз понятие энтропии, которое мы вводили на первой лекции, так как оно понадобится нам при работе с кодами.

Определение: Энтропией дискретного ансамбля X= {x, p(x)} называется

$$H(X) = \mathrm{M}\left[-\log p(x)\right] = -\sum_{x \in X} p(x) \log p(x)$$

Энтропия – количественная мера неопределенности о том, какое из сообщений будет порождено источником.

Свойства энтропии

1. $H(X) \geq 0.$

2. $H(X) \leq \log |X|$ равенство возможно, когда все элементы X равновероятны.

$$\begin{aligned} H(X) - \log |X| &\overset{(a)}{=} -\sum_{x \in X} p(x) \log p(x) - \sum_{x \in X} p(x) \log |X| = \\ &\overset{(b)}{=} \sum_{x \in X} p(x) \log \frac{1}{p(x)|X|} \leq \\ &\overset{(c)}{\leq} \log e \left[\sum_{x \in X} p(x) \left(\frac{1}{p(x)|X|} - 1 \right) \right] = \\ &= \log e \left(\sum_{x \in X} \frac{1}{|X|} - \sum_{x \in X} p(x) \right) = 0 \quad . \end{aligned}$$

3. X = {x; p(x)} и Y = {y = f (x), p(y)}, тогда H(Y) ≤H(X), равенство выполняется если f(x) отличается от p(x) только порядком следования элементов

4. Если X и Y независимы, то H(XY)=H(X)+H(Y)

5. Энтропия - выпуклая функция распределения вероятностей на элементах X

6. Пусть X={x, p(x)} и A входит в X введем X’ ={x, p’(x)} p’(x):

$$p'(x) = \begin{cases} \frac{P(A)}{|A|}, x \in A, \\ p(x), x \notin A. \end{cases}$$

тогда H(X’) ≥ H(X)

7. Задан X и на множестве его элементов определена функция g(x), введем

Y={y=g(x)}. Тогда H(Y) ≤ H(X). Равенство имеет место когда g(x) обратима.

5. ЗАКЛЮЧЕНИЕ

На данной лекции мы познакомились с математическими основами теории кодирования, ввели понятие ошибки, понятие корректирующих кодов и кодов обнаруживающих ошибку.