Теория кодирования

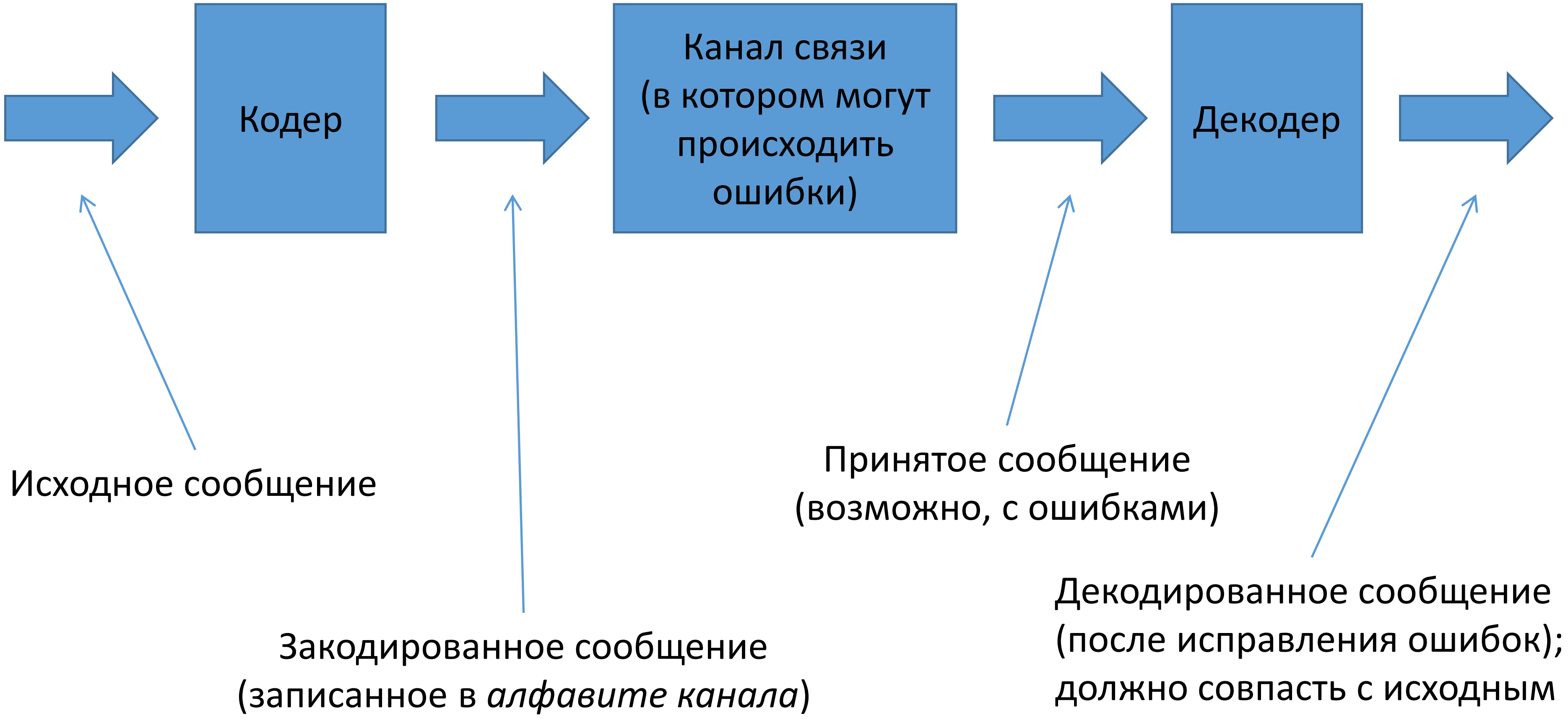
МФТИ, осень 2013

Александр Дайняк

www.dainiak.com

Коды, исправляющие ошибки

Основная модель канала связи:



Типы ошибок

- Ошибки замещения: муха → мука
  - Симметричные
  - Несимметричные
- Ошибки стирания: муха → му?а
- Ошибки выпадения: муха → уха
- Ошибки вставки: мука → мурка
- Комбинации перечисленных типов

Типы ошибок

Всегда задаются ограничения на «ненадёжность» канала, например:

- верхняя оценка числа ошибок на одно сообщение (детерминированные ограничения)
- вероятность возникновения ошибки на один символ сообщения (вероятностные ограничения)

Чаще всего алфавит канала двоичный: $\{0,1\}$

Коды

Пусть $\mathbb{A}_q$ — алфавит канала, $|\mathbb{A}_q| = q$.

q-*ичным кодом* называется любое подмножество

$$C \subseteq \mathbb{A}_q^n$$

n — *длина кода* (длина кодовых слов)

$|C|$ — *мощность кода* (число кодовых слов)

Чаще всего рассматривают двоичные коды, т.е. когда $q = 2$ и $\mathbb{A}_q = \{0,1\}$.

Для произвольного двоичного слова $\boldsymbol{a}$ будем через $\|\boldsymbol{a}\|$ обозначать *вес слова*, т.е. величину

$$\#\{i \mid a_i \neq 0\}$$

Обнаружение/исправление ошибок

Пусть $\boldsymbol{a}$ и $\boldsymbol{b}$ — слова в алфавите канала.

Обозначим через $\tilde{d}(\boldsymbol{a}, \boldsymbol{b})$ минимальное число ошибок, в результате которых $\boldsymbol{a}$ может перейти в $\boldsymbol{b}$.

Способ кодирования позволяет *обнаруживать* k *ошибок*, если для любых различных кодовых сообщений $\boldsymbol{a}'$ и $\boldsymbol{a}''$
при передаче в канал $\boldsymbol{a}'$ на выходе не может получиться $\boldsymbol{a}''$
(если в канале произошло не более k ошибок).

Иначе говоря, $\tilde{d}(\boldsymbol{a}', \boldsymbol{a}'') > k$.

Обнаружение/исправление ошибок

Способ кодирования позволяет *исправлять k ошибок*, если при передаче в канал различных кодовых сообщений $\boldsymbol{a}'$ и $\boldsymbol{a}''$ на выходе из канала будут получаться различные сообщения (при условии, что с каждым отдельным сообщением в канале происходит не более k ошибок).

Формально:

$$\nexists \boldsymbol{a}', \boldsymbol{a}'' \in C, \boldsymbol{a}: \quad \left(\boldsymbol{a}' \neq \boldsymbol{a}'' \quad \wedge \quad \tilde{d}(\boldsymbol{a}', \boldsymbol{a}) \leq k \quad \wedge \quad \tilde{d}(\boldsymbol{a}'', \boldsymbol{a}) \leq k\right)$$

Метрика

Особенно удобно, когда $\tilde{d}$ является *метрикой*:

- $\forall \boldsymbol{a}, \boldsymbol{b} \quad \tilde{d}(\boldsymbol{a}, \boldsymbol{b}) = \tilde{d}(\boldsymbol{b}, \boldsymbol{a})$
- $\forall \boldsymbol{a} \neq \boldsymbol{b} \quad \tilde{d}(\boldsymbol{a}, \boldsymbol{b}) > 0$
- $\forall \boldsymbol{a} \quad \tilde{d}(\boldsymbol{a}, \boldsymbol{a}) = 0$
- $\forall \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \quad \tilde{d}(\boldsymbol{a}, \boldsymbol{b}) \leq \tilde{d}(\boldsymbol{a}, \boldsymbol{c}) + \tilde{d}(\boldsymbol{c}, \boldsymbol{b})$

Так бывает не всегда. Например, если в канале есть только ошибки вставки и никаких других, то при $\boldsymbol{a} \neq \boldsymbol{b}$ по крайней мере одна из двух величин $\tilde{d}(\boldsymbol{a}, \boldsymbol{b}), \tilde{d}(\boldsymbol{b}, \boldsymbol{a})$ вовсе не определена.

Метрика Хемминга

Если рассматриваются слова одной и той же длины, а в канале возможны только ошибки типа замещения (любые), то $\tilde{d}(\boldsymbol{a}, \boldsymbol{b}) = d_{\mathrm{X}}(\boldsymbol{a}, \boldsymbol{b})$, где

$$d_{\mathrm{X}}(\boldsymbol{a}, \boldsymbol{b}) \coloneqq \#\{i \mid a_i \neq b_i\}$$

Функционал d_{X} — *метрика Хемминга*,

$d_{\mathrm{X}}(\boldsymbol{a}, \boldsymbol{b})$ — *расстояние Хемминга между* $\boldsymbol{a}$ и $\boldsymbol{b}$

Метрика Левенштейна

Если в канале происходят ошибки выпадения/вставки,
то канал описывается *метрикой Левенштейна*:

$$d_{\text{Л}}(\boldsymbol{a}, \boldsymbol{b}) \coloneqq \min \# \text{ выпадений и вставок, переводящих } \boldsymbol{a} \text{ в } \boldsymbol{b}$$

Например:

- $d_{\text{Л}}(«aba», «aa») = 1$
- $d_{\text{Л}}(«abbaba», «abaab») = 3$

Кодовое расстояние

Пусть $\tilde{d}(\cdot,\cdot)$ — метрика и C — код.

Кодовым расстоянием кода C называется величина

$$\tilde{d}(C) \coloneqq \min_{\substack{\boldsymbol{a} \neq \boldsymbol{b} \\ \boldsymbol{a}, \boldsymbol{b} \in C}} \tilde{d}(\boldsymbol{a}, \boldsymbol{b})$$

Кодовое расстояние определяет устойчивость к ошибкам:

- C обнаруживает t ошибок $\Leftrightarrow$ $\tilde{d}(C) > t$
- C исправляет t ошибок $\Leftrightarrow$ $\tilde{d}(C) > 2t$

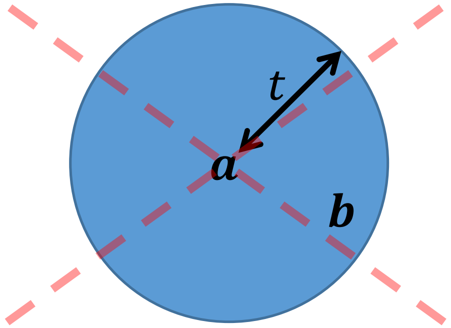
Геометрическая интерпретация

Шар радиуса r *с центром в* $\boldsymbol{a}$ — это множество

$$S_r(\boldsymbol{a}) \coloneqq \{\boldsymbol{b} \mid \tilde{d}(\boldsymbol{a}, \boldsymbol{b}) \le r\}$$

Если в канал передавалось $\boldsymbol{a}$, то на выходе из канала может быть любое слово $\boldsymbol{b} \in S_t(\boldsymbol{a})$.

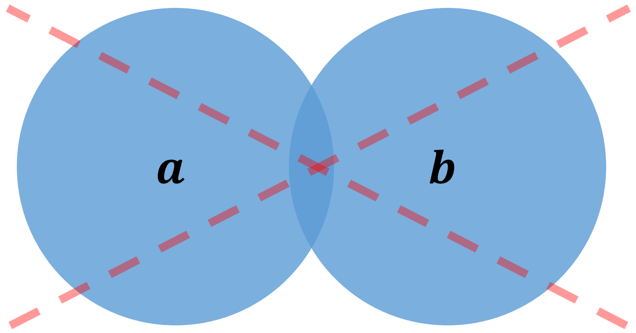
Значит, код обнаруживает t ошибок т. и т.т., когда никакое кодовое слово не попадает в шар радиуса t с центром в другом кодовом слове:



Геометрическая интерпретация

Код исправляет t ошибок т. и т.т., когда при передаче в канал различных кодовых слов на выходе получаются различные слова,

то есть когда шары радиуса t с центрами в кодовых словах не пересекаются:



Основные задачи теории кодов, исправляющих ошибки

Основная задача: строить коды, для которых

- число кодовых слов как можно больше,
- кодовое расстояние как можно больше,
- длина кодовых слов как можно меньше.

Задачи, связанные с ресурсами:

- Процессы кодирования и декодирования (исправление ошибок) должны быть возможно менее трудоёмкими по количеству операций и по памяти

Основные задачи теории кодов, исправляющих ошибки

Основная задача: строить коды, для которых

- число кодовых слов как можно больше,
- кодовое расстояние как можно больше,
- длина кодовых слов как можно меньше.

Геометрически, это **задача об упаковке**

- возможно большего числа шаров,
- возможно большего радиуса,
- в пространстве возможно меньшей размерности.

Коды Варшамова—Тененгольца

- Пример кодов, исправляющих ошибки выпадения/вставки
- Простой алгоритм исправления ошибок

Коды Варшамова—Тененгольца

Код Варшамова—Тененгольца длины n:

$$C \coloneqq \left\{ a_1 a_2 \dots a_n \mid \sum_{i=1}^{n} i a_i \equiv 0 \pmod{(n+1)} \right\}$$

Для мощности кода справедлива формула (без доказательства)

$$|C| = \frac{1}{2(n+1)} \sum_{\substack{d \mid (n+1) \\ d \text{ нечётно}}} \phi(d) 2^{(n+1)/d}$$

Асимптотически это максимально возможная мощность кода, исправляющего одну ошибку выпадения/вставки символа.

Коды Варшамова—Тененгольца: исправление одной ошибки выпадения

Пусть C — код В.—Т. длины n, и пусть $\boldsymbol{a} \in C$.

Пусть в канал передали $\boldsymbol{a} = a_1 \dots a_n$, и на выходе получили слово

$$\boldsymbol{a}' \coloneqq a_1' \dots a_{n-1}' = a_1 \dots a_{k-1} a_{k+1} \dots a_n$$

(символ a_k выпал).

Наша задача: по $\boldsymbol{a}'$ восстановить $\boldsymbol{a}$.

Коды Варшамова—Тененгольца: исправление одной ошибки выпадения

Восстановить $\boldsymbol{a}$ — не то же самое, что восстановить пару (k, a_k).

Например, если $\boldsymbol{a}' = 1001,$ то $\boldsymbol{a} = 10001$, но мы не узнаем, какой именно из нулей выпал.

Положим

$$n_0 \coloneqq \#\{\, i > k \mid a_i = 0 \,\}$$
$$n_1 \coloneqq \#\{\, i > k \mid a_i = 1 \,\}$$

Заметим, что если $a_k = 0$, то $\boldsymbol{a}$ можно восстановить по $\boldsymbol{a}'$, если известно n_1.

Аналогично, если $a_k = 1$, то $\boldsymbol{a}$ можно восстановить по $\boldsymbol{a}'$, если известно n_0.

Коды Варшамова—Тененгольца: исправление одной ошибки выпадения

Рассмотрим суммы

$$S \coloneqq \sum_{i=1}^{n} ia_i \quad \text{и} \quad S' \coloneqq \sum_{i=1}^{n-1} ia_i'$$

Заметим, что

$$S - S' = \sum_{i=1}^{n} ia_i - \left(\sum_{i=1}^{k-1} ia_i + \sum_{i=k}^{n-1} ia_{i+1} \right) = \sum_{i=k}^{n} ia_i - \sum_{i=k+1}^{n} (i-1)a_i$$

$$= ka_k + \sum_{i=k+1}^{n} a_i$$

Коды Варшамова—Тененгольца: исправление одной ошибки выпадения

Получаем

$$S' = S - \left(ka_k + \sum_{i=k+1}^{n} a_i\right) = S - ka_k - n_1$$

Так как $S \equiv 0 \pmod{(n+1)}$, то

$$S' \equiv -n_1 - ka_k \pmod{(n+1)}$$

Если $a_k = 0$, то $-S' \equiv n_1$.

Если $a_k = 1$, то

$$-S' \equiv n_1 + k = (n - k - n_0) + k = n - n_0$$

Коды Варшамова—Тененгольца: исправление одной ошибки выпадения

Итак,

- если $a_k = 0$, то $(-S') \bmod (n+1) = n_1$,
- если $a_k = 1$, то $(-S') \bmod (n+1) = n - n_0$.

Осталось определить, чему равно a_k.

Заметим, что $\|\boldsymbol{a}'\| \geq n_1$,

$$\|\boldsymbol{a}'\| \leq (n-1) - n_0$$

Отсюда $n_1 \leq \|\boldsymbol{a}'\| < n - n_0$.

То есть, если $(-S') \bmod (n+1) \leq \|\boldsymbol{a}'\|$, то это n_1, а в противном случае это $n - n_0$.

Коды Варшамова—Тененгольца: исправление одной ошибки выпадения

Итоговый алгоритм восстановления $\boldsymbol{a}$ по $\boldsymbol{a}'$:

- Вычисляем величину

$$T \coloneqq \left(-\sum_{i=1}^{n-1} i a_i'\right) \bmod (n+1)$$

- Если $T \le \|\boldsymbol{a}'\|,$ то в слово $\boldsymbol{a}'$ вставляем перед T-й с конца единицей символ 0.
- Если $T > \|\boldsymbol{a}'\|,$ то в слово $\boldsymbol{a}'$ вставляем перед $(n-T)$-м с конца нулём символ 1.

Коды Варшамова—Тененгольца: исправление одной ошибки вставки

Теперь рассмотрим задачу, когда $\boldsymbol{a}'$ получено из $\boldsymbol{a}$ вставкой символа:

$$\boldsymbol{a}' = \ldots a_k \, x a_{k+1} \ldots$$

(Если $k = 0$, то $\boldsymbol{a}' = x\boldsymbol{a}$; если $k = n$, то $\boldsymbol{a}' = \boldsymbol{a}x$)

Тогда $S' = S + (k+1)x + \sum_{i>k} a_i$, и значит

$$S' \equiv (k+1)x + \sum_{i>k} a_i = (k+1)x + n_1$$

Положим $T \coloneqq S' \bmod (n+1)$.

Коды Варшамова—Тененгольца: исправление одной ошибки вставки

$$\boldsymbol{a}' = \dots a_k \, x a_{k+1} \dots$$
$$S' \equiv (k+1)x + n_1$$

Положим $T \coloneqq S' \bmod (n+1)$.

Есть два случая, когда $T = 0$:

- $(k+1)x + n_1 = 0$. Тогда $x = 0$ и $a_{k+1} = \cdots = a_n = 0$.
- $(k+1)x + n_1 = n + 1$. Тогда $x = 1$ и $a_{k+1} = \cdots = a_n = 1$.

В обоих случаях $\boldsymbol{a}$ получается из $\boldsymbol{a}'$ удалением последнего символа.

Коды Варшамова—Тененгольца: исправление одной ошибки вставки

$$\boldsymbol{a}' = \dots a_k \, x a_{k+1} \dots$$
$$T \equiv (k+1)x + n_1$$

Теперь рассмотрим случай, когда $T = \|\boldsymbol{a}'\| > 0$.

Это возможно только в одном из двух случаев:

- $a_1 = \cdots = a_k = x = 0$
- $a_1 = \cdots = a_k = x = 1$

В любом случае, если $T = \|\boldsymbol{a}'\|$, то $\boldsymbol{a}$ получается из $\boldsymbol{a}'$ удалением первого символа.

Коды Варшамова—Тененгольца: исправление одной ошибки вставки

$$\boldsymbol{a}' = \dots a_k \, x a_{k+1} \dots$$
$$T \equiv (k+1)x + n_1$$

Остался случай $0 < T \neq \|\boldsymbol{a}'\|$.

- Если $x = 0$, то $T = n_1 < \|\boldsymbol{a}'\|$.
- Если $x = 1$, то $T = k + 1 + n_1 > \|\boldsymbol{a}'\|$.
  При этом оказывается, что
  $$T = k + 1 + (n - k - n_0) = n + 1 - n_0.$$

В обоих случаях нужная для восстановления $\boldsymbol{a}$ информация у нас есть.

Коды Варшамова—Тененгольца

Возможные обобщения кодов В.—Т.:

- Произвольный фиксированный модуль $l > n$:

$$\{a_1 a_2 \dots a_n \mid \sum_{i=1}^{n} i a_i \equiv 0 \pmod{l}\}$$

- Дополнительные соотношения, например:

$$\{a_1 a_2 \dots a_n \mid \sum_{i=1}^{n} i a_i \equiv \sum_{i=1}^{n} i^2 a_i \equiv 0 \pmod{l}\}$$

Ошибки замещения

- Односторонние ошибки. Например, если в двоичном канале возможны только замещения вида $0 \to 1$ или только $1 \to 0$. Канал связи в этом случае называется *несимметричным*.
- Двусторонние (симметричные) ошибки. Если возможно замещение символов $b_1 \to b_2$, то возможно и замещение $b_2 \to b_1$.
- Канал связи, в котором ошибки только симметричные, называется *симметричным*.

Ошибки замещения

Коды Варшамова—Тененгольца могут исправлять единичные односторонние ошибки замещения:

- Если $\boldsymbol{a}'$ получается из $\boldsymbol{a}$ замещением i-го символа с 0 на 1, то $S' \equiv i \pmod{(n+1)}$
- Если $\boldsymbol{a}'$ получается из $\boldsymbol{a}$ замещением i-го символа с 1 на 0, то $S' \equiv -i \pmod{(n+1)}$

(Исправлять оба типа ошибок одновременно коды Варшамова—Тененгольца не могут.)

Обозначение кодов

Далее будем изучать коды, исправляющие ошибки замещения, значит, метрика по умолчанию — метрика Хемминга.

Обозначение кода с заданными параметрами

Если C — q-ичный код с длиной слов n, числом слов M и кодовым расстоянием d, то пишут:

«C является $(n, M, d)_q$-кодом»

Если код двоичный, то символ q не указывают.

Граница сферической упаковки

Теорема. (Граница Хемминга (R.W. Hamming), граница сферической упаковки)
Для любого $(n, M, d)_q$-кода имеем

$$M \leq \frac{q^n}{\left|S_{\lfloor (d-1)/2 \rfloor}(\mathbf{0})\right|}$$

В двоичном случае

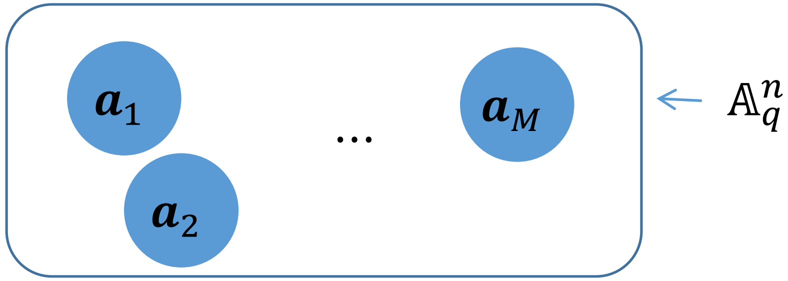
$$M \leq \frac{2^n}{\sum_{k=0}^{\lfloor (d-1)/2 \rfloor} \binom{n}{k}}$$

Коды, достигающие эту границу, называются *совершенными* или *плотно упакованными*.

Граница сферической упаковки

Доказательство теоремы:
Пусть $C = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_M\}$ — $(n, M, d)_q$-код.

Так как $d(C) = d$, то шары радиуса $\lfloor (d-1)/2 \rfloor$ с центрами в кодовых словах не пересекаются:



Отсюда $q^n \geq \sum_{j=1}^{M} \left| S_{\lfloor (d-1)/2 \rfloor}(\boldsymbol{a}_j) \right| = M \cdot \left| S_{\lfloor (d-1)/2 \rfloor}(\boldsymbol{0}) \right|$

«Анти-Хемминг»

Теорема. (В некотором смысле, обратная границе Хемминга)
Пусть числа $q, n, M, d \in \mathbb{N}$ таковы, что

$$M \leq \frac{q^n}{|S_d(\mathbf{0})|}.$$

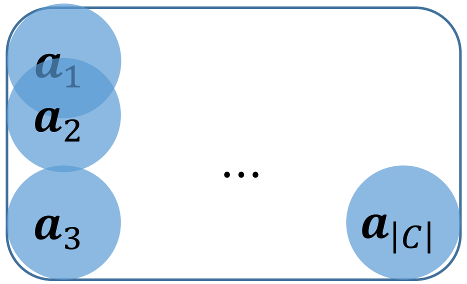
Тогда существует $(n, M, d)_q$-код.

«Анти-Хемминг»

Доказательство теоремы:
Пусть $C = \{\boldsymbol{a}_1, \dots, \boldsymbol{a}_{|C|}\}$ — код максимальной мощности с кодовым расстоянием d и длиной слов n.

Тогда шары радиуса d с центрами в кодовых словах покрывают целиком множество $\mathbb{A}_q^n$ (иначе код C можно было пополнить любым из слов, не лежащих ни в одном из этих шаров).

Отсюда $\sum_{j=1}^{|C|} \left|S_d(\boldsymbol{a}_j)\right| \geq q^n$,
следовательно $|C| \geq M$.



Граница Синглтона

Теорема. (R.C. Singleton)
Для любого $(n, M, d)_q$-кода имеем
$$|C| \leq q^{n-d+1}$$

Коды, на которых достигается граница Синглтона, называются *MDS-кодами* (maximum distance separable codes).

Граница Синглтона

Доказательство:

Пусть $C = \{\boldsymbol{a}_1, \dots, \boldsymbol{a}_M\}$ — $(n, M, d)_q$-код.

Рассмотрим слова $\{\boldsymbol{a}'_i\}_{i=1}^{M}$, где $\boldsymbol{a}'_i$ получено из $\boldsymbol{a}_i$ отбрасыванием $(d-1)$ последних координат.

Так как $d(\boldsymbol{a}_i, \boldsymbol{a}_j) \geq d$ для любых i, j, то все слова $\boldsymbol{a}'_i$ различны. Их количество не превосходит числа всех q-ичных слов длины $(n-d+1)$.

Поэтому и $M \leq q^{n-d+1}$.

На лекции мы рассмотрели:

- Исправление ошибок, кодовое расстояние
- Коды Варшамова—Тененгольца
- Простые границы мощностей кодов