О задачах поиска

1. Общие задачи поиска

1.1. **Поиск монет.** Имеется n монет, часть из которых – фальшивые. Известно, что фальшивая монета весит 9 гр., а настоящая – 10 гр. Обозначим через $A(n)$ минимальное число взвешиваний на электронных весах (они показывают точный вес того, что на них положили), позволяющих найти все фальшивые монеты, а через $A(n, \leq t)$ и $A(n, t)$ минимальное число взвешиваний при условии, что фальшивых монет не более t и. соответственно, ровно t. Если "план" взвешиваний составлен заранее и не зависит от результатов предшествующих взвешиваний (неадаптивный поиск, или *без обучения*), то соответствующие величины будем помечать * сверху, т.е. $A^*(n)$, $A^*(n, \leq t)$, $A^*(n, t)$.
Упражнение.(0.5 балла) "Постановка" * не уменьшает соответствующую величину, т.е. $A^*(n) \geq A(n)$ и т.д.

Задача 1. (2 балла) Докажите, что:

a)(1.5 балла)

$$A(n) \geq \frac{n}{\log_2(n+1)},$$

b)$A(n, t) \geq \log_{t+1} C_n^t$ (1.5 балла)

Задача 2. (6 баллов) Докажите, что

$$A^*(n) \geq 2\frac{n}{\log_2(n+1)}(1 + o(1))$$

Задача 3. (5 баллов) Постройте план взвешиваний без обучения с "примерно" $2n/\log_2(n+1)$ взвешиваниями, т.е. докажите, что

$$A^*(n) \leq \frac{2n}{\log_2(n+1)}(1 + o(1))$$

1.2. **Жесткость пространств Хэмминга.** Обозначим через H_q^n множество всех q^n слов длины n над конечным алфавитом мощности q. Определим расстояние Хэмминга $d(x, y)$ между словами $x = (x_1, \ldots, x_n)$ и $y = (y_1, \ldots, y_n)$ как число позиций (координат) где эти слова различны, т.е.

$$d(x, y) = |\{i : x_i \neq y_i\}|$$

Множество точек $a^{(1)}, \ldots, a^{(N)}$ метрического пространства $\mathcal{M}$ называется его "репером", если любая точка пространства однозначно определяется ее расстояниями до точек репера. Минимальная мощность репера называется жесткостью пространства и обозначается $\chi(\mathcal{M})$.

Задача 4. (1 балл) Докажите, что $|A(n) - \chi(H_2^n)| \leq 1$, т.е. что жесткость двоичного пространства Хэмминга и минимальное число взвешиваний для n монет – это почти одно и тоже.

Задача 5. (7 баллов) Докажите, что для $q = 3$ и $q = 4$

$$\chi(H_q^n) = 2\frac{n}{\log_q(n+1)}(1 + o(1))$$

Гипотеза 1. (∞ баллов) Для произвольного q

$$\chi(H_q^n) = 2\frac{n}{\log_q(n+1)}(1 + o(1))$$

1.3. **Различные модели поиска.** Имеется множество A из n элементов и его неизвестное подмножество $X \subset A$. Принято множество X рассматривать как множество дефектных элементов A. Как найти X за минимальное число вопросов? Будем как и в разделе 1.1 обозначать соответствующие числа через $A(n),\ A(n, \leq t),\ A(n, t)$, указывая какая модель поиска рассматривается.

Вопросом является любое подмножество $Q \subset A$. Ответ является функцией от $X \cap Q$, которая и определяет ту или иную постановку задачи поиска. Наиболее известны следующие функции-ответы:

1)$f(X \cap Q) = |X \cap Q|$, т.е. в ответ говорят сколько дефектных элементов есть в Q. Эта задача уже описана в разделе 1.1

2)$f(X \cap Q) = 0$ если $X \cap Q = \emptyset$, и $f(X \cap Q) = 1$ – в противном случае. Т.е. в ответ говорят есть или нет дефектные элементы в Q. Это так называемая дизъюнктивная ($\bigvee$) модель. В качестве примера этой модели рассмотрим n лампочек, часть которых дефектна (не горит), и n патронов, соединенных *последовательно*. В качестве вопроса выбранные лампочки вставляются в патроны, а затем включается ток. Очевидно, что если среди выбранных лампочек есть неисправная, то лампочки не загорятся, а если же среди выбранных их нет, то лампочки будут гореть.

3)$f(X \cap Q) \equiv |X \cap Q| mod 2$, т.е. в ответ говорят четно ли число дефектных элементов в Q. На этой странной модели "стоит" теория кодирования.

Очень интересным является вопрос "адаптивный поиск vs неадаптивный" для каждой из моделей.

Задача 6. (3 балла) Докажите, что для дизъюнктивной модели:

а) $A_{\bigvee}(n, 2) = 2 \log_2 n(1 + o(1))$ при $t = 2$(2 балла);

б) $A_{\bigvee}(n, t) = t \log_2 n(1 + o(1))$ при любом фиксированном t(3 балла)

Ответ для этой задачи при неадаптивном поиске мне неизвестен.

2. Конечные поля и линейная алгебра над ними

2.1. **Построение конечных полей.** Поле это множество, на котором определены сложение и умножение такие, что все множество является группой по сложению, а множество без нуля (нейтрального элемента по сложению) является группой по умножению, обе группы коммутативны ($a+b = b+a$ и $ab = ba$), и операции связаны дистрибутивным законом: $a(b + c) = ac + bc$. Хорошо известные вам примеры полей: поле рациональных чисел $\mathbb{Q}$, поле действительных чисел $\mathbb{R}$ и поле комплексных чисел $\mathbb{C}$. Множество $\mathbb{Z}_n$

вычетов целых чисел по модулю числа n является полем тогда и только тогда, когда n – простое число. Действительно, если $n = n_1 n_2$ составное число, то $n_1 n_2 \equiv 0(\mod n)$, т.е. произведение двух ненулевых элементов равно нулю, а это в поле невозможно (почему?). Если же $n = p$ – простое число, то надо убедиться, что для любого ненулевого остатка a элементы ax различны и отличны от нуля (по модулю), когда x nробегает все $p-1$ ненулевых остатков и, следовательно, найдется x такой, что $ax = 1$.

Но как построить поле из 4-х элементов? $\mathbb{Z}_4$, как мы выяснили, не годится. Попробуем строить это поле "перебором". А именно, в поле должны быть 0 и 1, при этом $1 + 1 = 0$ (иначе снова получим вычеты по модулю 4). Т.е. в нашем поле есть подполе $\mathbb{Z}_2$. Обозначим один из двух отличных от них элементов через α. Так как $1+\alpha$ это тоже элемент поля, то это и есть четвертый элемент поля (проверьте). Но поле замкнуто не только по сложению, но и по умножению, отсюда следует, что $\alpha(1 + \alpha) = 1$ (проверьте, что другие варианты несовместимы с аксиомами поля). Так как $1 + 1 = 0$, то получаем соотношение $\alpha(1 + \alpha) + 1 = 0$, которое и определяет наше поле. Многочлен $\alpha^2 + \alpha + 1$ является *неприводимым многочленом* над полем коэффициентов $\mathbb{Z}_2$, т.е. он не разлагается в произведение двух многочленов (каждый ненулевой степени). Очевидно, что неприводимость – это аналог понятия простого числа, только для многочленов.

Задача 7. (1 балл) Докажите, что для многочленов не более 3-й степени неприводимость равносильна отсутствию корней (в поле коэффициентов).

Обобщим приведенное выше построение поля из 4-х элементов (обозначение $\mathbb{F}_4$ или $GF(4)$). Рассмотрим $\mathbb{Z}_p = \mathbb{F}_p$ в качестве поля коэффициентов и пусть многочлен $f(x) = f_o + f_1 x + \ldots + f_m x^m$ ($f_i \in \mathbb{F}_p$) неприводим. В качестве поля $\mathbb{F}_{p^m}$ возьмем множество остатков многочленов (с коэффициентами из $\mathbb{Z}_p = \mathbb{F}_p$) по модулю многочлена $f(x)$ с естественными операциями. Оказывается, что других конечных полей нет!

Задача 8. (1.5 балла) а) Постройте поле из 25 элементов. Например, как целые комплексные числа по модулю 5.(1 балл)

б)Постройте поле из 27 элементов.

Сейчас, как попутный результат из теории векторных пространств, мы покажем, что мощность любого конечного поля есть степень простого числа.

2.2. **Векторные пространства над конечными полями.** Множество V называется векторным пространством над полем K, если над элементами V, называемых векторами, определены операции сложения и умножения на элементы поля K, удовлетворяющие следующим свойствам:

1) векторы образуют коммутативную группу по сложению;

2) $\lambda(\mu a) = (\lambda\mu)a,\ (\lambda+\mu)a = \lambda a + \mu a$ и $\lambda(a+b) = \lambda a + \lambda b$, где $\lambda, \mu \in K,\ a, b \in V$.

Векторы $a^{(1)}, \ldots, a^{(N)}$ называются линейно зависимыми (сокращенно, ЛЗ), если $\exists \lambda_1, \ldots, \lambda_N \in K$ не равные тождественно нулю, такие что

$$\sum_{i=1}^{N} \lambda_i a^{(i)} = \overrightarrow{0} \tag{1}$$

Если же (1) возможно только если все λ_i равных 0, то такие векторы называются линейно независимыми (ЛНЗ).

Задача 9. (2 балла) Рассмотрим поле действительных чисел $\mathbb{R}$ как векторное пространство над полем рациональных чисел $\mathbb{Q}$. "Векторы" $\sqrt{2}$ и $\sqrt{3}$ ЛЗ или ЛНЗ?

Множество векторов $E = \{e^{(1)}, \ldots, \}$ называется *базисом* пространства V если для любого $x \in V$ существует и единственный набор коэффициентов $x_1, \ldots, x_n$ такой, что

$$\sum_{j=1}^{n} x_j e^{(i_j)} = x \tag{2}$$

Попробуйте доказать следующие утверждения.

Теорема 1. (2 балла) Если множество векторов $E = \{e^{(1)}, \ldots, \}$ ЛНЗ, но далее не расширяемо с сохранением этого свойства (т.е. $E \cup v$ уже ЛЗ для любого $v \in V$), то E -

базис.

Теорема 2. (1 балла) У любого векторного пространства есть базис.

Теорема 3. (2 балла) Все базисы данного векторного пространства имеют одинаковую мощность (называемую *размерностью* пространства).

Докажем теперь обещанное – мощность любого конечного поля есть степень простого числа. Пусть K - конечное поле. Определим *характеристику* $\chi(K)$ поля как минимальное натуральное число $n > 0$, такое что $n \times 1 = 0$.

Задача 9. (1 балл) $\chi(K)$ –простое число p и K содержит $\mathbb{F}_p = \mathbb{Z}_p$ в качестве подполя.

Теперь рассмотрим поле K как векторное пространство над $\mathbb{F}_p$. Тогда у него есть размерность m и, следовательно, число векторов в K равно p^m.

Конечные поля обладают рядом замечательных свойств. Например,

Теорема 4. (3 балла) Конечное поле (без 0) является циклической группой по умножению, т.е. $\exists \alpha \in K$ такое, что $\{\alpha^0, \alpha^1, \ldots, \alpha^{|K|-2}\} = K \setminus \{0\}$.

Конечные поля хороши еще и тем, что позволяют многое подсчитать.

Задача 10. (2 балла) Сколько в n-мерном векторном пространстве над полем $\mathbb{F}_q$ имеется k-мерных векторных подпространств? Если не сделали задачу в общем случае, то попробуйте (за 1 балл) частный случай $k = 1$, т.е. сколько "прямых проходящих через 0.

3. Об одной задаче Эрдёша и случайном кодировании

3.1. **2-дизъюнктивные семейства подмножеств.** Рассмотрим следующую задачу, ставшую хорошо известной благодаря двум работам П.Эрдёша, но известную и задолго до него как задача о дизъюнктивных кодах (Каутц и Синглетон). Семейство

$C_1, \ldots, C_M$ подмножеств n-элементного множества A называется 2-дизъюнктивным, если из $C_i \subseteq C_j \cup C_k$ следует, что $i = j$ или $i = k$. Иными словами, объединение двух множеств из семейства не покрывает никакое третье множество из этого же семейства. Дайте сами определение r-дизъюнктивного семейства подмножеств.

Как найти 2-дизъюнктивное семейство с максимальным числом подмножеств? Обозначим это максимальное число через $M(2, n)$. Сопоставим каждому множеству C_i его характеристический вектор $c^{(i)} \in \{0, 1\}^n$, а все множество этих векторов будем называть *кодом* (2-дизъюнктивным). Рассмотрим $M \times n$-матрицу $\mathcal{C}$, строками которой являются векторы $c^{(i)}$. Условие 2-дизъюнктивности можно переформулировать таким образом:

Матрица $\mathcal{C}$ называется 2-дизъюнктивной, если для любой строки i и других двух строк j и k существует как минимум один столбец q такой, что в нем i-ая строка имеет значение 1, а две другие - значение 0.

Мы не умеем явно строить "достаточно хорошие" 2-дизъюнктивные семейства множеств или 2-дизъюнктивные матрицы с "достаточно большим" числом строк. Вместо этого мы покажем, что почти все матрицы "достаточно хороши"!

Рассмотрим случайную $M \times n$-матрицу, т.е. ее элементы принимают значения 0 и 1 равновероятно и независимо друг от друга. Подсчитаем вероятность P_{good} того, что условие 2-дизъюнктивности для данной "тройки" $i;\ j, k$ выполнено в q-ом столбце. Очевидно, что $p_{good} = 1/8$ и, следовательно, вероятность того, что условие 2-дизъюнктивности для данной "тройки" $i;\ j, k$ невыполнено в q-ом столбце есть $p_{bad} = 7/8$. Так как столбцов n, то вероятность того, что условие 2-дизъюнктивности для данной "тройки" $i;\ j, k$ невыполнено ни в одном столбце равна $P_{bad} = (7/8)^n$. Так как всего "троек" $i;\ j, k$ не более $M^3/2$, то вероятность того, что есть хотя бы одна "плохая" тройка не превышает $M^3/2(7/8)^n$. Следовательно, если $M^3/2(7/8)^n < 1$, то вероятность "хорошей"матрицы отлична от нуля. В частности, существует 2-дизъюнктивная $\lfloor (8/7)^{n/3} \rfloor \times n$ матрица, и значит $M(2, n) \geq (8/7)^{n/3}$. *Это и есть метод случайного кодирования!*

Что означает утверждение "почти все матрицы достаточно хороши"? Уменьшим незначительно величину M (мощность 2-дизъюнктивного кода). Например, положим $M = \lfloor (8/7)^{n/3}/\ln n \rfloor$ или же $M = \lfloor (8/7)^{(n-\ln n)/3} \rfloor$. Теперь вероятность того, что есть хотя бы одна "плохая" тройка, стремится к 0 с ростом n, и следовательно, доля хороших матриц стремится к 1.

Улучшим метод случайного кодирования. Действительно, мы возможно выбрасываем матрицы, в которых совсем немного "плохих" троек. Давайте вместо этого выбрасывать "плохие" тройки. Вероятность "тройки" $i;\ j,k$ быть "плохой" (т.е. условие 2-дизъюнктивности не выполнено ни в одном столбце) равна $P_{bad} = (7/8)^n$. Поэтому *среднее* число "плохих" троек меньше или равно $(7/8)^n M^3/2$. Почему?!

Чтобы доказать это “почему” рассмотрим случайную величину $\xi_{i;j,k}$ от матрицы, равную 1, если тройка $i;j,k$ плохая для этой матрицы, и 0 - в противном случае. По определению среднего $\mathbb{E}(\xi_{i;j,k}) = (7/8)^n$. Так как число плохих троек для матрицы равно $\xi = \sum_{i;j,k} \xi_{i;j,k}$, то среднее число плохих троек равно

$$\mathbb{E}(\xi) = \mathbb{E}(\sum_{i;j,k} \xi_{i;j,k}) = \sum_{i;j,k} \mathbb{E}(\xi_{i;j,k}) = MC_M^2 (7/8)^n$$

Вернемся к доказательству. Найдется матрица с числом "плохих" троек не более среднего. Из этой матрицы выкинем с каждой плохой тройкой один из элементов этой тройки (например, i). Тогда плохих троек не останется, а число строк новой матрицы будет не меньше чем $M - (7/8)^n M^3/2$. Выберем M так, чтобы не выбрасывать более половины строк, т.е. $(7/8)^n M^3/2 = M/2$ или $M = (8/7)^{n/2}$. Следовательно существует 2-дизъюнктивная $2^{-1}\lfloor (8/7)^{n/2} \rfloor \times n$ матрица и $M(2,n) \geq 2^{-1}(8/7)^{n/2}$. *Это метод случайного кодирования с “выбрасыванием”.*

Еще одно возможное улучшение метода случайного кодирования состоит в необязательно равновероятном выборе 0 и 1 при порождении случайных матриц.

Задача 11. (2 балла) Рассмотрите случайные матрицы, в которых 1 порождается с вероятностью p, а 0 - с вероятностью $1-p$. Повторите все проделанные выше рассуждения и выберите параметр p оптимально! Какая оценка на $M(2,n)$ у вас получится?

3.2. **Поиск в дизъюнктивной модели с 2 дефектами и 2-дизъюнктивные семейства подмножеств.** Построим из 2-дизъюнктивного семейства $C_1,\dots,C_M$ подмножеств n элементного множества A план из n вопросов неадаптивного поиска двух дефектных элементов в множестве из M элементов для 2-ой (дизъюнктивной) модели. Точке $a\in A$ сопоставим вопрос-множество $Q_a\subseteq\{1,\dots,M\}$, состоящий из $i\in\{1,\dots,M\}$ таких, что $a\in C_i$. Пусть j,k - это дефектные элементы. Тогда вектором ответа $S=(s_1,\dots,s_n)$ на все n вопросов будет характеристический вектор множества $C_j\cup C_k$ или, что тоже самое, $S=c_j\bigvee c_k$. Рассмотрим вектор S из 0 и 1 как характеристический вектор некоторого множества $\mathcal{S}$, которое, как мы знаем, равно $C_j\cup C_k$. Тогда C_j и C_k являются подмножествами $\mathcal{S}$, тогда как 2-дизъюнктивность означает, что для любого $i\notin\{j,k\}$ множество C_i не является подмножеством $\mathcal{S}$ (так как существует точка-вопрос $a\in A$ такая, что $a\in C_i$, но $a\notin C_j$ и $a\notin C_k$). Это дает одновременно и достаточно простой алгоритм нахождения дефектных позиций. Здесь выигрыш в сложности по сравнение с переборным алгоритмом не так велик (M вместо M^2), но для r-дизъюнктивности это уже будет выигрыш M вместо M^r.

Задача 12. (2 балла) Докажите, что

$$A^*_{\bigvee}(M(2,n),2)\le n \tag{3}$$

Воспользуйтесь этим неравенством, чтобы получить для дизъюнктивной модели оценку типа $A^*_{\bigvee}(n,2)\le c\log_2 n(1+o(1))$. Какое c вы можете получить на основе результатов этого раздела?

Задача 13. (3 балла) Проведите самостоятельно рассуждения этого раздела для 3-дизъюнктивных семейств подмножеств.

4. Коды, исправляющие ошибки, или задача поиска для 3-ей модели

4.1. **Коды, ошибки и пространство Хэмминга.** Зададимся вопросом как передавать сообщения по каналу связи, в котором при передаче n символов происходит не более t ошибок, таким образом, чтобы на приемной стороне было возможно однозначное восстановление (*декодирование*) переданного сообщения? Будем передавать не все возможные n символьные слова, а только их часть, называемую *корректирующим кодом* , или кодом, исправляющим t ошибок. Математическая модель такова:
по каналу передаются символы из q-ичного алфавита H_q (чаще всего $q = 2$);
не более t ошибок означает, что переданное слово c и принятое слово y отличаются не более чем в t позициях, т.е. расстояние Хэмминга $d(c, y)$ между c и y не более t;
однозначное восстановление означает, что шары (в метрике Хэмминга) радиуса t с центрами в разных словах кода не пересекаются, т.е. что $d(c, c') > 2t$ для любых двух различных слов кода c и c'.
Итак, *q-ичный код длины n, исправляющий t ошибок,* это произвольное множество $C \subset H_q^n$ такое, что

$$d(C) := \min_{c \neq c' \in C} d(c, c') \geq 2t + 1$$

Величина $d(C)$ называется *расстоянием* кода. Иначе говоря, код, исправляющий t ошибок, это центры упаковки шаров радиуса t в метрическом пространстве Хэмминга.
Ограничимся двоичным случаем ($q = 2$) и вернемся к 3-й модели, когда ответом на вопрос Q относительно неизвестного множества $E \subset \{1, 2, \ldots, n\}$ является

$$|E \cap Q| \mod 2 = \sum_{i=1}^{n} q_i e_i \mod 2, \tag{4}$$

где $e = (e_1, \ldots, e_n)$ и $q = (q_1, \ldots, q_n)$ это характеристические векторы множеств E и Q, соответственно. В качестве множества E возьмем множество позиций, на которых произошли ошибки, т.е. $E = \{i : c_i \neq y_i\}$, и следовательно, $y = c + e \mod 2$.
Пусть $Q^{(1)}, \ldots, Q^{(r)}$ - набор вопросов, осуществляющих неадаптивный поиск для 3-й

модели при числе дефектов не более t. Зададим с их помощью код C, исправляющий t ошибок, системой уравнений

$$\sum_{i=1}^{n} q_i^{(j)} c_i \equiv 0 \mod 2, \quad j = 1, \ldots, r, \tag{5}$$

где $q^{(j)} = (q_1^{(j)}, \ldots, q_n^{(j)})$ это характеристический вектор множества-вопроса $Q^{(j)}$. Эта система линейных уравнением записывается в матричной форме как

$$C = \{x \in H_2^n : Hx^T = 0\}, \tag{6}$$

где $r \times n$ матрица H, строками которой являются векторы $q^{(j)}$, $j = 1, \ldots, r$, называется *проверочной* матрицей кода C, а сам код C будем называть для краткости кодом с r проверками.

Для принятого вектора y вычислим его *синдром* $S(y) = (s_1(y), \ldots, s_r(y))$, где по определению

$$s_j(y) = \sum_{i=1}^{n} q_i^{(j)} y_j \mod 2 \tag{7}$$

Тогда код C - это множество векторов, синдром которых равен 0. А так как синдром линеен, т.е. $S(a+b) = S(a) + S(b)$, а $y = c + e \mod 2$ и $S(c) = 0$, то получаем, что

$$S(y) = S(e). \tag{8}$$

Следовательно, по синдрому принятого вектора мы однозначно найдем вектор-ошибку и, тем самым, восстановим передававшийся кодовой вектор как $c := y+e \mod 2$. Код C, задаваемый уравнениями (5), является линейным подпространством (проверьте аксиомы!), и поэтому называется линейным кодом. Тем самым, мы установили *эквивалентность двух понятий: план неадаптивного поиска для 3-й модели при числе дефектов не более t и линейный код, исправляющий t ошибок.* Осталось за малым - научиться строить для 3-й модели планы неадаптивного поиска :-) Но благодаря тому, что ответ на вопрос в этой модели является линейным, см. (4), ситуация намного лучше, чем для

1-й или 2-й модели.

Заметим, что число решений системы (5), т.е. число слов в коде, не меньше чем 2^{n-r}, а если уравнения ЛНЗ – то ровно 2^{n-r}. Поэтому передавая слова кода, а они длины n, мы передаем $k = n - r$ "полезных" (информационных) бит, и величину $R = k/n$, которую естественно рассматривать как КПД метода передачи, принято называть *скоростью* кода.

Обозначим $h^{(1)}, \ldots, h^{(n)}$ столбцы матрицы H. Тогда слово $c = (c_1, \ldots, c_n)$ принадлежит коду $\Leftrightarrow \sum_{i=1}^{n} c_i h^{(i)} = 0$. Т.е. кодовые слова это тоже самое, что соотношения линейной зависимости для вектров $h^{(1)}, \ldots, h^{(n)}$. Тем самым мы доказали

Критерий Боуза. Расстояние кода C с проверочной матрицей H не меньше $d \Leftrightarrow \forall\ (d-1)$ столбцов H ЛНЗ.

Отсюда, в частности, следует, что $r \times (2^r - 1)$ матрица H, столбцами которой являются все $2^r - 1$ ненулевых r-столбцов, задает код, исправляющий одиночные ошибки, – это знаменитый код Хэмминга.

Напомним, что код, исправляет t ошибок $\Leftrightarrow$ шары радиуса t (в метрике Хэмминга) с центрами в словах кода непересекаются. Так как всего точек в пространстве Хэмминга 2^n, а в каждом шаре $V(n,t) = \sum_{i=0}^{t} C_n^i$ точек, то число шаров–кодовых слов не превышает $2^n/V(n,t)$ – это называется границей Хэмминга или границей плотной упаковки. В частности, при $t = 1$ получаем что мощность *любого* (не обязательно линейного) кода не превосходит $2^n/(n+1)$, и эта граница достигается на кодах Хэмминга. Т.е. коды Хэмминга дают плотнейшую (без дыр) упаковку пространства Хэмминга (булева куба) размерности $2^r - 1$ шарами радиуса 1! Еще известны две плотнейшие (совершенные) упаковки: двоичный код Голея длины 23, исправляющий три ошибки, и троичный код Голея, длины 11, исправляющий две ошибки. Одним из самых замечательных результатов теории кодирования является доказанная в начале 70-х годах XX века гипотеза, что других совершенных кодов, исправляющих t ошибок, при $t > 1$ не существует.

4.2. **Как строить коды, исправляющие t ошибок? При $t = \lambda n$ – случайно.** Рассмотрим случайную проверочную $r \times n$ матрицу H, т.е. ее элементы – это случайные величины, принимающие значения 0 и 1 с вероятностью $1/2$ независимо друг от друга. Вероятность того, что некоторое *ненулевое* слово $x = (x_1, \ldots, x_n)$ удовлетворяет конкретному уравнению из системы (5), равна $1/2$ (почему?). Так как таких уравнений r и их коэффициенты независимые (!) случайные величины, то вероятность того, что слово x удовлетворяет системе (5), равна $1/2^r$. Следовательно, если $\sum_{i=0}^{d-1} 2^{-r} < 1$, то существует код с r проверками и расстоянием не меньше d. Эта граница *существования* кодов известна как граница Гилберта (не Гильберта!).

Задача 14. (1 балл) Докажите тоже самое *подсчетом*. Например, рассмотрите коэффициенты матрицы H как "неизвестные" в уравнении $Hx^T = 0$ (а $x_1, \ldots, x_n$ – коэффициенты уравнения), и покажите, что число решений не превосходит $2^{(n-1)r}$.

Воспользуйтесь критерием Боуза и решите

Задача 15. (2 балла) Докажите, что если $\sum_{i=0}^{d-2} 2^{-r} < 1$, то существует код с r проверками и расстоянием не меньше d (граница Варшамова).

Воспользуемся известным неравенством на биномиальные коэффициенты

$$\sum_{i=0}^{\delta n} C_n^i \leq 2^{nH_2(\delta)}, \tag{9}$$

где $\delta \leq 1/2$ и $H_2(x) = -(x \log_2 x + (1-x)\log_2(1-x))$, чтобы выписать асимптотическую форму этих границ.

Граница *существования* кодов (Варшамова-Гилберта)

$$R(n, \delta n) \geq 1 - H_2(\delta) + o(1) \tag{10}$$

Граница *несуществования* кодов (Хэмминга)

$$R(n, \delta n) \leq 1 - H_2(\delta/2) + o(1) \tag{11}$$

Эти две границы сильно расходятся.

Обозначим через $\mathbb{A}(n, d)$ максимальную мощность кода (не обязательно линейного) длины n и расстоянием d.

Задача 16. (3 балла) Докажите, что $\mathbb{A}(n, d) \leq \frac{2d}{2d-n}$ при $d > n/2$.

Задача 17. (2 балла) Рассмотрим строки проверочной матрицы кода Хэмминга как базис некоторого линейного кода (называемого двойственным). Докажите, что расстояние нового кода $d = (n + 1)/2$ и он достигает границу предыдущей задачи.

Таким образом, для относительного кодового расстояния $\delta = d/n$ точка 0.5 является *критической*: если $\delta < 0.5$, то скорость лучших кодов отделена от нуля ($R \geq 1 - H_2(\delta) > 0$) и, тем самым, мощность лучших кодов растет *экспоненциально* с длиной кода, а если же $\delta \geq 0.5$, то мощность кода растет не быстрее чем линейно от длины кода (задача 16).

Известно, что предел $\lim_{n\to\infty} R(n, \delta n) = \mathcal{R}(\delta)$ существует, но как в действительности ведет себя функция $\mathcal{R}(\delta)$ при $0 < \delta < 1/2$ науке неизвестно.

Гипотеза 2. (2^∞ баллов)

$$\mathcal{R}(\delta) = 1 - H_2(\delta)$$

Кабатянский Григорий Анатольевич, Институт проблем передачи информации РАН, kaba@iitp.ru