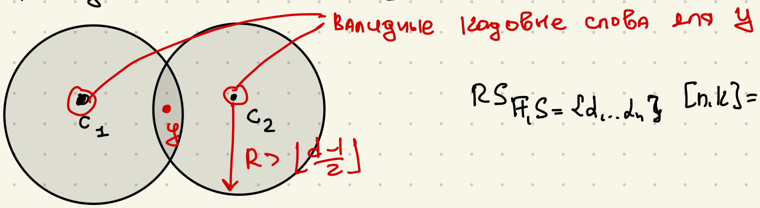
Лекция № 9

Списочное декодирование
Приложение теории кодирования в вычислительной биологии

I. Списочное декодирование

Линейный $[n,k,d]$ код исправляет $\lfloor \frac{d-1}{2} \rfloor$ ошибок



$RS_{\mathbb{F},S}$, $S=\{d_1 \dots d_n\}$ $[n,k] = \{(p(d_1) \dots p(d_n)) \in \mathbb{F}^n \mid p \in \mathbb{F}[x], \deg p \le k-1\}$

Задача поиска двух многочленов: Положим $p_1(x), p_2(x) \in \mathbb{F}[X]$,

$\deg p_1(x) = \deg p_2(x) = k-1$; $n \ge 4k$ — чётное; $d_1 \dots d_n$ — различны.

$\exists T \subset \{1..n\}$, $|T| = \frac{n}{2}$. Далее, пусть нам дан $y \in \mathbb{F}^n$

$$y_i = \begin{cases} p_1(d_i), & i \in T \\ p_2(d_i), & i \in \{1..n\} \setminus T \end{cases}$$

Задача состоит в вычислении $p_1(x), p_2(x)$ по данным парам $\{(d_i, y_i)\}_{i \le n}$

Связь с декодированием RS: $(y_i)_{i \le n}$ — полученное слово, полагая, что $p_1(x)$ - исходное сообщение. Однако, $p_1(x)$ не совпадает с y на $\frac{n}{2}$ значениях; $\frac{n}{2} > \frac{d-1}{2} = \left\lfloor \frac{n-k+1-1}{2} \right\rfloor = \left\lfloor \frac{n-k}{2} \right\rfloor$. Поэтому обычные алг-мы декодирования кода RS не подходят.

Идея списочного декодирования $(y_i - p_1(d_i))(y_i - p_2(d_i)) = 0 \quad \forall i$

Положим, мн-н $Q(x,y) = (y - p_1(x))(y - p_2(x)) = y^2 - \underbrace{(p_1(x) + p_2(x))}_{B(x)} y + \underbrace{p_1(x) \cdot p_2(x)}_{C(x)}$

$= y^2 - B(x) y + C(x)$, где $\deg B(x) = k-1$, $\deg C(x) = 2(k-1)$

$Q(d_i, y_i) = 0 \quad \forall i \le n.$

Алгоритм

Шаг 1 Составить систему лин. ур-ий из $Q(d_i, y_i) = 0$;

$\{\underbrace{b_0 \ldots b_{k-1}}_{\text{коэф-ты } B(x)}, \underbrace{c_0 \ldots C_{2(k-1)}}_{\text{коэф-ты } C(x)}\}$ - неизвестные; и n ур-ий, т.к. $n \geq 4K$,

система будет иметь решение. Получим $B(x), C(x)$ в явном виде

Сложность: $\Theta(n^{\omega})$, $2 < \omega \leq 3$

Шаг 2 Факторизуем мн-н $Q(x, y) = (y - f_1(x))(y - f_2(x))$.
Вернуть $f_1(x), f_2(x)$.

Сложность: факторизация мн-ов от двух переменных степени d над $\mathbb{F}$: $\Theta(d^5 \cdot \lg|F|)$

Корректность На шаге I мы всегда отыщем какие-либо $B(x), c(x)$, $\exists$ решение $B(x) = p_1(x) + p_2(x)$, $c(x) = p_1(x) \cdot p_2(x)$.
Докажем, что на II-м шаге мы получим корректные $p_1(x), p_2(x)$

Лемма $\forall Q(x, y)$, полученного на шаге I, справедливо

$$(y - p_1(x)) \mid Q(x, y) \quad \text{и} \quad (y - p_2(x)) \mid Q(x, y)$$

◁ Док-м утверждение для $p_1(x)$

Заметим, что $Q(x, y)$ - унитарный (от y) мн-н. Для того, чтобы показать, что $(y - \beta)$ делит Q, достаточно показать, что β - корень $Q \Rightarrow$ $Q(\beta) = 0$. Чтобы показать, $y - p_1(x) \mid Q(x, y)$, покажем, что $Q(x, p_1(x)) = 0$

$$\text{✻ } R(x) := Q(x, p_1(x)), \quad \deg R(x) \leq 2(k-1)$$

$$\quad \| $$

$$p_1(x)^2 - (p_1(x) + p_2(x)) \cdot p_1(x) + p_1(x) p_2(x)$$

Заметим, что $\exists \frac{n}{2} \geq 2K$ d_i, т.ч. $p_1(d_i) = y_i$. Для таких d_i, справедливо:
$R(d_i) = Q(d_i, p_1(d_i)) = Q(d_i, y_i) = 0 \Rightarrow$ мы нашли $\frac{n}{2} \geq 2K$ корней мн-на степени $\leq 2(k-1) \Rightarrow R(x) \equiv 0$. ▶

II Задача: групповое тестирование (group testing)

Пример Даны N человек, $s \ll N$ из которых заражены.
Задача: выявить заражённых за min. число тестов

Тривиальное решение сделать N тестов может быть дорогим.

Можно ли смешать образцы крови так, чтобы можно было выявить больных / здоровых за меньшее число тестов

Идея

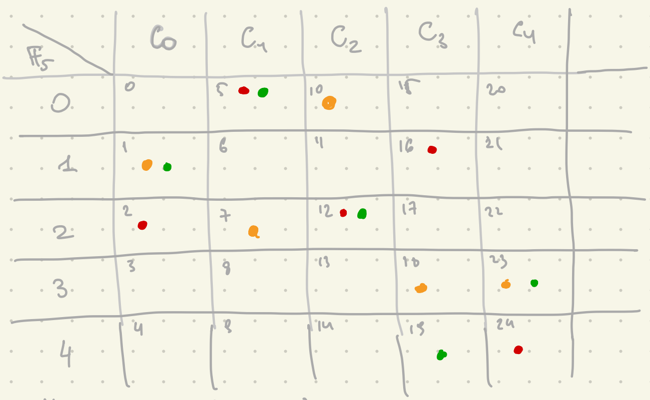
- Ассоциируем с каждым человеком кодовое слово $c \in RS_{\mathbb{F}_1 S}[n,k]$
- имеем $n \cdot |\mathbb{F}|$ колб / тестов, сформированных в матрицу $|\mathbb{F}| \times n$, где столбцы пронумерованы $\{c_i\}_i$
- образцы крови человека, ассоциированным с кодовым словом c, помещается в колбы $(c[i], c_i)$
- $d = n-k+1$ – min. расстояние RS $\Rightarrow \forall$ два c_i, c_j отличаются на как min d позициях.

Пример $RS_{\mathbb{F}_5}$, $n=5$, $k=3$, $d=3$, $s \le 2$; $|RS_{\mathbb{F}_5}| = 5^3 = 125 \Rightarrow$ макс. 125 человек

$c^* = (c_0^*, c_1^*, c_2^*, c_3^*, c_4^*) = (2, 0, 2, 1, 4)$ – заражённый №1

$c^* = (1, 2, 0, 3, 3)$ – заражённый №2

$c = (1, 0, 2, 4, 3)$ – здоровый



Клетка = колба (всего 25)

Если $s = 1 \Rightarrow \exists$ 3 негативных теста для $\forall$ здорового;

Если $s = 2 \Rightarrow \exists$ 1 негативный тест — // —

Замечание Можно показать, что для $1 \le s \ll N$, такое групповое тестирование работает корректно, если число тестов удовлетворяет

$$T = O\left(s^2 \cdot \left(\frac{\lg N}{\lg s}\right)^2\right)$$