16

Лекция 22–23 декабря

Сжатие и границы

Константин Чухарев

Дискретная математика — Осень 2026

Код Хаффмана и границы

> *Основная задача связи состоит в точном или приближенном воспроизведении в некотором месте сообщения, выбранного для передачи в другом месте.*
>
> ***— Клод Шеннон***

Защита и сжатие

Код Хэмминга защищает от ошибок, добавляя избыточность.

Код Хаффмана решает обратную задачу — сжать данные до минимального размера без потерь.

> Частым символам — короткие кодовые слова, редким — длинные.

Азбука Морзе делает то же самое: буква E — одна точка, а Q — четыре знака.

Префиксный код

Определение 1 (Префиксный код)

Код называется **префиксным**, если ни одно кодовое слово не является началом другого.

Пример

Код $\{0, 10, 11\}$ префиксный: ни одно слово не начало другого.

Код $\{0, 01, 011\}$ нет: слово 0 может продолжиться как 01 или как 011 — декодер не знает, кончилось ли слово.

Декодирование без задержки

Пример

Слова $A = 0$, $B = 10$, $C = 110$, $D = 1110$, $E = 1111$.

Поток `1101111` читается слева направо: `110` — это C, сразу за ним `1111` — это E.

Другого разбиения нет: конец слова виден в момент прочтения.

Префиксность устраняет неоднозначность: декодер не возвращается и не ждёт.

Алгоритм Хаффмана

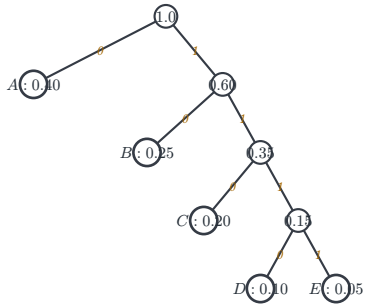
Алгоритм жадный и строит дерево снизу вверх.

Пример

Частоты: $A = 0.40$, $B = 0.25$, $C = 0.20$, $D = 0.10$, $E = 0.05$.

1. Сливаем два наименьших веса: D и E — узел 0.15.
2. Дальше C и DE — узел 0.35.
3. Дальше B и CDE — узел 0.60.
4. Наконец A и BCDE — корень 1.00.

Дерево Хаффмана



Путь от корня к листу — кодовое слово: левый переход даёт 0, правый — 1.

Кодовые слова: $A = 0$, $B = 10$, $C = 110$, $D = 1110$, $E = 1111$.

Средняя длина

Определение 2 (Средняя длина кода)

$$L = \sum_i f_i l_i,$$

где f_i — частота символа, l_i — длина его слова.

Пример

$$L = 0.40 \cdot 1 + 0.25 \cdot 2 + 0.20 \cdot 3 + 0.10 \cdot 4 + 0.05 \cdot 4 = 2.10$$

бита на символ — против $\lceil \log_2 5 \rceil = 3$ битов у равномерного кода на пять символов.

Оптимальность

Теорема 1 (Оптимальность кода Хаффмана)

Для заданных частот код Хаффмана имеет минимальную среднюю длину среди всех префиксных кодов.

- Идея доказательства — обменный аргумент: два редчайших символа можно обменять с любыми соседними листьями максимальной глубины, и средняя длина не вырастет.
- Код Хаффмана не единственный: паре сливаемых поддеревьев нули и единицы раздаются свободно, а при равных весах оптимальных деревьев несколько.

Словарное сжатие: LZ77 и LZ78

Хаффман видит только частоты, а главный резерв реальных данных — повторения.

Словарное сжатие заменяет повтор ссылкой: канал передаёт адрес, а не копию.

Определение 3 (Тройка LZ77)

Кодер ищет совпадения в *скользящем окне* — уже закодированном хвосте входа. Повтор кодируется тройкой (o, l, a): смещение назад, длина совпадения, первый несовпавший символ.

LZ78 копит *словарь фраз* вместо окна: выход — пары из номера фразы и нового символа.

DEFLATE (gzip, zip, PNG) — LZ77 плюс код Хаффмана: LZ выучивает повторения, Хаффман ужимает остаток.

Сколько кодов вообще возможно?

Исправление ошибок — это упаковка шаров в двоичном кубе.

Теорема 2 (Граница Хэмминга)

Если код исправляет t ошибок, то

$$|C| \cdot \sum_{i=0}^{t} \binom{n}{i} \leq 2^n.$$

Объём шара радиуса t — число строк на расстоянии не больше t: $\sum_{i=0}^{t} \binom{n}{i}$.

Доказательство

Шары радиуса t вокруг кодовых слов не пересекаются, и все они лежат в кубе из 2^n строк.

Совершенный код

Пример

Для кода Хэмминга $(7, 4)$ объём шара радиуса 1 равен $1 + 7 = 8$.

$$16 \cdot 8 = 128 = 2^7.$$

Шары покрывают всё пространство без зазоров.

Код, достигающий границы Хэмминга, называется **совершенным**.

Код Хэмминга $(7, 4)$ — совершенный.

Энтропия

Определение 4 (Энтропия)

Для символов с вероятностями $p_1, ..., p_n$:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i.$$

Пример

Два равновероятных символа: $H = 1$ бит.

Достоверный символ ($p = 1$): $H = 0$ — сообщать не о чем.

Энтропия — минимум среднего числа битов на символ: любой префиксный код имеет $L \geq H$.

Пропускная способность канала

Канал передаёт биты и с вероятностью p переворачивает каждый независимо — двоичный симметричный канал.

Определение 5 (Пропускная способность)

$$C = 1 - H(p) = 1 + p\log_2 p + (1-p)\log_2(1-p).$$

p	C	Канал
0	1	без ошибок
0.1	0.53	умеренный шум
0.2	0.28	шум заметен
0.5	0	бесполезен

При $p = 0.5$ принятый бит не зависит от переданного — канал не несёт ничего.

Теорема Шеннона

Теорема 3 (Теорема Шеннона о кодировании (1948))

При любой скорости $R < C$ существуют коды с вероятностью ошибки, стремящейся к нулю.

При любой скорости $R > C$ надёжная передача невозможна.

Доказательство неконструктивно: случайный код большой длины хорош с высокой вероятностью, но из доказательства не извлечь сам код.

Построение явных кодов у границы — программа на семь десятилетий: Рид—Соломон, турбо-коды, LDPC, полярные коды.

Сжать, потом защитить

Защита и сжатие двигают одну величину — долю полезного в переданном.

Данные сжимают кодом Хаффмана, затем защищают кодом Хэмминга.

Сжатие убирает лишнюю избыточность, канальный код добавляет её назад — дозированно и там, где её ждёт декодер.

Обзор: циклические коды

Определение 6 (Циклический код)

Линейный код, замкнутый относительно циклического сдвига:

$$(c_0, ..., c_{n-1}) \to (c_{n-1}, c_0, ..., c_{n-2}).$$

Слово — многочлен $c(x) = c_0 + c_1 x + ... + c_{n-1} x^{n-1}$, а сдвиг — умножение на x по модулю $x^n - 1$.

Кодовые слова — в точности кратные порождающего многочлена $g(x)$, который делит $x^n - 1$.

Деление с остатком

Пример

Для кода $(7, 4)$: $g(x) = 1 + x + x^3$.

Проверка над $\mathrm{GF}(2)$:

$$x^7 - 1 = (1 + x + x^3)(1 + x + x^2 + x^4).$$

Удвоенные члены обнуляются: $x^i + x^i = 0$.

Кодирование: $c(x) = x^{n-k} m(x) - r(x)$, где $r(x)$ — остаток от деления на $g(x)$.

Декодирование — то же деление: ненулевой остаток — сигнал об ошибке.

CRC в Ethernet

1. Отправитель вычисляет остаток $r(x)$ от деления данных на $g(x)$ и дописывает его в конец пакета.
2. Получатель делит принятое на $g(x)$: ненулевой остаток — ошибка.
3. CRC-32 живёт в Ethernet, PNG и gzip.

CRC обнаруживает, но не исправляет: повреждённый пакет отбрасывается, а доставку берёт на себя транспортный уровень.

Обзор: коды Рида—Соломона

Определение 7 (Код Рида—Соломона)

Над полем $\mathrm{GF}(2^8)$ сообщение $(m_0, \ldots, m_{k-1})$ задаёт многочлен степени меньше k.

Кодовое слово — значения этого многочлена в $n = 255$ точках поля.

Многочлен степени меньше k однозначно восстанавливается по любым k значениям — это интерполяция.

Кодовое расстояние $d = n - k + 1$: два разных многочлена совпадают не более чем в $k - 1$ точках.

Маленький РС-код

Пример

$\mathrm{RS}(7, 3, 5)$ над $\mathrm{GF}(8)$, где $8 = 2^3$.

Три символа кодируются семью, расстояние $d = 7 - 3 + 1 = 5$.

Исправляет до $\frac{7-3}{2} = 2$ символьных ошибок, хотя каждый символ — 3 бита.

Две символьные ошибки — до шести испорченных битов: алгебра многочлена чинит их одной операцией над символами.

Пакетные ошибки

Царапина портит биты подряд.

Код Рида—Соломона считает ошибки символами: байт испорчен целиком или нет, и до $\frac{n-k}{2}$ испорченных символов исправляются.

Пример

RS$(255, 239)$: скорость $R = \frac{239}{255} \approx 0.94$.

Избыточность всего 6% — и до 8 байтовых ошибок исправляются.

Где живут РС-коды?

- QR-коды — RS над $\mathrm{GF}(256)$ с уровнями коррекции от 7% до 30% повреждённой площади.
- CD и DVD — каскад CIRC: два слоя RS с перемежением против царапин и пыли.
- Цифровое телевидение DVB — $\mathrm{RS}(204, 188)$: до 8 байтовых ошибок на блок.
- RAID 6 — два проверочных диска: данные выживают при отказе двух дисков разом.

Лекция 16: что мы поняли?

- Код Хаффмана строит оптимальный префиксный код: частым символам — короткие слова.
- Префиксность даёт однозначное декодирование без задержки.
- Словарное сжатие LZ заменяет повторения ссылками: DEFLATE — это LZ77 плюс Хаффман.
- Граница Хэмминга ограничивает число кодов упаковкой шаров, и код $(7, 4)$ — совершенный.
- Энтропия H — минимум битов на символ, а теорема Шеннона делит скорости на возможные и невозможные.
- Циклические коды — алгебра многочленов: CRC в Ethernet обнаруживает ошибки делением на $g(x)$.
- Рид—Соломон работает с символами-байтами и чинит пакетные ошибки: QR-коды, диски, RAID.

Вопросы для самопроверки

1. Постройте код Хаффмана для частот $(0.4, 0.3, 0.2, 0.1)$ и посчитайте среднюю длину.
2. Почему код $\{0, 10, 110, 111\}$ префиксный, а $\{0, 01, 011\}$ — нет?
3. Сожмите «ABRACADABRA» тройками LZ77 с окном в весь уже закодированный префикс. Какая тройка кодирует финальное «ABRA»?
4. Посчитайте энтропию распределения $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$.
5. Оцените границей Хэмминга число слов кода длины 9, исправляющего две ошибки.
6. Проверьте, что $g(x) = 1 + x + x^3$ делит $x^7 - 1$ над $\mathrm{GF}(2)$.
7. Сколько байтовых ошибок исправляет $\mathrm{RS}(255, 239)$? Почему он хорош против царапины на диске?

Чтение и экзамен

- m13 — «Код Хаффмана», «Словарное сжатие: LZ77 и LZ78», «Граница Шеннона», «Циклические коды», «Коды Рида—Соломона».
- Последние занятия — 22–23 декабря.
- Экзамен за семестр 1 — в зимнюю сессию (январь). Даты задаёт расписание вуза.

Семестр 1: что мы прошли?

- Логика и множества — язык, на котором говорят остальные темы.
- Отношения, порядок, функции, счётность — словарь структур и первая встреча с бесконечностью.
- Формальная логика — вывод и семантика: доказательства как объекты, силлогизмы как системы.
- Булева алгебра и схемы — от формул к железу: ДНФ, минимизация, верификация через SAT.
- Коды — избыточность как математика: расстояние, код Хэмминга, код Хаффмана, границы Шеннона.
- Семестр 2: графы, автоматы, машины Тьюринга, комбинаторика и остальное — с февраля.