# „Теория информации“. Лекции 1-4

А.В. Смаль

16 марта 2017 г.

## Комбинаторный подход

### Информация по Хартли

Пусть задано некоторое конечное множество $A$ — *множество исходов*.

**Определение 1** (1928)**.** Определим *количество информации в* $A$ как $\chi(A) = \log_2 |A|$ (мы будем измерять количество информации в битах, поэтому все логарифмы будут по основанию 2, для байтов основание нужно было бы заменить на 256).

Если про некоторый $x \in A$ стало известно, что $x \in B$, то теперь для идентификации $x$ нам достаточно $\chi(A \cap B) = \log |A \cap B|$ битов, т.е. нам сообщили $\chi(A) - \chi(A \cap B)$ битов информации.

*Пример* 1.1. Предположим, что мы хотим узнать некоторое неизвестное упорядочение множества $\{a_1, a_2, \ldots, a_5\}$. Нам стало известно, что $a_1 > a_2$ или $a_3 > a_4$. Сколько битов информации мы узнали? Множество $A$ состоит из 5! перестановок, множество $B$ — из перестановок, которые удовлетворяют новому условию. Легко проверить, что $|B| = 90$. Итого мы узнали $\log 120 - \log 90 = \log(4/3)$ битов.

Пусть $A \subset \{0,1\}^* \times \{0,1\}^*$. Обозначим через $\pi_1(A)$ и $\pi_2(A)$ проекции множества $A$ на первую и вторую координату соответственно, а $\chi_1(A) = \log |\pi_1(A)|$ и $\chi_2(A) = \log |\pi_2(A)|$ — их сложность по Хартли.

**Теорема 1.1.** $\chi(A) \leq \chi_1(A) + \chi_2(A)$.

**Определение 2.** Количество информации в второй компоненте $A$ при известной первой

$$\chi_{2|1} = \log \left( \max_{a \in \pi_1(A)} |A_a| \right),$$

где $A_a = \{(a, x) \mid x \in \pi_2(A)\}$.

**Теорема 1.2.** $\chi(A) \leq \chi_1(A) + \chi_{2|1}(A)$.

**Теорема 1.3.** *Для* $A \subset \{0,1\}^* \times \{0,1\}^* \times \{0,1\}^*$

$$2 \cdot \chi(A) \le \chi_{12}(A) + \chi_{13}(A) + \chi_{23}(A).$$

**Следствие 1.1.** *Квадрат объёма трёхмерного тела не превосходит произведение площадей его проекций на координатные плоскости.*

**Утверждение 1.1.** *Если* $f : X \to Y$

1. *является сюръекцией, то* $\chi(Y) \le \chi(X)$,
2. *является инъекцией, то* $\chi(X) \le \chi(Y)$.

### Применение: игра в 10 вопросов

Сколько вопросов на ДА/НЕТ нужно задать, чтобы определить загаданное число от 1 до $N$, если (а) можно задавать вопросы адаптивно; (б) вопросы нужно написать на бумажке заранее.

Оценка $\lceil \log N \rceil$ достигается в обоих случаях, если задавать вопросы про биты двоичного представления загаданного числа.

Докажем нижнюю оценку. Пусть $A = [N]$. Множество $Q = \{(q_1, q_2, \dots, q_k)\}$ — множество протоколов (ответы на вопросы). Можно рассматривать $A$ и $Q$ как проекции некоторого множества исходов игры $S$ на разные координаты. Тогда верны следующие неравенства:

- $\chi_Q(S) = \chi(Q) \le \chi_1(Q) + \chi_2(Q) + \dots + \chi_k(Q) \le k$,
- $\chi_A(S) = \chi(A) \le \chi(S) \le \chi_Q(S) + \chi_{A|Q}(S) \le k + 0 = k$.

Таким образом получаем, что $\log N = \chi(A) \le k$.

### Цена информации

Пусть имеется некоторое неизвестное число от 1 до $n$ (где $n \ge 2$). Разрешается задавать любые вопросы с ответами ДА/НЕТ. При ответе ДА мы заплатим 1 рубль, а при ответе НЕТ — два рубля. Сколько необходимо и достаточно заплатить для отгадывания числа?

**Верхняя оценка.** Давайте задавать вопросы так, чтобы отрицательные ответы приносили в два раза больше информации, чем положительные. Тогда за каждый бит информации мы заплатим $c \log n$ для некоторой константы $c$. Пусть вопросы будут вида „$x \in T$?“. Тогда требуется

$$2(\log |X| - \log |X \cap T|) = \log |X| - \log |X \cap \overline{T}|.$$

Пусть $|X \cap T| = \alpha |X|$, тогда $|X \cap \overline{T}| = (1-\alpha)|X|$, т.о. $\alpha^2 = 1 - \alpha$, $\alpha = (\sqrt{5}-1)/2$. При любом ответе мы заплатим $c = 1/(-\log \alpha) \approx 1.44$ рублей за бит, а в целом — $\log n/(-\log \alpha)$ рублей.

**Нижняя оценка.** Применим рассуждение про злонамеренного противника (adversary argument). Пусть противник выбирает ответ ДА/НЕТ в зависимости от того, какое из двух значение $1/(\log|X| - \log|X \cap T|)$ и $(2/\log|X| - \log|X \cap \overline{T}|)$ больше. При любых $X$, $T$ одно из этих значений не меньше $c = 1/(-\log\alpha)$. Таким образом мы заставляем алгоритм платить не менее $c$ рублей за бит, а значит любой алгоритм в худшем случае заплатит $\lceil c \log n \rceil$ рублей.

### Применение: упорядочивание камней по весу

### Верхняя и нижняя оценки для произвольного $N$

Сколько сравнений нужно сделать для того, чтобы упорядочить $N$ камней по весу?

**Нижняя оценка.** Потребуется $\lceil \chi(S_N) \rceil = \lceil \log n! \rceil$ сравнений.

**Верхняя оценка.** Будем сортировать вставкой с бинарным поиском места вставки. Количество сравнений:

$$\lceil \log 2 \rceil + \lceil \log 3 \rceil + \cdots + \lceil \log n \rceil \le \log n! + n - 1 = n \log n + O(n).$$

### Точные оценки для маленьких $N$

*Упражнение* 1.1. Сколько нужно взвешиваний, чтобы упорядочить $N$ камней по весу? Найдите точный ответ на этот вопрос для $N = 2, 3, 4, 5$. Указание: воспользуйтесь жадной стратегией, при которой каждое взвешивание приносит максимум информации.

### Применение: поиск фальшивой монетки

- 20 монет, одна фальшивая легче остальных.

  Каждое взвешивание даёт не более $\log 3$ битов. Итого $k \ge \log N / \log 3 = \log_3 N$.

- 13 монет, одна фальшивая (с неизвестным относительным весом), 3 взвешивания.

  Два варианта первого шага:

  - если взвешиваем по 4, то при равенстве нельзя из 5 за два взвешивания найти фальшивую (остаётся 10 исходов),
  - если взвешиваем по 5, то при неравенстве остаётся 10 возможных исходов.

- 15 монет, одна фальшивая, три взвешивания. Не требуется узнавать относительный вес монеты.

  Всего исходов $2 \cdot 14 + 1 > 27$, т.к. только в случае трёх равенств мы можем не узнать относительный вес фальшивой монеты.

- 14 монет, одна фальшивая, три взвешивания. Не требуется узнавать относительный вес монеты.

  Всего исходов $2 \cdot 13 + 1 \leq 27$, но определить тем не менее нельзя. Аппарата информации по Хартли недостаточно.

### Логика знаний

В этом разделе мы будем называть множество исходов $A$ множеством *миров*. Пусть $f$ — это некоторая функция из $A$ в некоторое множество $I$ (будем воспринимать это как информация о мире). Нам не важно какие значения принимает $f$, нам будут важны лишь классы эквивалентности, на которые $f$ разбивает $A$: каждый класс эквивалентности будет состоять из миров $A$ с одинаковым значением $f$.

*Пример* 1.2. Пусть $A = \{1, 2, 3, 4, 5\}$, а $f(x) = x \bmod 3$. Тогда $f$ разбивает $A$ на три класса эквивалентности $\{1, 4\}$, $\{2, 5\}$ и $\{3\}$.

Пусть $B \subset A$ — это некоторое *утверждение* о мирах. $B$ *истинно* в мире $x$, если $x \in B$. В противном случае $B$ *ложно* в $x$. В мире $x$ мы *знаем, что* $B$ *истинно*, если $y \in B$ для всех $y \sim x$.

*Пример* 1.3. Пусть $A = \{1, 2, 3, 4, 5\}$, а $f(x) = x \bmod 3$. Тогда в мирах 1, 4 и 3 мы знаем, что мир меньше 5. А в мирах 2 и 5 — не знаем.

*Замечание* 1.1. „Не знаем“ мы будем понимать в смысле „не верно, что знаем“.

К утверждениям о мирах можно применять обычные логические связки: «И» (пересечение), «ИЛИ» (объединение), «НЕ» (дополнение).

**Утверждение 1.2.** *Если в мире $x$ мы знаем $B$, то в мире $x$ мы знаем, что мы знаем $B$. Аналогично, если в мире $x$ мы не знаем $B$, то в мире $x$ мы знаем, что не знаем $B$.*

Пусть теперь у нас есть $k$ человек со своими знаниями о мире. Они определяют $k$ отношений эквивалентности $\sim_1, \sim_2, \ldots, \sim_k$ и, соответственно, $k$ разбиений на классы эквивалентности.

*Пример* 1.4. Пусть множество миров $A = \{1, 2, 3, 4, 5\}$ и есть два человека, Алиса и Боб. Алиса знает значения $f_A(x) = x \bmod 3$, а Боб знает $f_B(x) = x \bmod 2$. Тогда классы эквивалентности Алисы: $\{1, 4\}$, $\{2, 5\}$ и $\{3\}$, а классы эквивалентности Боба: $\{1, 3, 5\}$ и $\{2, 4\}$. В мире 1 Алиса знает, что мир меньше 5, а Боб не знает. В мире 4 они оба это знают. В мире 1 Алиса не знает, что Боря не знает, что мир меньше 5 (действительно, в мире 4, который с точки зрения Алисы эквивалентен 1, Боря это знает).

## Вероятностный подход

### Энтропия Ше́ннона

Энтропия Шеннона определяет количество информации $H(\alpha)$ в распределении вероятностей для некоторой случайной величины $\alpha$. Пусть $\alpha$ принимает значения из множества $\{a_1, a_2, \ldots, a_k\}$ с вероятностями $\{p_1, p_2, \ldots, p_k\}$, $p_i \geq 0$, $\sum_i p_i = 1$.

Нам бы хотелось, чтобы это определение согласовывалось с определением Хартли, т.е. имеют место следующие „граничные условия“:

- если $p_1 = \cdots = p_k$, то $H(\alpha) = \log k$,
- если $p_1 = 1$, $p_2 = \cdots = p_k = 0$, то $H(\alpha) = 0$.

Будем искать $H(\alpha)$ в виде математического ожидания „удивления“ от исхода случайной величины („удивление“ зависит от вероятности данного исхода).

$$H(\alpha) = \sum_i p_i \cdot \mathrm{impress}(p_i).$$

Граничные условия однозначно определяют функцию $\mathrm{impress}(p_i) = \log \frac{1}{p_i} = -\log p_i$.

**Определение 3** (1948)**.** Энтропия Шеннона случайной величины $\alpha$

$$H(\alpha) = \sum_{i=1}^{k} p_i \cdot \log \frac{1}{p_i}.$$

(По непрерывности доопределим $0 \cdot \log \frac{1}{0} = 0$.)

Можно вывести это соотношение из определения Хартли более формально. Пусть $W_n$ — это множество всех слов длины $n$ состоящих из букв $\{a_1, a_2, \ldots, a_k\}$, где каждая буква $a_i$ встречается ровно $n_i = p_i \cdot n$ раз (будем считать, что вероятности $p_i$ рациональны, и что множество $W_n$ определено только тогда, когда все $n_i$ целые). Информация по Хартли в $W_n$

$$\chi(W_n) = \log |W_n| = \log \frac{n!}{n_1! n_2! \cdots n_k!}.$$

Это выражение можно оценить при помощи формулы Стирлинга.

$$\begin{aligned}
\chi(W_n) &= \log \frac{\mathrm{poly}(n) \cdot (n/e)^n}{\mathrm{poly}(n) \cdot (n_1/e)^{n_1} \cdot (n_1/e)^{n_2} \cdots (n_k/e)^{n_k}} = \\
&= \log \left( \left(\frac{n}{n_1}\right)^{n_1} \cdot \left(\frac{n}{n_2}\right)^{n_2} \cdots \left(\frac{n}{n_k}\right)^{n_k} \right) + O(\log n) = \\
&= \log \left( \left(\frac{1}{p_1}\right)^{p_1 \cdot n} \cdot \left(\frac{1}{p_2}\right)^{p_2 \cdot n} \cdots \left(\frac{1}{p_k}\right)^{p_k \cdot n} \right) + O(\log n) = \\
&= n \cdot \sum_{i=1}^{k} p_i \cdot \log \frac{1}{p_i} + O(\log n).
\end{aligned}$$

В среднем на один символ приходится $\chi(W_n)/n$ битов информации. В пределе получаем

$$\lim_{n \to \infty} \frac{\chi(W_n)}{n} = \sum_{i=1}^{k} p_i \cdot \log \frac{1}{p_i} = H(\alpha)$$

(предел нужно брать по бесконечной подпоследовательности натуральных чисел $n$ таких, для которых все $\{n_i\}$ — целые).

**Лемма 2.1.** *Для энтропии Шеннона выполняются следующие соотношения.*

- $H(\alpha) \geq 0$, *причём* $H(\alpha) = 0 \iff$ *распределение* $\alpha$ *вырождено.*
- $H(\alpha) \leq \log k$, *причём* $H(\alpha) = \log k \iff$ *величина* $\alpha$ *распределена равномерно.*

Для доказательства нам потребуется следующая теорема.

**Теорема 2.1** (Неравенство Йенсена)**.** *Пусть функция* $f(x)$ *является вогнутой на некотором промежутке* $\mathcal{X}$ *и числа* $q_1, q_2, \ldots, q_n > 0$ *таковы, что* $q_1 + \ldots + q_n = 1$. *Тогда для любых* $x_1, x_2, \ldots, x_n$ *из промежутка* $\mathcal{X}$ *выполняется неравенство:*

$$\sum_{i=1}^{n} q_i f(x_i) \leq f\left(\sum_{i=1}^{n} q_i x_i\right).$$

*Доказательство леммы 2.1.* Первое свойство следует напрямую из определения: каждый член суммы $H(\alpha)$ неотрицателен и равен нулю только в случае, если $p_i = 0$ или $p_i = 1$.

Для доказательства второго неравенства перенесём всё в левую часть и применим неравенство Йенсена:

$$H(\alpha) - \log k = \sum_{i=1}^{k} p_k \cdot \log \frac{1}{p_i} - \sum_{i=1}^{k} p_i \cdot \log k = \sum_{i=1}^{k} p_k \cdot \log \frac{1}{p_i k} \leq \log\left(\sum_{i=1}^{k} p_i \frac{1}{p_i k}\right) = \log 1 = 0.$$

□

Энтропию совместного распределения пары случайных величин $\alpha$ и $\beta$ будем обозначать $H(\alpha, \beta)$.

**Лемма 2.2.** *Выполняются следующие свойства:*

- $H(\alpha, \beta) \leq H(\alpha) + H(\beta)$, *причём равенство достигается тогда и только тогда, когда случайные величины независимы;*
- $H(\alpha) \leq H(\alpha, \beta)$, *причём равенство достигается тогда и только тогда, когда* $\beta$ *полностью определяется значением* $\alpha$, *т.е.* $\beta = f(\alpha)$.

*Доказательство.* Введём обозначения для вероятностей событий совместного распределения вероятностей $(\alpha, \beta)$. Пусть пара $(a_i, b_j)$ имеет вероятность $p_{i,j}$, событие $[\alpha = a_i]$ имеет вероятность $p_{i,*} = p_{i,1} + \cdots + p_{i,n}$, а событие $[\beta = b_j]$ — вероятность $p_{*,j} = p_{1,j} + \cdots + p_{k,j}$. В этих обозначениях неравенство $H(\alpha, \beta) \leq H(\alpha) + H(\beta)$ переписывается как

$$\sum_{i,j} \cdot p_{i,j} \cdot \log \frac{1}{p_{i,j}} \leq \sum_i \sum_j p_{i,j} \cdot \log \frac{1}{p_{i,*}} + \sum_j \sum_i p_{i,j} \cdot \log \frac{1}{p_{*,j}}.$$

Перенесём всё в левую часть и применим неравенство Йенсена.

$$\sum_{i,j} p_{i,j} \cdot \log \frac{p_{i,*} \cdot p_{*,j}}{p_{i,j}} \leq \log \left( \sum_{i,j} p_{i,j} \cdot \frac{p_{i,*} \cdot p_{*,j}}{p_{i,j}} \right) = \log \left( \sum_{i,j} p_{i,*} \cdot p_{*,j} \right) =$$

$$= \log \underbrace{\left( \sum_{i} p_{i,*} \right)}_{1} \cdot \underbrace{\left( \sum_{j} p_{*,j} \right)}_{1} = 0.$$

Равенство в неравенстве Йенсена для $f(x) = \log(x)$ достигается только, если все точки равны, т.е. для любых $i, j$ $\frac{p_{i,*}p_{*,j}}{p_{i,j}} = c$ для некоторой константы $c$. Несложно заметить, что $c = 1$, т.к. выполняется следующее равенство $\sum_{i,j} p_{i,*}p_{*,j} = c \sum_{i,j} p_{i,j}$ в котором обе суммы равны 1. Таким образом в случае равенства $\alpha$ и $\beta$ независимы.

Доказательство второго свойства мы получим как следствие из свойств условной энтропии. □

**Определение 4.** Энтропия $\alpha$ при условии $\beta = b_j$

$$H(\alpha \mid \beta = b_j) = \sum_i \Pr[\alpha = a_i \mid \beta = b_j] \cdot \log \frac{1}{\Pr[\alpha = a_i \mid \beta = b_j]}.$$

**Определение 5.** *Условная (относительная) энтропия* $\alpha$ относительно $\beta$

$$H(\alpha \mid \beta) = \sum_j \Pr[\beta = b_j] \cdot H(\alpha \mid \beta = b_j).$$

Другими словами

$$H(\alpha \mid \beta) = \mathop{\mathrm{E}}_{b_j \leftarrow \beta}[H(\alpha \mid \beta = b_j)].$$

Если подставить определение 4, то можно получить выражение для условной энтропии через отдельные вероятности событий.

$$H(\alpha \mid \beta) = \sum_j \Pr[\beta = b_j] \cdot \sum_i \Pr[\alpha = a_i \mid \beta = b_j] \cdot \log \frac{1}{\Pr[\alpha = a_i \mid \beta = b_j]} = \sum_{i,j} p_{i,j} \cdot \log \frac{p_{*,j}}{p_{i,j}}.$$

**Лемма 2.3.** *Условная энтропия обладает следующими свойствами.*

- $H(\alpha \mid \beta) \geq 0$.
- $H(\alpha \mid \beta) = 0 \iff$ *$\alpha$ однозначно определяется по $\beta$.*
- $H(\alpha, \beta) = H(\beta) + H(\alpha \mid \beta) = H(\alpha) + H(\beta \mid \alpha)$.

*Доказательство.* Первое свойство выполняется, т.к. условная энтропия это матожидание неотрицательной случайной величины. Второе свойство объясняется тем, что для любого $j$ распределение $\langle \alpha \mid \beta = b_j \rangle$ имеет нулевую энтропию, т.е. распределение вырождено и каждому $b_j$ соответствует ровно один $a_i$. Третье свойство следует из следующего равенства.

$$\sum_{i,j} p_{i,j} \cdot \log \frac{1}{p_{i,j}} = \sum_{i,j} p_{i,j} \cdot \log \frac{1}{p_{*,j}} + \sum_{i,j} p_{i,j} \cdot \log \frac{p_{*,j}}{p_{i,j}}.$$

(Нужна аккуратность, если есть строки, которые состоят из одних нулей, т.е $p_{*,j} = 0$ — такие строки не нужно включать в эти суммы.) □

**Следствие 2.1.** *$H(\alpha, \beta) \geq H(\alpha)$, причём равенство достигается тогда и только тогда, когда $\beta = f(\alpha)$.*

*Доказательство.* $H(\alpha, \beta) - H(\alpha) = H(\beta \mid \alpha) \geq 0$. По второму свойству условной энтропии равенство достигается тогда и только тогда, когда $\beta = f(\alpha)$. □

### Взаимная информация

**Определение 6.** *Информация в $\alpha$ о величине $\beta$* определяется следующим соотношением:

$$I(\alpha : \beta) = H(\beta) - H(\beta \mid \alpha).$$

Эту величину так же *называют взаимной информацией случайных величин $\alpha$ и $\beta$.*

**Лемма 2.4.** *Для взаимной информации выполняются следующие соотношения.*

1. $I(\alpha : \beta) \leq H(\alpha)$.
2. $I(\alpha : \beta) \leq H(\beta)$.
3. $I(\alpha : \alpha) = H(\alpha)$.
4. $I(\alpha : \beta) = I(\beta : \alpha)$.
5. $I(\alpha : \beta) = H(\alpha) + H(\beta) - H(\alpha, \beta)$.

**Определение 7.** Пусть $\alpha, \beta, \gamma$ — случайные величины. Определим *взаимную информацию в $\alpha$ о $\beta$ при условии $\gamma$.*

1. $I(\alpha : \beta \mid \gamma) = H(\beta \mid \gamma) - H(\beta \mid \alpha, \gamma)$.
2. $I(\alpha : \beta \mid \gamma) = \sum_\ell I(\alpha : \beta \mid \gamma = c_\ell) \cdot \Pr[\gamma = c_\ell]$.
3. $I(\alpha : \beta \mid \gamma) = H(\alpha \mid \gamma) + H(\beta \mid \gamma) - H(\alpha, \beta \mid \gamma)$.
4. $I(\alpha : \beta \mid \gamma) = H(\alpha, \gamma) + H(\beta, \gamma) - H(\alpha, \beta, \gamma) - H(\gamma)$.

**Лемма 2.5.** *Все определения условной взаимной информации эквивалентны.*

*Доказательство.* (3) $\iff$ (4).

$$(3) = H(\alpha \mid \gamma)+H(\beta \mid \gamma)-H(\alpha,\beta \mid \gamma) = H(\alpha,\gamma)-H(\gamma)+H(\beta,\gamma)-H(\gamma)-H(\alpha,\beta,\gamma)+H(\gamma).$$

□

### Применение: опять о поиске фальшивой монетки

Теперь у нас достаточно знаний, чтобы доказать, что за три взвешивания нельзя найти одну фальшивую монету из 14, даже если не нужно определять её относительный вес.

*Доказательство.* Предположим, что существует способ найти фальшивую монету за три взвешивания. Тогда протокол взвешивания можно представить в виде полного троичного дерева, где каждый лист помечен номером монетки, которая оказалась фальшивой (у нас как раз ровно $3^3 = 27$ исходов).

Давайте введём следующее распределение вероятностей $\alpha$. Пусть монета, номер которой находится в листе, соответствующем трём равенствам (такой лист только один), имеет номер $i$. В нашем распределении вероятностей монета с номером $i$ будет фальшивой с вероятностью $1/27$. Оставшиеся монеты оказываются фальшивыми с вероятностями $2/27$, причём с вероятностью $1/27$ монета оказывается легче, чем настоящая, и с такой же вероятностью она оказывается тяжелее настоящей.

$$H(\alpha) = \log 27 = 3 \log 3.$$

Пусть случайные величины $\beta_1$, $\beta_2$, $\beta_3$ соответствуют результатам первого, второго и третьего взвешивания соответственно. Значение $\alpha$ однозначно определяется после трёх взвешиваний: $H(\alpha \mid \beta_1, \beta_2, \beta_3) = 0$, а следовательно

$$H(\alpha) \leq H(\beta_1, \beta_2, \beta_3) \leq H(\beta_1) + H(\beta_2) + H(\beta_3) \leq 3 \log 3.$$

Таким образом каждое взвешивание должно иметь энтропию ровно $\log 3$. Рассмотрим первое взвешивание. Пусть на чашах весов лежит по $k$ монет. Вероятность каждого исхода взвешивания ($<$, $>$, $=$) относительно распределения $\alpha$ должна быть ровно $1/3$.

$$\Pr[<] = \frac{k}{27} + \frac{k}{27} = \frac{1}{3}.$$

Таким образом $2k = 9$, а значит нет такого целого $k$. □

## Кодирование

### Однозначно декодируемые коды

**Определение 8.** Будем называть *кодом* функцию $C : \{a_1, a_2, \ldots, a_n\} \to \{0,1\}^*$, сопоставляющую буквам некоторого алфавита *кодовые слова*. Если любое сообщение, которое получено применением кода $C$, декодируется однозначно (т.е. только единственным образом разрезается на образы $C$), то такой код называется *однозначно декодируемым*.

**Определение 9.** Код называется *префиксным (беспрефиксным, prefix-free)*, если никакое кодовое слово не является префиксом другого кодового слова.

**Теорема 3.1** (Неравенство Крафта-Макмилана)**.** *Для любого однозначно декодируемого кода со множеством кодовых слов* $\{c_1, c_2, \ldots, c_n\}$ *выполняется следующее неравенство:*

$$\sum_{i=1}^{n} 2^{-|c_i|} \leq 1.$$

**Лемма 3.1.** *Для префиксных кодов верно неравенство Крафта-Макмилана.*

*Доказательство.* Рассмотрим дерево префиксного кода и посчитаем суммарную меру поддеревьев, которые соответствуют кодовым словам. □

**Утверждение 3.1.** *Для префиксных кодов верно и обратное: если есть набор целых чисел* $\{\ell_1, \ell_2, \ldots, \ell_n\}$*, удовлетворяющие неравенству Крафта-Макмилана*

$$\sum_{i=1}^{n} 2^{-\ell_i} \leq 1,$$

*то существует префиксный код с кодовыми словами* $\{c_1, c_2, \ldots, c_n\}$*, где* $|c_i| = \ell_i$*.*

*Доказательство.* Отсортируем $\ell_i$ по возрастанию и будем развешивать их в бесконечном двоичном дереве, выбирая каждый раз самый левый свободный узел соответствующей меры. Можно заметить, что мы всегда сможем найти такой узел. □

**Следствие 3.1.** *Для любого однозначно декодируемого кода существует префиксный код с теми же длинами кодовых слов.*

*Доказательства теоремы 3.1.* Сопоставим кодовым словам $\{c_i\}$ мономы $\{p_i\}$ от переменных $x$ и $y$ таким образом, что каждый '0' в кодовом слове соответствует $x$, а каждая '1' — $y$:

$$c_i = 0110101 \implies p_i(x, y) = xyyxyxy.$$

Рассмотрим следующее выражение для некоторого $L$.

$$\left(\sum_{i=1}^{n} p_i(x, y)\right)^L = \sum_{\ell=L}^{\max|c_i|\cdot L} M_\ell(x, y),$$

где $M_\ell$ обозначает сумму всех получившихся одночленов степени $\ell$. Заметим, что в каждом $M_\ell$ не более $2^\ell$ одночленов: в противном случае код не был бы однозначно декодируемым — каждый одночлен (без учёта коммутативности и ассоциативности) мог получиться не более одного раза.

Теперь рассмотрим значение этого выражения при $x = y = \frac{1}{2}$.

$$\left(\sum_{i=1}^{n} p_i\left(\tfrac{1}{2}, \tfrac{1}{2}\right)\right)^L = \sum_{\ell=L}^{\max|c_i|\cdot L} M_\ell\left(\tfrac{1}{2}, \tfrac{1}{2}\right) \leq \sum_{\ell=L}^{\max|c_i|\cdot L} (2^{-\ell} \cdot 2^\ell) \leq L \cdot \max|c_i| = O(L). \qquad (1)$$

Предположим теперь, что неравенство Крафта-Макмилана не выполняется, т.е.

$$q = \sum_{i=1}^{n} p_i(1/2, 1/2) = \sum_{i=1}^{n} 2^{-|c_i|} > 1.$$

Сравнивая это с (1) получаем противоречие: $q^L = O(L)$ (левая часть растёт экспоненциально, а правая — линейно). □

Пусть для каждого символа алфавита задана вероятность $p_i$. Нас будут интересовать самые короткие в среднем коды, т.е. такие, что

$$\sum_{i=1}^{n} p_i \cdot |c_i| \to \min .$$

**Теорема 3.2** (Шеннон)**.** *Для любого однозначно декодируемого кода выполняется*

$$\sum_{i=1}^{n} p_i \cdot |c_i| \geq \sum_{i=1}^{n} p_i \cdot \log \frac{1}{p_i}.$$

*Доказательство.* Перенесём всё в правую часть и применим неравенство Йенсена:

$$\sum_{i=1}^{n} p_i \cdot \log \frac{2^{-|c_i|}}{p_i} \leq \log \sum_{i=1}^{n} \left( p_i \frac{2^{-|c_i|}}{p_i} \right) = \log \sum_{i=1}^{n} 2^{-|c_i|} \leq \log 1 = 0.$$

□

**Теорема 3.3** (Шеннон)**.** *Для любого распределения вероятностей $\{p_1, p_2, \ldots, p_n\}$ существует однозначно декодируемый/префиксный код $\{c_1, c_2, \ldots, c_n\}$, такой что*

$$\sum_{i=1}^{n} p_i \cdot |c_i| \leq \sum_{i=1}^{n} p_i \cdot \log \frac{1}{p_i} + 1.$$

*Замечание* 3.1. От '+1' в правой части никак не избавиться: например, если у нас только два символа в алфавите, то $\sum p_i \cdot |c_i| = 1$, в то время как $\sum p_i \log \frac{1}{p_i}$ может быть сколько угодно близко к нулю.

*Доказательство.* Покажем, что найдутся $\{c_1, c_2, \ldots, c_n\}$ такие, что $|c_i| = \left\lceil \log \frac{1}{p_i} \right\rceil$. Код существует, т.к. для длин $c_i$ выполняется неравенство Крафта-Макмилана:

$$\sum_{i=1}^{n} 2^{-|c_i|} = \sum_{i=1}^{n} 2^{-\left\lceil \log \frac{1}{p_i} \right\rceil} \leq \sum_{i=1}^{n} 2^{-\log \frac{1}{p_i}} = \sum_{i=1}^{n} p_i = 1.$$

Теперь оценим среднюю длину кода:

$$\sum_{i=1}^{n} p_i \cdot |c_i| = \sum_{i=1}^{n} p_i \cdot \left\lceil \log \tfrac{1}{p_i} \right\rceil < \sum_{i=1}^{n} p_i \cdot \left( \log \tfrac{1}{p_i} + 1 \right) = \left( \sum_{i=1}^{n} p_i \cdot \log \tfrac{1}{p_i} \right) + 1.$$

□

### Код Шеннона-Фано

Упорядочим вероятности символов по убыванию: $p_1 \geq p_2 \geq \cdots \geq p_n$. Уложим на прямой без пропусков отрезки длиной $p_1$, $p_2$,..., $p_n$ и обозначим $i$-ый отрезок через $S_i$, а их объединение — через $S$. Коды тех букв $a_i$, для которых отрезок $S_i$ попал в левую половину $S$, будут начинаться с '0', а коды тех букв, для которых отрезок $S_i$ попал в правую часть $S$ — с '1'. Центральный отрезок может не попасть целиком в одну из половин $S$. Если центральный отрезок является первым или последним, то начнём его код, соответственно, с '0' или '1'. В противном случае отнесём его в произвольную половину $S$. Далее применяем эту стратегию отдельно для букв из левой половины $S$ и отдельно для правой половины $S$. Повторяем так пока не получим уникальные коды для всех символов.

**Определение 10.** Будем называть кодирование, при котором для некоторой константы $c$ и для всех $i$ выполняется $|c_i| \leq -\log p_i + c$, *сбалансированным.*

**Теорема 3.4** (Шеннон)**.** *Средняя длина кода Шеннона-Фано близка к энтропии, но не обязательно оптимальна:*

$$\sum_{i=1}^{n} p_i \cdot |c_i| = H + O(1).$$

### Код Хаффмана

**Определение 11.** Будем строить *код Хаффмана* по индукции. При $n = 2$ коды $c_1 = \langle 0 \rangle$, $c_2 = \langle 1 \rangle$. При $n > 2$ будем предполагать, что вероятности упорядочены по убыванию $p_1 \geq p_2 \geq \cdots \geq p_n$. Заменим символы $a_{n-1}$ и $a_n$ на символ $a'_{n-1}$ с вероятностью $p'_{n-1} = p_{n-1} + p_n$. Построим код Хаффмана для $n-1$ символа. Для символов $a_{n-1}$ и $a_n$ возьмём коды $c_{n-1} = c'_{n-1}0$ и $c_n = c'_{n-1}1$.

**Лемма 3.2.** *Средняя длина кодового слова для кода Хаффмана оптимальна, т.е. не превосходит средней длины любого другого префиксного кода (а значит и любого однозначно декодируемого).*

**Следствие 3.2.** *Для кода Хаффмана выполняется неравенство из теоремы Шеннона 3.3.*

*Замечание* 3.2. На энтропию случайной величины иногда удобно смотреть как на среднюю длину кода Хаффмена.

### Блоковое кодирование

Для того, чтобы нивелировать неустранимую '+1' в средней длине кода, мы будем кодировать не отдельные символы, а блоки символов. Пусть каждый блок состоит из $k$

символов. Пусть случайные величины $\alpha_1, \alpha_2, \ldots, \alpha_k$ распределены как $\alpha$ и соответствуют буквам в блоке.

$$H(\alpha_1, \alpha_2, \ldots, \alpha_k) = \sum_{i=1}^{k} H(\alpha_i) = k \cdot H(\alpha).$$

Тогда по теоремам Шеннона получается следующее ограничение на среднюю длину кода символа в блоке:

$$H(\alpha) \leq (\text{средняя длина кода буквы в блоке}) \leq H(\alpha) + \frac{1}{k}.$$

При кодировании блоков длины 100 мы получаем отклонение от энтропии не более, чем на 0.01. Однако мы не можем применить код Хаффмена, т.к. на вход алгоритму его построения нужно было бы передать $n^{100}$ частот символов.

### Арифметическое кодирование

Мы построим код со следующим ограничением на среднюю длину:

$$\sum_{i=1}^{n} p_i \cdot |c_i| \leq \sum_{i=1}^{n} p_i \cdot \log \frac{1}{p_i} + 2,$$

что хуже, чем в теореме Шеннона.

**Определение 12.** Будем называть полуинтервал *стандартным*, если он имеет вид $[0.v0_2, 0.v1_2)$, где $v$ — это некоторая последовательность битов, а числа записаны в двоичной системе счисления. Будем сопоставлять каждому стандартному интервалу $[0.v0_2, 0.v1_2)$ код $v$.

Для первой буквы кода на отрезке [0,1] мы отложим слева направо непересекающиеся интервалы длины $p_i$. Пусть первая буква блока — это $a_{i_1}$, тогда для второй буквы кода мы внутри интервала соответствующего $p_{i_j}$ повторим эту операцию (отложим непересекающиеся интервалы), но длины интервалом будут уже масштабированы с коэффициентом $p_i$. Повторим эту операцию $k$ раз. Получившемуся интервалу в качестве его кода сопоставим код наибольшего стандартного интервала, который полностью содержится внутри него.

**Утверждение 3.2.** *В интервале $[a, b)$ всегда найдётся стандартный интервал длины $2^{-k}$, где $\frac{b-a}{4} < 2^{-k} \leq \frac{b-a}{2}$, т.е. длина кода любого интервала при арифметическом кодировании не превосходит* $\log \frac{4}{b-a} = \log \frac{1}{p_i} + 2$.

*Замечание* 3.3. В случае Марковской цепи можно строить код с соответствующими условными вероятностями.

**Блоковые коды с ошибками**

Пусть $\alpha_1, \alpha_2, \ldots, \alpha_n$ — независимые одинаково распределённые на $\{a_1, a_2, \ldots, a_k\}$ случайные величины с вероятностями $p_1, p_2, \ldots, p_k$. Рассмотрим блоковое кодирование, заданное функциями $E_n$ и $D_n$:

$$E_n : \{a_1, a_2, \ldots, a_k\}^n \to \{0,1\}^{L_n},$$

$$D_n : \{0,1\}^{L_n} \to \{a_1, a_2, \ldots, a_k\}^n,$$

**Определение 13.** *Вероятность ошибки* $\varepsilon_n$ — это вероятность следующего события: $[(\alpha_1, \alpha_2, \ldots, \alpha_n) = (a_{i_1}, a_{i_2}, \ldots, a_{i_n}) \mid D_n(E_n(a_{i_1}, a_{i_2}, \ldots, a_{i_n})) \neq (a_{i_1}, a_{i_2}, \ldots, a_{i_n})]$.

**Теорема 3.5** (Шеннон)**.** *При блоковом кодировании допускающем ошибки выполняются следующие соотношения.*

1. *Если* $h > H(\alpha) = \sum_{i=1}^{k} p_i \log \frac{1}{p_i}$, *то существует функции* $(E_n, D_n)$ *для* $L_n = \lceil h \cdot n \rceil$, *такие что* $\varepsilon_n \to 0$ *при* $n \to \infty$.
2. *Если* $h < H(\alpha) = \sum_{i=1}^{k} p_i \log \frac{1}{p_i}$, *то для любых функций* $(E_n, D_n)$ *для* $L_n = \lceil h \cdot n \rceil$ *вероятность ошибки* $\varepsilon_n \to 1$ *при* $n \to \infty$.

**Определение 14.** Будем называть слово $w = \langle a_{i_1}, a_{i_2}, \ldots, a_{i_n} \rangle$ $\delta$*-типичным*, если каждая буква $a_j$ встречается в нём $t_j$ раз, причём

$$\begin{cases} t_j \leq (p_j + \delta) \cdot n, \\ t_j \geq (p_j - \delta) \cdot n. \end{cases}$$

**Лемма 3.3.** *Для* $\delta = n^{-0.49} = \frac{n^{0.01}}{\sqrt{n}}$ *вероятность не* $\delta$*-типичного не превосходит* $\varepsilon_n$, *для* $\varepsilon_n \to 0$.

*Доказательство.* Применить неравенство Чебышева

$$\mathrm{P}[|X - \mu| \geq \delta n] \leq \frac{\sigma^2}{(\delta n)^2} = \frac{n p_i (1 - p_i)}{\delta^2 n^2} = O(n^{-0.02}).$$

□

**Лемма 3.4.** *Для* $\delta = n^{-0.49}$ *количество* $\delta$*-типичных слов не превосходит* $2^{h \cdot n}$ *(при достаточно больших* $n$*).*

*Доказательство.* Давайте для начала рассмотрим слова определённого *типа*, в которых буква $i$ встречается $n_i$ раз, $n_1 + n_2 + \cdots + n_k = n$. Сначала оценим количество слов типа, в котором $n_i = n \cdot p_i$. Таких слов

$$\frac{n!}{n_1! n_2! \cdots n_k!}.$$

По формуле Стирлинга $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \cdot (1+o(1))$.

$$\log \frac{n!}{n_1!n_2!\cdots n_k!} \approx \log \frac{\text{poly}(n)\left(\frac{n}{e}\right)^n}{\text{poly}(n)\left(\frac{n_1}{e}\right)^{n_1}\cdots\left(\frac{n_k}{e}\right)^{n_k}} =$$

$$= \log\left(\frac{n}{n_1}\right)^{n_1}\cdots\left(\frac{n}{n_k}\right)^{n_k} + O(\log n) = \sum_{i=1}^{k} \underbrace{np_i}_{n_i} \cdot \log\tfrac{1}{p_i} + O(\log n) < h\cdot n. \quad (2)$$

Мы оценили это только для конкретного типа слов. Давайте оценим для произвольного $\delta$-типичного слова с $n_i = n\cdot(p_i+\Delta_i)$, где $|\Delta_i| \le \delta$. Тогда (2) изменится следующим образом:

$$\cdots = \sum_{i=1}^{k} n(p_i+\Delta_i)\cdot\log\tfrac{1}{p_i+\Delta_i} + O(\log n) = n\cdot\sum_{i=1}^{k} p_i\cdot\log\tfrac{1}{p_i} + O(\log n) + n\cdot O(\delta) < h\cdot n.$$

(Действительно, энтропия — это непрерывная функция, а значит при небольшом отклонении она изменяется на $c\cdot\max_i\Delta_i$, где $c$ зависит от производной функции энтропии.) Итого общее количество $\delta$-типичных слов можно оценить как количество типов умноженное на количество $\delta$-типичных слов одного типа:

$$\text{poly}(n)\cdot 2^{n\cdot H(\alpha)+n\cdot O(\delta)+O(\log n)} < 2^{h\cdot n}.$$

□

*Доказательство теоремы 3.5.*

1. Если мы будем кодировать только $\delta$-типичные слова, то по лемме 3.4 нам будет достаточно длины кода $L_n$, а вероятность всех не типичных слов будет стремиться к нулю.

2. Обозначим за $\hat{\varepsilon}_n$ вероятность ошибки при декодировании $\delta$-типичных слов. Мы хотим показать, что $\hat{\varepsilon}_n \to 1$. Давайте рассмотрим конкретное $\delta$-типичное слово $w = \langle a_{i_1}, a_{i_2}, \ldots, a_{i_n}\rangle$. Пусть $p'_1, p'_2, \ldots, p'_n$ — это частоты букв $a_1, a_2, \ldots, a_n$. Оценим вероятность появления $w$:

   $$\Pr[\langle a_{i_1}, a_{i_2}, \ldots, a_{i_n}\rangle = w] = p_1^{p'_1\cdot n}\cdot\ldots\cdot p_k^{p'_k\cdot n} = 2^{-(\sum_i p'_i\log\frac{1}{p_i})\cdot n} \le 2^{-(\sum_i p_i\log\frac{1}{p_i})\cdot n + O(\delta_n\cdot n)}.$$

   Всего мы может корректно закодировать не более $2^{L_n}$ $\delta$-типичных слов, т.е. вероятность корректно декодировать $\delta$-типичное слово

   $$1-\hat{\varepsilon}_n \le 2^{L_n}\cdot 2^{-H(\alpha)\cdot n+O(\delta_n\cdot n)} \le 2^{h\cdot n+1}\cdot 2^{-H(\alpha)\cdot n+O(\delta_n\cdot n)} \to 0.$$

   Таким образом $\hat{\varepsilon}_n \to 1$. Вместе с леммой 3.3 получаем, что $\varepsilon_n \to 1$.

□

*Замечание* 3.4. Используя предыдущую теорему можно, например, получить альтернативное доказательство неравенства $H(\alpha,\beta) \le H(\alpha) + H(\beta)$. В левой части стоит асимптотическая средняя длина кода при блоковом кодировании $(\alpha,\beta)$, а справа сумма средних длин кодов при блоковом кодировании $\alpha$ и $\beta$ отдельно друг от друга. Т.к. мы можем рассмотреть кодирование $(\alpha,\beta)$ как конкатенацию кодов для $\alpha$ и $\beta$, то неравенство выполняется.

## Свойства распределений

### Энтропийные профили

**Утверждение 4.1.** *Для любого $h \ge 0$ существует распределение $\alpha$: $H(\alpha) = h$.*

*Доказательство.* Возьмём некоторое целое $n$: $0 \le h \le \log n$. Искомое распределение — это линейная комбинация распределений с вероятностями $(1,0,\dots,0)$ и $(\frac{1}{n},\frac{1}{n},\dots,\frac{1}{n})$. □

Каким может быть совместное распределение двух случайных величин $\alpha$ и $\beta$? Рассмотрим как может быть устроен *энтропийный профиль* $(H(\alpha), H(\beta), H(\alpha,\beta))$.

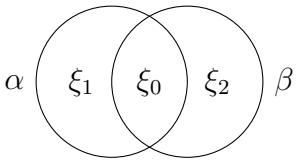
**Утверждение 4.2.** *Для любых чисел $h_1, h_2, h_{12} \ge 0$, которые удовлетворяют следующим соотношениям*

$$\begin{cases} h_{12} \le h_1 + h_2 & \iff \quad t_0 = I(\alpha:\beta) \ge 0, \\ h_2 \le h_{12} & \iff \quad t_1 = H(\alpha \mid \beta) \ge 0, \\ h_1 \le h_{12} & \iff \quad t_2 = H(\beta \mid \alpha) \ge 0. \end{cases}$$

*существует пара случайных величин $(\alpha,\beta)$ с энтропийным профилем $(h_1,h_2,h_{12})$.*

*Доказательство.* Пусть $\xi_0, \xi_1, \xi_2$ — независимые случайные величины с энтропиями $t_0, t_1, t_2$ соответственно. Тогда $\alpha = (\xi_0,\xi_1)$ и $\beta = (\xi_0,\xi_2)$ будут искомыми величинами.

$$\begin{cases} H(\xi_0) = t_0 = h_1 + h_2 - h_{12}, \\ H(\xi_1) = t_1 = h_{12} - h_2, \\ H(\xi_2) = t_2 = h_{12} - h_1. \end{cases}$$




□

Давайте попробуем разобраться с аналогичным вопросом для троек случайных величин. Энтропийный профиль для тройки $(\alpha,\beta,\gamma)$ будет задаваться 7 числами:

$$\big(H(\alpha), H(\beta), H(\gamma), H(\alpha,\beta), H(\alpha,\gamma), H(\beta,\gamma), H(\alpha,\beta,\gamma)\big).$$

Для случайных величин $(\alpha,\beta,\gamma)$ можно записать 9 независимых неравенств.

$$\begin{array}{lll} H(\alpha \mid \beta,\gamma) \ge 0, & I(\alpha:\beta) \ge 0, & I(\alpha:\beta \mid \gamma) \ge 0, \\ H(\beta \mid \gamma,\alpha) \ge 0, & I(\beta:\gamma) \ge 0, & I(\beta:\gamma \mid \alpha) \ge 0, \\ H(\gamma \mid \alpha,\beta) \ge 0, & I(\gamma:\alpha) \ge 0, & I(\gamma:\alpha \mid \beta) \ge 0. \end{array}$$

**Определение 15.** Определим общую информацию трёх случайных величин

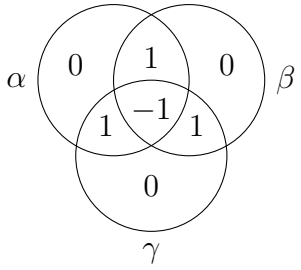
$$I(\alpha : \beta : \gamma) = I(\alpha : \beta) - I(\alpha : \beta \mid \gamma).$$

**Утверждение 4.3.** *Общая информация трёх случайных величин может быть отрицательной.*

*Доказательство.* Пусть $\alpha$ и $\beta$ будут независимыми равномерно распределёнными на $\{0,1\}$ случайными величинами. Случайная величина $\gamma$ будет принимать значение из $\{0,1\}$ в соответствии со следующим соотношением:

$$\alpha \oplus \beta \oplus \gamma = 0.$$

Мы получим следующую картину:

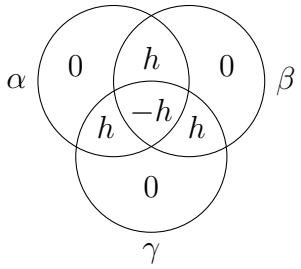



□

**Утверждение 4.4.** *Других неравенств для троек нет.*

**Утверждение 4.5.** *Есть профили, которые не реализуются никакими распределениями, но их мера 0.*

*Упражнение* 4.1. Доказать, что следующий профиль реализуется только при $h = \log n$ для некоторого целого $n$.




**Утверждение 4.6.** $2H(\alpha, \beta, \gamma) \leq H(\alpha, \beta) + H(\alpha, \gamma) + H(\beta, \gamma)$.

**Следствие 4.1** (Теорема 1.3)**.** *Для* $A \subset \{0,1\}^* \times \{0,1\}^* \times \{0,1\}^*$

$$2\chi(A) \leq \chi_{12}(A) + \chi_{13}(A) + \chi_{23}(A).$$

*Доказательство.* Пусть $(\alpha, \beta, \gamma)$ равномерно распределены на $A$.

$$2\chi(A) = 2H(\alpha, \beta, \gamma) \leq \underbrace{H(\alpha, \beta)}_{\leq \chi_{12}(A)} + \underbrace{H(\alpha, \gamma)}_{\leq \chi_{13}(A)} + \underbrace{H(\beta, \gamma)}_{\leq \chi_{23}(A)}.$$

□