Заметки к курсу „Теория информации“. Лекции 1-3

А. Смаль

23 января 2020 г.

Аннотация

Курс посвящён изучению подходов к определению понятия „количество информации“. Последовательность изложения материала данного курса основана на классической статье Колмогорова „Три подхода к определению понятия количества информации“ (1965).

В курсе будет рассмотрено три подхода к определению „количества информации“: комбинаторный (информация по Хартли), вероятностный (энтропия Шеннона) и алгоритмический (Колмогоровская сложность). Кроме этого мы поговорим про различные применения аппарата теории информации в различных областях компьютерных наук: в криптографии, в коммуникационной сложности, в теории кодирования, в теории конечных автоматов, в теории сложности вычислений и некоторых других.

Содержание

1. Комбинаторный подход

1.1. Информация по Хартли

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

Определение 1.1 (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 измерения в байтах нужно выбрать основание 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 $\chi_2(A) = \log |\pi_2(A)|$ — количество информации в них по Хартли.

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

Определение 1.2. Количество информации в второй координате $A \subset \{0,1\}^* \times \{0,1\}^*$ при известной первой

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

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

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

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

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* $\chi(Y) \leq \chi(X)$,
2. *является инъекцией, то* $\chi(X) \leq \chi(Y)$.

1.2. 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 $Q = \{(q_1, q_2, \ldots, q_k)\}$ — множество протоколов (ответы на вопросы). Можно рассматривать A и Q как проекции некоторого множества исходов игры S на разные координаты. Тогда верны следующие неравенства:

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

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

1.3. Цена информации

Пусть загадано некоторое целое 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. Будем задавать вопросы так, чтобы отрицательные ответы приносили в два раза больше информации, чем положительные. Тогда за каждый бит информации мы заплатим некоторое константное количество рублей c. Пусть все вопросы будут вида „$x \in T$?“. Потребуем, чтобы

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

Пусть $|X \cap T| = \alpha |X|$, тогда $|X \cap \overline{T}| = (1-\alpha)|X|$, таким образом получается уравнение

$$2 \log(1/\alpha) = \log(1/(1-\alpha)),$$

эквивалентное квадратному уравнению

$$\alpha^2 = 1 - \alpha.$$

Из двух корней нас интересует тот, что меньше 1, т.е. $\alpha = (\sqrt{5}-1)/2$. Следовательно при любом ответе мы заплатим $c = 1/(-\log \alpha) \approx 1.44$ рублей за бит, а в целом — $c \log n$ рублей.

В этой оценке мы полностью проигнорировали вопросы округления. Действительно, у нас никогда получится разделить множество из n элементов на два в отношении

$\alpha : (1-\alpha)$, т.к. α — иррациональное. Поэтому на каждом вопросе будет накапливаться некоторая ошибка округления. Давайте вместо вопросов принадлежности некоторому подмножеству T множества X будем задавать вопрос о принадлежности отрезку с вещественными координатами. Начнём с отрезок $S = [1, n]$ и будем каждый раз уменьшать его в $1/\alpha$ раз, т.е. первым вопросом спросим, принадлежит ли x отрезку $S' = [1, 1 + \alpha(n-1)]$. Длина отрезка S' в $1/\alpha$ раз меньше длины отрезка S. Продолжим действовать так же до тех пор, пока длина отрезка не станет меньше 1 — в этом случае x определено однозначно. После каждого вопроса длина отрезка уменьшается максимум в $1/(1-\alpha) = 1/\alpha^2$, поэтому длина последнего отрезка не меньше α^2. Таким образом длина отрезка сократится не более, чем в $(n-1)/\alpha^2$ раз. Поскольку мы каждый раз выбирали отрезки так, чтобы платить c рублей за уменьшение $\log|S|$ на 1, то в сумме заплатим не более

$$c\log((n-1)/\alpha^2) = c\log(n-1) - 2c\log\alpha = c\log(n-1) + 2.$$

При любом исходе мы заплатим целое число рублей, поэтому эту оценку можно уточнить до $\lfloor c\log(n-1)\rfloor + 2$.

Нижняя оценка. Применим рассуждение про злонамеренного противника (adversary argument). Пусть противник выбирает ответ ДА/НЕТ в зависимости от того, какое из двух значений $1/(\log|X| - \log|X\cap T|)$ 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.

2. Вероятностный подход

2.1. Энтропия Ше́ннона

Энтропия Шеннона определяет количество информации $H(\alpha)$ в распределении вероятностей для некоторой случайной величины α. Пусть α 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 информации, которую мы получаем от каждого исхода.

$$H(\alpha) = \sum_i p_i \cdot (\text{информация в } a_i).$$

Как оценить, сколько информации в исходе a_i? Пусть U — всё пространство элементарных исходов, все исходы которого равновероятны. Тогда событию $\alpha = a_i$ соответствует множеству элементарных исходов меры p_i. Соответственно, если случилось событие $\alpha = a_i$, то размер множества согласованных с этим событием исходов уменьшается с $|U|$ до $p_i|U|$, т.е. событие $\alpha = a_i$ сообщает нам $\log|U| - \log(p_i|U|) = \log\frac{1}{p_i}$ битов информации. Пусть теперь элементарные исходы не равновероятны. В этом случае событие $\alpha = a_i$ сообщает нам информацию, которая уменьшает *меру* множества возможных исходов в $1/p_i$ раз, т.е. опять получаем $\log 1 - \log p_i = \log\frac{1}{p_i}$. Это приводит нас к следующему определению.

Определение 2.1 (1948)**.** Энтропия Шеннона случайной величины α

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

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

Можно вывести это соотношение из определения информации по Хартли другим способом. 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* α *вырождено.*
- $H(\alpha) \leq \log k$, *причём* $H(\alpha) = \log k \iff$ *величина* α *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* $q_1 + \cdots + q_n = 1$. *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 α и β 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* β *полностью определяется значением* α, *т.е.* $\beta = f(\alpha)$.

Доказательство. Введём обозначения для вероятностей событий совместного распределения вероятностей (α, β). 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

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

Перенесём всё в левую часть и применим неравенство Йенсена.

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

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

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 α и β независимы.

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

Определение 2.2. Энтропия α 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 2.3. *Условная (относительная) энтропия* α относительно β

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

Другими словами

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

Если подставить определение 2.2, 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.*

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

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 $\beta = f(\alpha)$. □

2.2. Взаимная информация

Определение 2.4. *Информация в α о величине β* определяется следующим соотношением:

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

Эту величину так же *называют взаимной информацией случайных величин α и β.*

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 2.5. Пусть α, β, γ — случайные величины. Определим *взаимную информацию в α о β при условии γ.*

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).

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

□

Утверждение 2.1 (chain rule for mutual information)**.** *Имеют место следующие соотношения:*

1. $I((\alpha,\beta):\gamma) = I(\alpha:\gamma) + I(\beta:\gamma \mid \alpha)$.
2. $I((\alpha,\beta):\gamma \mid \delta) = I(\alpha:\gamma \mid \delta) + I(\beta:\gamma \mid \alpha,\delta)$.

3. Кодирование

3.1. Однозначно декодируемые коды

Определение 3.1. 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 3.2. 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 ℓ_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_ℓ обозначает сумму всех получившихся мономов степени ℓ. Заметим, что в каждом M_ℓ не более 2^ℓ мономов: в противном случае код не был бы однозначно декодируемым — каждый моном (без учёта коммутативности и ассоциативности) мог получиться не более одного раза.

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

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

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*

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

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

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

□

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:

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

□

3.2. 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 3.3. 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:*

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

3.3. Код Хаффмана

Определение 3.4. 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.