# Схема арифметического кодирования

- *По исходному значению вероятностей строим таблицу, состоящую из пересекающихся в граничных точках отрезков.*

Объединение этих отрезков должно образовывать интервал [0;1], а их длины пропорциональны вероятностям кодируемых значений. Алгоритм заключается в построении отрезка, однозначно определяющего данную последовательность значений.

- *По мере поступления входных символов сужаем отрезок, определяющий сообщение*

Если имеется отрезок сообщения длиной n-1, то для построения отрезка сообщения длиной n, предыдущий интервал разбивается на столько частей, сколько значений включает алфавит источника.

- *Начало и конец каждого нового отрезка определяется путём прибавления к началу предыдущего интервала произведения его ширины на значения границ отрезка, отвечающего текущему символу.*
- *Из полученных отрезков выбирается тот, который соответствует конкретной последовательности длиной n.*
- *Для построенного отрезка находится число, принадлежащее этому отрезку, равное целому числу, делённому на минимально возможную степень 2.*

Это вещественное число и будет кодом для рассматриваемой последовательности.

**Пример**: закодировать по схеме алфавитного кодирования сообщение «МАТЕМАТИКА»

Алфавит сообщения {М, А, Т, Е, И, К}

| символ | вероятность | интервал |
|---|---|---|
| **М** | **0,2** | **[0;0,2)** |
| **А** | **0,3** | **[0,2;0,5)** |
| **Т** | **0,2** | **[0,5;0,7)** |
| **Е** | **0,1** | **[0,7;0,8)** |
| **И** | **0,1** | **[0,8;0,9)** |
| **К** | **0,1** | **[0,9;1)** |

Последовательность интервалов, соответствующих кодируемому сообщению

| Символ – интервал | Интервал сообщения | Ширина интервала |
|---|---|---|
| ***М*** - $[0; 0,2)$ | $[0; 0,2)$ | $0,2$ |
| ***А*** - $[0,2; 0,5)$ | $[0,04; 0,1)$ | $0,6$ |
| ***Т*** - $[0,5; 0,7)$ | $[0,07; 0,082)$ | $0,012$ |
| ***Е*** - $[0,7; 0,8)$ | $[0,0784; 0,0796)$ | $0,0012$ |
| ***М*** - $[0; 0,2)$ | $[0,0784; 0,07864)$ | $0,00024$ |
| ***А*** - $[0,2; 0,5)$ | $[0,078448; 0,07852)$ | $0,72 \times 10^{-4}$ |
| ***Т*** - $[0,5; 0,7)$ | $[0,078484; 0,0784984)$ | $0,144 \times 10^{-4}$ |
| ***И*** - $[0,8; 0,9)$ | $[0,07849552; 0,07849696)$ | $0,144 \times 10^{-5}$ |
| ***К***- $[0,9; 1,0)$ | $[0,078496816; 0,07849696)$ | $0,144 \times 10^{-6}$ |
| ***А*** - $[0,2; 0,5)$ | $[0,0784968448; 0,078496888)$ | $0,432 \times 10^{-7}$ |

Результат кодирования сообщения «МАТЕМАТИКА» - вещественное число, принадлежащее интервалу [0,078496448; 0,078496888]

Целое число, делённое на минимальную степень 2, принадлежащее данному отрезку

$$0{,}07849687 = 1316959/2^{24}$$

Двоичный 24-разрядный код числа

$$1316959_{10} = 000101000001100001011111_2$$

Этот код и есть арифметический код сообщения «МАТЕМАТИКА»

Длина кода L(x)=24 бита

Средняя длина кода

$$\overline{L}(x) = \frac{24}{10} = 2{,}4$$

Декодирование арифметического кода сообщения происходит по следующему алгоритму:

Шаг 1. По таблице отрезков символов алфавита определяется интервал, содержащий текущий код — и по этому интервалу из той же таблицы однозначно определяется символ исходного сообщения. Если это маркер конца сообщения, то конец, иначе — переход к шагу 2.

Шаг 2. Из текущего кода вычитается нижняя граница содержащего его интервала. Полученная разность делится на длину этого интервала. Полученное значение считается новым текущим кодом. Переход к шагу 1. Рассмотрим пример декодирования сообщения, сжатого по алгоритму арифметического кодирования.

**Пример 3.** Длина исходного сообщения 10 символов. Двоичный арифметический код сообщения $000101000001100001011111_2 = 1316259_{10}$. Вещественное число, принадлежащее интервалу, однозначно определяющему закодированное сообщение, $1316959/_{2^{24}} = 0,07849687$.

Это число и будет текущим кодом сообщения.

По исходной таблице значений д.с.в. и назначенных им интервалов (Таблица 2.7) определяется отрезок, которому принадлежит это число, - $[0; 0,2)$, и соответственно, что первый закодированный символ — это "***M***".

Исключим из результата кодирования влияние теперь уже известного первого символа "***M***": *для этого вычтем из текущего кода нижнюю границу отрезка, отведенного для раскодированного символа, и разделим полученный результат на ширину этого отрезка*, т.е. следующим декодируемым числом будет $\frac{0,07849687 - 0}{0,2} = 0,39248435$. Это число принадлежит отрезку $[0,2; 0,5)$, отведенному символу "***A***", следовательно вторым символом декодированной последовательности будет "***A***". Исключим из итогового интервала $[0,2; 0,5)$ влияние буквы "***A***". Для этого вычтем из текущего кода нижнюю границу этого интервала и разделим на его ширину: $\frac{0,39248435 - 0,2}{0,3} = 0,6416145$. Результат принадлежит отрезку $[0,5; 0,7)$, отведенному для символа "***T***", — это очередной декодируемый символ.

Исключив из результата декодирования влияние буквы “***T***”, получим $\frac{0,6416145-0,5}{0,2}=0,7080725$. Результат принадлежит отрезку буквы “***E***” $[0,7;0,8)$ и т. д., пока не декодируем все символы (Таблица 2.10).

Описанный алгоритм декодирования арифметического кода сообщения выглядит следующим образом.

Таблица 2.10

| Декодируемое число | Символ на выходе | Интервал | Ширина интервала |
|---|---|---|---|
| $0,07849687$ | М | $[0;0,2)$ | $0,2$ |
| $0,39248435$ | А | $[0,2;0,5)$ | $0,3$ |
| $0,6416145$ | Т | $[0,5;0,7)$ | $0,2$ |
| $0,7080725$ | Е | $[0,7;0,8)$ | $0,1$ |
| $0,080725$ | М | $[0;0,2)$ | $0,2$ |
| $0,403625$ | А | $[0,2;0,5)$ | $0,3$ |
| $0,67875$ | Т | $[0,5;0,7)$ | $0,2$ |
| $0,89375$ | И | $[0,8;0,9)$ | $0,1$ |
| $0.9375$ | К | $[0,9;1,0)$ | $0,1$ |
| $0.375$ | А | $[0,2;0,5)$ | $0,3$ |