A Very Short Introduction to Information Theory

Bostan Viorel

Information

Information can be of two types:

- Discrete (Digital)
  - Discrete information is characterized by discrete values of a certain quantity
  - For example: A Bernoulli trials process:

$$SSFSSSFFSSSSF$$

  - Discrete values of some function

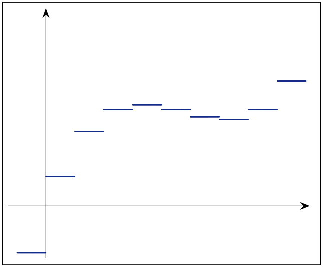
$$12345678$$

- Continuous (Analog)
  - Continuous information is characterized by continuous change of a certain quantity
  - For example: Pressure sensor indicators
  - Continuous values of velocity as a function of time

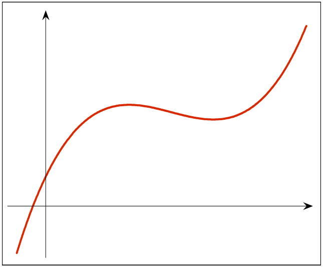
$$V(t) = 16t^2 - 10t + 2$$

Discrete (Digital) vs Continuous (Analog)

Discrete (Digital)



Continuous (Analog)



Discrete (Digital) vs Continuous (Analog)

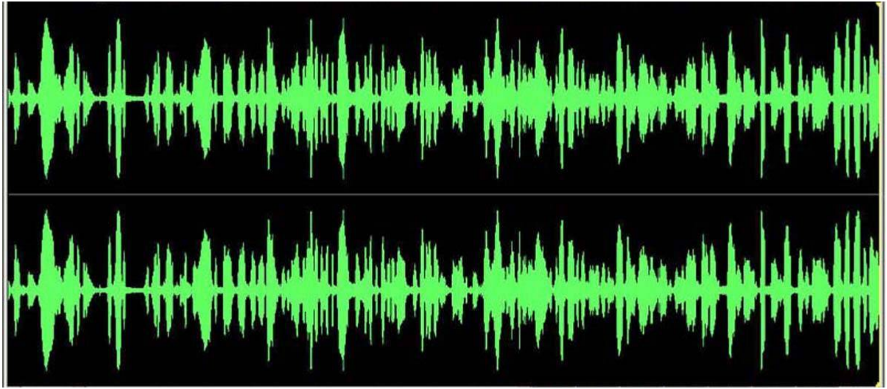
- It is more convenient to deal with discrete information.
- On the other hand plenty of information comes as continuous.
- Thus, it is necessary to convert the continuous information to discrete and viceversa.
    - For example, a **modem** (name comes from *modulation-demodulation*) is such translating device:
    - it converts digital data from computer to sound (electromagnetic wawe oscillations) and viceversa.
- While converting continuous information to discrete a very important feature is the **discretisation frequency** ν and the **discretisation period** T :

$$T = \frac{1}{\nu}$$

- In signal analysis discretisation is also called **sampling**.

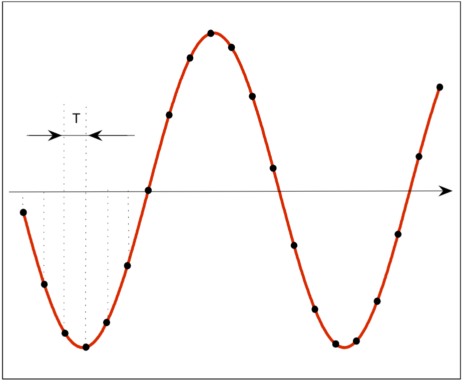
Analog to Digital Conversion

Usually sound waves look like this (especially in sound processing)



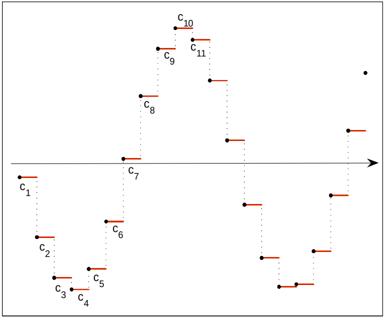
Analog to Digital Conversion

Original continuous signal



Analog to Digital Conversion

Discretized signal



Digital signal

$$c_1\, c_2\, c_3\, c_4\, c_5\, c_6 \dots$$

Analog to Digital Conversion

- Obviuosly, the higher the discretisation frequency the more precise is the conversion of the continuous information into discrete one.
- On the other hand, if we will increase ν, then the quantity of digital data increases too, which in turn raises the storage issues and time for processing and transmission of these data.
- It turns out that in order to increase the precision of conversion it is not necessary to increase the discretisation frequency infinitely.
- It is rerasonable to increase this frequency up to a certain limit, a limit established by the Sampling Theorem, known as Nyquist Law.
- Recall that any continuous quantity (i.e. function, signal) can be described as a sum of wave functions of the form $A\sin(\omega t + \varphi)$ (called harmonics), where A is the amplitude, ω is the frequency, φ is the phase and t denotes time.

Analog to Digital Conversion

- **Nyquist Law** states that in order to discretize exactly some continuous signal, the discretisation frequency ν should be at least twice as the biggest frequency of the harmonics contained in the signal.
- An exapmle of application of the Nyquist Law are music CDs.
- The higher the discretisation frequency the more "precise" (clear, full) is the recorded sound.
- Since human ears can distinguish sounds with frequencies up to $20KHz$ ($20,000$ oscillations per sec), it is nonsense to record sounds with higher frequency.
- Therefore, according to Nyquist Law, the discretisation frequency should be at least $40KHz$.
- Actually, the industry standard is $44.1KHz$

Analog to Digital Conversion

- In conversion of the discrete information into a continuous one, one of the main characteristics is the rate (speed) of conversion: the higher the rate, the more high-frequency harmonics will be obtained.
- On the other hand, a continuous signal wiht high-frequency harmonics is more difficult to work with. For example, usual phone lines can be used for transmission of sounds with frequency only up to $3KHz$.
- Devices for conversion of continuous information into discrete one and viceversa are called *ADC* (Analog to Digital Convertor, A/D) and *DAC* (Digital to Analog Convertor, D/A).
- When buying different electronic devices and appliances, the customer should look for A/D and/or D/A characteristics.

Information Storage, Measure and Transmission

- Information is measured in bits, just as length is measured in meters and time is measured in seconds.
- Knowing the amount of information, in bits, is not the same as knowing the information itself, what it means, or what it implies.
- How is information quantified?
- Consider a situation or experiment that could have any of several possible outcomes. Examples might be flipping a coin or selecting a card from a deck of playing cards.
- How compactly could one person **A** tell another person **B** the outcome of such an experiment or observation?
- First consider the case of the two outcomes of flipping a coin, and let us suppose they are equally likely.If **A** wants to tell **B** the result of the coin toss, **A** could use several possible techniques, but they are all equivalent, in terms of the amount of information conveyed, to saying either “heads” or “tails” or to saying 0 or 1.

Information Storage, Measure and Transmission

- We say that the information so conveyed is one bit. If **A** flipped two coins, **A** could say which of the four possible outcomes actually happened, by saying 0 or 1 twice.
- Similarly, the result of an experiment with eight equally likely outcomes could be conveyed with three bits, and more generally 2^n outcomes with n bits. Thus, the amount of information is the $\log_2 n$ of the number n of equally likely outcomes.
- Note that conveying information requires two phases.
- First is the "setup" phase, in which **A** and **B** agree on what they will communicate about, and exactly what each sequence of bits means. This common understanding is called the **code**.
- For example, to convey the suit of a card chosen from a deck, their code might be that 00 means clubs, 01 diamonds, 10 hearts, and 11 spades. Agreeing on the code may be done before any observations have been made, so there is not yet any information to be sent.

Information Storage, Measure and Transmission

- The setup phase can include informing the recipient that there is new information.
- Then, there is the "outcome" phase, where actual sequences of 0 and 1 representing the outcomes are sent. These sequences are the data.
- Using the agreed-upon code, **A** draws the card, and tells **B** the suit by sending two bits of data. **A** could do so repeatedly for multiple experiments, using the same code.
- After **B** knows that a card is drawn, but before receiving **A**'s message, **B** is uncertain about the suit. His uncertainty, or lack of information, can be expressed in bits.
- Upon hearing the result, his uncertainty is reduced by the information he receives. **B**'s uncertainty rises during the setup phase and then it is reduced during the outcome phase.

Information Storage, Measure and Transmission

Note some important things about information, some of which are illustrated in previuos example:

- Information can be learned through observation, experiment, or measurement;
- Information is subjective, or “observer-dependent.” What **A** knows is different from what **B** knows (if information were not subjective, there would be no need to communicate it);
- A person’s uncertainty can be increased upon learning that there is an observation about which information may be available, and then can be reduced by receiving that information;
- Information can be lost, either through loss of the data itself, or through loss of the code;
- The physical form of information is localized in space and time. As a consequence,
    - Information can be sent from one place to another;
    - Information can be stored and then retrieved later.

Information Storage, Measure and Transmission

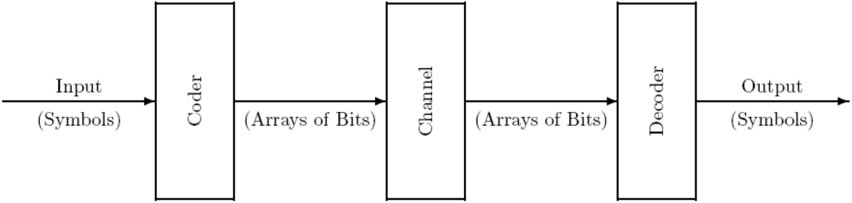
Bit is the unit of measuring information. Oftenly, *byte* is being used which equals 8 bits. As for usual measurement units, we have prefixes K (Kilo), M (Mega), G (Giga),T (Tera), P (Penta) and so on. For bits and bytes, these are not powers of 10, but 2:

$$\begin{aligned} 1Kbit &= 2^{10} = 1024 \approx 10^3 \\ 1Mbit &= 2^{20} \approx 10^6 \\ 1Gbit &= 2^{30} \approx 10^9 \\ 1Tbit &= 2^{40} \approx 10^{12} \end{aligned}$$

A single bit is useful if exactly two answers to a question are possible. Examples include the result of a coin toss (heads or tails), the gender of a person (male or female), the verdict of a jury (guilty or not guilty), and the truth of an assertion (true or false).
Most situations in life are more complicated. Codes allow complex objects to be represented by arrays of bits.

Information Storage, Measure and Transmission

Processes of coding and decoding can be repeated several times. Errors during transmission of information are due to the noise in the channel of transmission (atmospheric noise, technical perturbations etc). also, errors can be introduced during the coding and decoding process itself. Information theory also studies ways to minimize such errors.



Information can be transmitted sequentially, i.e. bit after bit, or in paralel, i.e. blocks of data are send simultaneously.

Information Storage, Measure and Transmission

Codes

Some objects for which codes may be needed include:

- **Letters**: BCD (Binary Coded Decimals), EBCDIC, ASCII (American Standard Code for Information Interchange), Unicode, Morse Code
- **Integers**: Binary, Gray, 2's complement
- **Numbers**: Floating-Point
- **Proteins**: Genetic Code
- **Telephones**: MNP (Moldtelecom National Plan), International
- **Hosts**: Ethernet, IP Addresses,
- **Images**: TIFF, GIF, and JPEG
- **Audio**: MP3
- **Video**: MPEG

Information Storage, Measure and Transmission

Symbol Space Size

The first question to address is the number of symbols that need to be encoded. This is called the symbol space size. We can have symbol spaces of different sizes:

- 1
- 2
- Integral power of 2
- Finite
- Infinite, Countable
- Infinite, Uncountable

Use of Spare Capacity

In many situations there are some unused code patterns, because the number of symbols is not an integral power of 2. There are many strategies to deal with this.Here are some:

- Ignore
- Map to other values
- Reserve for future expansion
- Use for control codes
- Use for common abbreviations

Information Storage, Measure and Transmission

- The rate of transmission of the information through some transmission channel is measured in *bauds*:

  $$1 baud = 1 bit/s$$

  and prefixes K, M, G, etc are similar as bor bits and bytes.
- The capacity of the transmission channel can be approximatelly computed if you know the maximal frequency allowed in this channel.
- As a rule of thumb, you can say that the rate of transmission might be at least the maximal frequency allowed in the channel.
- For example, if maximal frequency is $1KHz$, then the transmission rate can be at least 1Kbauds.

Information Storage, Measure and Transmission

Examples of transmission channels and the maximal frequency allowed:

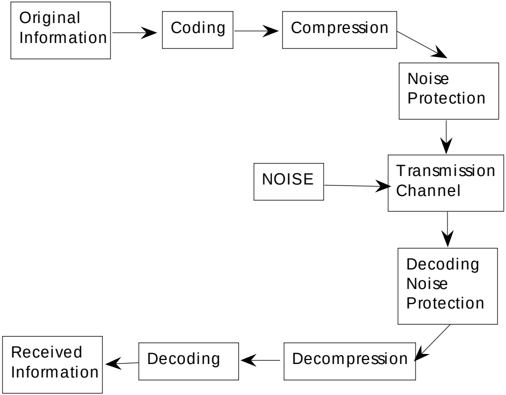
- telegraph : $140Hz$
- telephone : up to $3.1KHz$
- short waves SW $(10 - 100m)$: $3 - 30MHz$
- ultra short waves USW $(1 - 10m)$: $30 - 300MHz$
- satellites (centimeter waves) : up to $30\,GHz$
- optic fiber cable (infrared waves) : $0.15 - 400\,THz$
- optic fiber cable (light) : $400 - 700\,THz$
- optic fiber cable (ultraviolet waves) : $0.7 - 1.75PHz$

Information Storage, Measure and Transmission

- Usual transmission channels have been telegraph and telephone channels;
- Modern, implemented and fast develepping are optic fiber channels and digital telephone channel (ISDN, Integrated Services Digital Networks) – $57 - 128 Kbauds$;
- In optical fiber channels the achieved rate of transmission is still lower than theoretical limits, – up to $30 Gbauds$;
- Most used are still tlephone lines , up to $50 Kbauds$.

Information Storage, Measure and Transmission

A general scheme for information transmission:



Information Theory Introduction

- Equality $a = b$ contain information about the fact that quantity a equals quantity b.
- How about $a^2 = b^2$?
- Equality $a^2 = b^2$ contains **less** information, since $a = b \Rightarrow a^2 = b^2$
- Equality $a^3 = b^3$ contains the similar amount of information as $a = b$
- If some measurements are performed (with a ceratin accuracy), then more measurements will be performed then the more information about the measured quantity will be received.
- Expected value of a random variable contains some information about this random variable; If the random variable is distributed normally with a given standard deviation, then the expected value contains all the information about the random variable.

Information Theory Introduction

- Consider the process of transmision of some information. Mathematicalyy it can be written as

$$Y = X + Z$$

where X is the original transmitted information, Z is a random variable describing the noise in the channel and Y is the information received by receiver.

- In this context we can ask how much information is contained in Y with respect to X.
- Clearly, the smaller is the noise (standard deviation of Z is small), the more information about X is contained in Y.
- In case of zero noise, Y will contain all the information about X.

Information Theory Introduction

- In 1865, german physicist Rudolf Clausius introduced the notion of **enthropy** (degree of disorder) in physics.
- The **Second Law of Thermodynamics** states that enthropy is increasing!
- In 1921, british mathematician Ronald Fisher introduced the term **Information** in mathematics.
- In 1948 Clod Shennon have developed the formulas for information and enthropy. The name enthropy was suggested by the *father* of comuter science, american mathematician and physicist John von Neuman. Von Neumann observed that the formulas for information coincided miraculously with the fomulas being used in quantum physics.

Information and Enthropy

Definition

For random variables. X and Y, with distribution functions

$$P(X = X_i) = p_i, \quad and \quad P(Y = Y_j) = q_j$$

respecitvely, and joint distribution

$$P(X = X_i, Y = Y_j) = p_{ij}$$

information contained in X with respect to Y is

$$I(X, Y) = \sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_i q_j}$$

Information and Enthropy

Obviously

$$P(X = X_i, X = X_j) = \left\{ \begin{pmatrix} 0, & i \neq j \\ p_i, & i = j \end{pmatrix} \right.$$

and therefore,

$$I(X, X) = \sum_i p_i \log_2 \frac{p_i}{p_i p_i} = -\sum_i p_i \log_2 p_i$$

Definition

For a random variable X **enthropy** of X is defined as

$$H(X) = I(X, X) = -\sum_i p_i \log_2 p_i$$

Information and Enthropy

Theorem

$$\begin{array}{rrcl}
1) & I(X,Y) & > & 0, \quad \textit{and} \quad I(X,Y) = 0 \Leftrightarrow X \textit{ and } Y \textit{ are independent} \\
2) & I(X,Y) & = & I(Y,X) \\
3) & H(X) & = & 0 \Leftrightarrow X - \textit{constant} \\
4) & I(X,Y) & = & H(X) + H(Y) - H(X,Y), \ \textit{where} \\
 & H(X,Y) & = & -\sum_{i,j} p_{ij} \log_2 p_{ij} \\
5) & I(X,Y) & \leq & I(X,X)
\end{array}$$

Information and Enthropy

Proof. 1) Consider inequality

$$e^{x-1} \geq x \Leftrightarrow x - 1 \geq \ln x \Leftrightarrow \frac{x-1}{\ln 2} \geq \log_2 x$$

Then we have

$$\begin{aligned}
-I(X,Y) &= -\sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_i q_j} \\
&\leq \sum_{i,j} p_{ij} \frac{\frac{p_{ij}}{p_i q_j} - 1}{\ln 2} \\
&= \sum_{i,j} p_{ij} \frac{p_i q_j - p_{ij}}{\ln 2} \\
&= \frac{1}{\ln 2} \left(\sum_i p_i \sum_j q_j - \sum_{i,j} p_{ij} \right) \\
&= \frac{1}{\ln 2} (1 - 1) = 0
\end{aligned}$$

Information and Enthropy

Example. X_i = the result of a die roll, $i = 1, 2$. Also define $Y = X_1 + X_2$.

X_1	1	2	$\dots$	6
$P(X_1 = n)$	$1/6$	$1/6$	$\dots$	$1/6$

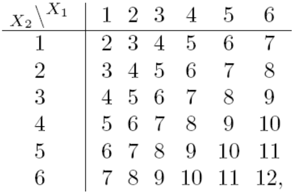
Same holds for X_2. Since X_1 and X_2 are independent

$$P(X_1 = n, X_2 = m) = P(X_1 = n)P(X_2 = m) = \frac{1}{36}$$

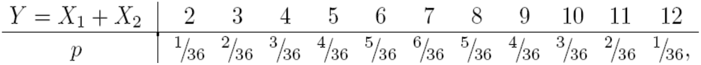
and

$$\begin{aligned} p_i &= P(Y = i) = P(X_1 + X_2 = i) = \sum_{n+m=i} P(X_1 = n)P(X_2 = m) \\ &= \sum_{n+m=i} \frac{1}{36} \end{aligned}$$

Information and Enthropy



$X_2 \backslash X_1$	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12,



$Y = X_1 + X_2$	2	3	4	5	6	7	8	9	10	11	12
p	$^1/_{36}$	$^2/_{36}$	$^3/_{36}$	$^4/_{36}$	$^5/_{36}$	$^6/_{36}$	$^5/_{36}$	$^4/_{36}$	$^3/_{36}$	$^2/_{36}$	$^1/_{36}$,

$$p_i = P(Y = i) = \frac{6 - |7 - i|}{36}$$

Information and Enthropy

Joint distribution of X_1 and Y :

$$p_{ij} = P(Y = i, X_1 = j) = P(Y = i | X_1 = j) P(X_1 = j)$$

For example,

$$\begin{aligned} P(Y &= 2, X_1 = 1) = P(Y = 2 | X_1 = 1) P(X_1 = 1) \\ &= P(X_2 = 1) P(X_1 = 1) = \frac{1}{36} \\ P(Y &= 3, X_1 = 4) = P(Y = 3 | X_1 = 4) P(X_1 = 4) \\ &= 0 \end{aligned}$$

$$p_{ij} = \begin{cases} 1/36, & 1 \leq i - j \leq 6 \\ 0, & \textit{otherwise} \end{cases}$$

Information and Enthropy

$X_1 \backslash Y$	2	3	4	5	6	7	8	9	10	11	12
1	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	0	0	0	0	0
2	0	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	0	0	0	0
3	0	0	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	0	0	0
4	0	0	0	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	0	0
5	0	0	0	0	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	0
6	0	0	0	0	0	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$	$^1/_{36}$

Information and Enthropy

$$
\begin{aligned}
I(Y, X_1) &= \sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_i q_j} \\
&= \sum_{j=1}^{6} \sum_{1 \le i-j \le 6} p_{ij} \log_2 \frac{p_{ij}}{p_i q_j} \\
&= \frac{1}{36} \sum_{j=1}^{6} \sum_{1 \le i-j \le 6} \log_2 \frac{1}{6 p_i} \\
& \quad \dots \\
&= \frac{1}{36} \left(10 + 24 \log_2 3 - 10 \log_2 5\right) \\
&\approx 0.69 \quad \mathit{bit\,/\,symbol}
\end{aligned}
$$

Information and Enthropy

$$\begin{aligned} H(X_1) &= I(X_1, X_1) = -\sum_{j=1}^{6} q_j \log_2 q_j \\ &= \frac{1}{6} \cdot 6 \cdot \log_2 6 = 1 + \log_2 3 \\ &\approx 2.58 \quad \mathit{bit\,/\,symbol} \end{aligned}$$

$$\begin{aligned} H(Y) &= I(Y, Y) = -\sum_{i=1}^{12} p_i \log_2 p_i \\ &= \frac{1}{36}(46 + 60 \log_2 3 - 10 \log_2 5) \\ &\approx 3.27 \quad \mathit{bit\,/\,symbol} \end{aligned}$$

Information and Enthropy

Let X be the result of a die roll, and $Y = 0$, if the result is odd and $Y = 1$, if it is even.

X	1	2	$\ldots$	6
$p_i = P(X = i)$	1/6	1/6	$\ldots$	1/6

Y	0	1
$q_j = P(Y = j)$	1/2	1/2

Joint distribuiton is

X	1	2	3	4	5	6	1	2	3	4	5	6
Y	0	1	0	1	0	1	1	0	1	0	1	0
p	1/6	1/6	1/6	1/6	1/6	1/6	0	0	0	0	0	0

$$p_{ij} = P(X = i, Y = j) = \left\{ \begin{array}{cc} 0, & i + j \text{ is even} \\ 1/6 & \text{otherwise} \end{array} \right.$$

Information and Enthropy

$$
\begin{aligned}
I(X,Y) &= I(Y,X) = \sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_i q_j} \\
&= 6 \cdot \frac{1}{6} \cdot \log_2 2 = 1 \quad bit\,/\,sym
\end{aligned}
$$

$$
\begin{aligned}
I(Y,Y) &= -\sum_{j=1}^{6} q_j \log_2 q_j = 2 \cdot \frac{1}{2} \log_2 2 \\
&= 1 \quad bit\,/\,sym
\end{aligned}
$$

The exact value of points on the die gives an exact information on evenness of the result.

Information and Enthropy

Enthropy of some random variable is the average number of bits that should be send through transmission channel about the current value of given random variable.

Example. In a horse race there are 4 horses..Suppose that these horses have equal chances of winning. Let RV X denote the number of winning horse: $1, 2, 3, 4$.

$$p_i = \frac{1}{4}, \quad i = 1, 2, 3, 4$$

Then

$$H(X) = -\sum_{i=1}^{4} p_i \log_2 p_i = 4 \cdot \frac{1}{4} \log_2 4 = 2 \quad bit/sym$$

Thus, after each race we can code the results as follows

$$1 - 00, \quad 2 - 01, \quad 3 - 10, \quad 4 - 11$$

So 2 bits will be sent in order to inform about the winning horse.

Information and Enthropy

Introduce function $Len(X) = length(code(X))$ that gives the length of the code that codifies the value of X. Then the expected value of $Len(X)$ is the average length of the code that codifies X.

Now, consider the case of variable X with the following distribution

X	1	2	3	4
$P(X=i)$	$\frac{3}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$

Then

$$\begin{aligned} H(X) &= -\sum_i p_i \log_2 p_i \\ &= \frac{3}{4}\log_2\frac{4}{3} + \frac{1}{8}\log_2 8 + 2\cdot\frac{1}{16}\log_2 16 \\ &= \frac{19}{8} - \frac{3}{4}\log_2 3 \approx 1.186 \quad bit\,/\,sym \end{aligned}$$

Information and Enthropy

So, the information about the winning horse can be sent using 1.186 bits. Clearly, the old code

$$1 - 00, \quad 2 - 01, \quad 3 - 10, \quad 4 - 11$$

will use

$$E(Len(X)) = \frac{3}{4} \cdot 2 + \frac{1}{8} \cdot 2 + \frac{1}{16} \cdot 2 + \frac{1}{16} \cdot 2 = 2 \quad bit\,/\,sym$$

Instead codify the result of the race as

$$1 - 0, \quad 2 - 10, \quad 3 - 110, \quad 4 - 111$$

i.e. such that each code is not the prefix of other codes. Such codes are called prefix codes. Then

$$\begin{aligned} E(Len(X)) &= \frac{3}{4} \cdot 1 + \frac{1}{8} \cdot 2 + \frac{1}{16} \cdot 3 + \frac{1}{16} \cdot 3 \\ &= \frac{11}{8} = 1.375 \quad bit\,/\,sym \end{aligned}$$

It can be proved that this code is the most efficient code in this case.