Lecture 4:
Entropy. Redundancy of information.

Bakibayev A.V.

Shannon's Information Theory

In 1948, Claude Shannon published a paper called A Mathematical Theory of Communication[1]. This paper heralded a transformation in our understanding of information. Before Shannon's paper, information had been viewed as a kind of poorly defined miasmic fluid. But after Shannon's paper, it became apparent that information is a well-defined and, above all, measurable quantity. Indeed, as noted by Shannon,

A basic idea in information theory is that information can be treated very much like a physical quantity, such as mass or energy. Caude Shannon, 1985.

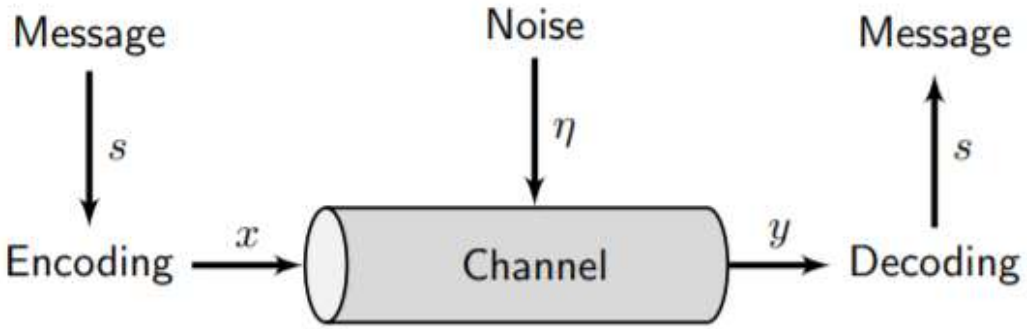


Figure 1: The communication channel. A message (data) is encoded before being used as input to a communication channel, which adds noise. The channel output is decoded by a receiver to recover the message.

Information theory defines definite, unbreachable limits on precisely how much information can be communicated between any two components of any system, whether this system is man-made or natural. The theorems of information theory are so important that they deserve to be regarded as the laws of information[2, 3, 4]. The basic laws of information can be summarised as follows. For any communication channel (Figure 1): 1) there is a definite upper limit, the channel capacity, to the amount of information that can be communicated through that channel, 2) this limit shrinks as the amount of noise in the channel increases, 3) this limit can very nearly be reached by judicious packaging, or encoding, of data.

Finding a Route, Bit by Bit

Information is usually measured in bits, and one bit of information allows you to choose between two equally probable, or equiprobable, alternatives. In order to understand why this is so, imagine you are standing at the fork in the road at point A in Figure 2, and that you want to get to the point marked D. The fork at A represents two equiprobable alternatives, so if I tell you to go left then you have received one bit of information. If we represent my instruction with a binary digit (0=left and 1=right) then this binary digit provides you with one bit of information, which tells you which road to choose.

Now imagine that you come to another fork, at point B in Figure 2. Again, a binary digit (1=right) provides one bit of information, allowing you to choose the correct road, which leads to C. Note that C is one of four possible interim destinations that you could

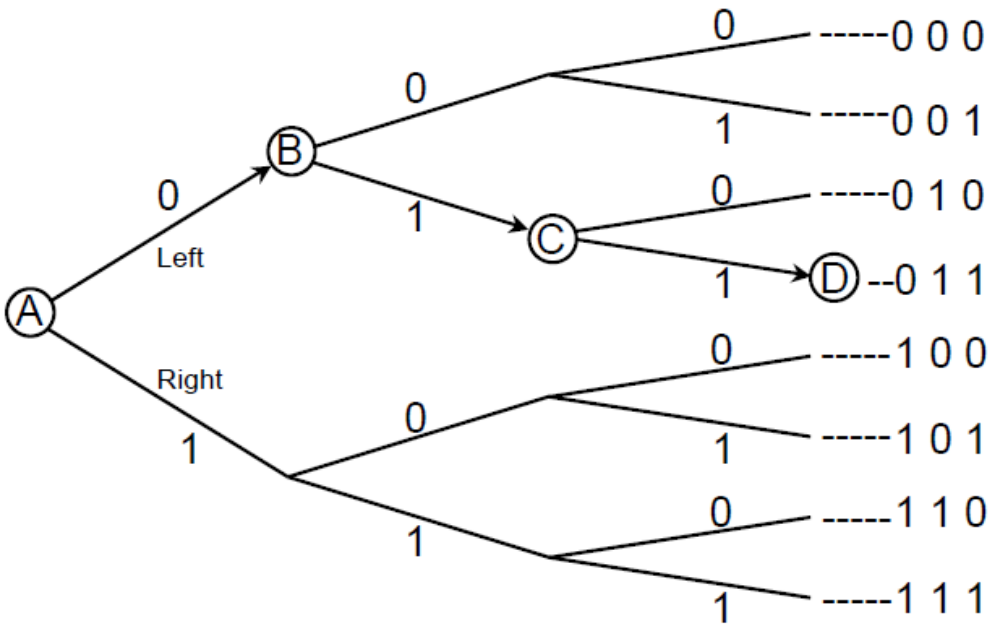


Figure 2: For a traveller who does not know the way, each fork in the road requires one bit of information to make a correct decision. The 0s and 1s on the right-hand side summarise the instructions needed to arrive at each destination; a left turn is indicated by a 0 and a right turn by a 1.

have reached after making two decisions. The two binary digits that allow you to make the correct decisions provided two bits of information, allowing you to choose from four (equiprobable) alternatives; 4 equals $2 \times 2 = 2^2$.

A third binary digit (1=right) provides you with one more bit of information, which allows you to again choose the correct road, leading to the point marked D. There are now eight roads you could have chosen from when you started at A, so three binary digits (which provide you with three bits of information) allow you to choose from eight equiprobable alternatives, which also equals $2x2x2 = 2^3 = 8$.

We can restate this in more general terms if we use n to represent the number of forks, and m to represent the number of final destinations. If you have come to n forks then you have effectively chosen from $m = 2_n$ final destinations. Because the decision at each fork requires one bit of information, n forks require n bits of information.

Viewed from another perspective, if there are m = 8 possible destinations then the number of forks is n = 3, which is the logarithm of 8. Thus, $3 = \log_2 8$ is the number of forks implied by eight destinations. More generally, the logarithm of m is the power to which 2 must be raised in order to obtain m; that is, $m = 2_n$. Equivalently, given a number m, which we wish to express as a logarithm, $n = \log_2 m$: The subscript 2 indicates that we are using logs to the base 2 (all logarithms in this book use base 2 unless stated otherwise).

Bits Are Not Binary Digits

The word bit is derived from binary digit, but a bit and a binary digit are fundamentally dierent types of quantities. A binary digit is the value of a binary variable, whereas a bit is an amount of information. To mistake a binary digit for a bit is a category error. In this case, the category error is not as bad as mistaking marzipan for justice, but it is analogous to mistaking a pint-sized bottle for a pint of milk. Just as a bottle can contain between zero and one pint, so a binary digit (when averaged over both of its possible states) can convey between zero and one bit of information.

Information and Entropy

Consider a coin which lands heads up 90% of the time (i.e. p(xh) = 0:9). When this coin is ipped, we expect it to land heads up (x = xh), so when it does so we are less surprised than when it lands tails up (x = xt). The more improbable a particular outcome is, the more surprised we are to observe it. If we use logarithms to the base 2 then the Shannon information or surprisal of each outcome is measured in bits (see Figure 3a)

$$\text{Shannon information} \quad = \quad \log \frac{1}{p(x_h)} \text{ bits}, \tag{1}$$

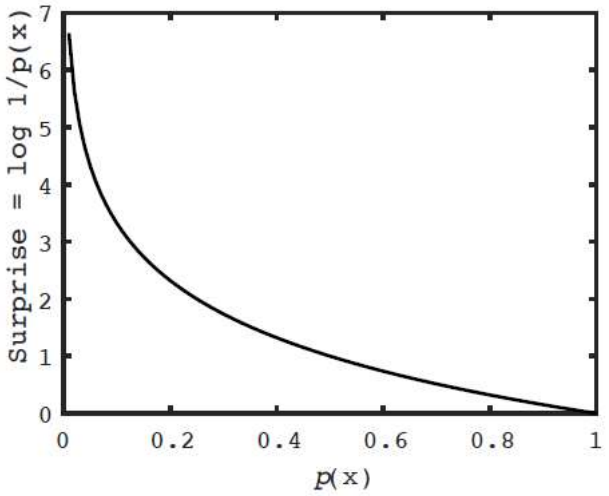
which is often expressed as: information $= -\log\, p(x_h)$ bits.

Entropy is Average Shannon Information. We can represent the outcome of a coin ip as the random variable x, such that a head is $x = x_h$ and a tail is $x = x_t$. In practice, we are not usually interested in the surprise of a particular value of a random variable, but we are interested in how much surprise, on average, is associated with the entire set of possible values. The average surprise of a variable x is dened by its probability distribution p(x), and is called the entropy of p(x), represented as H(x).

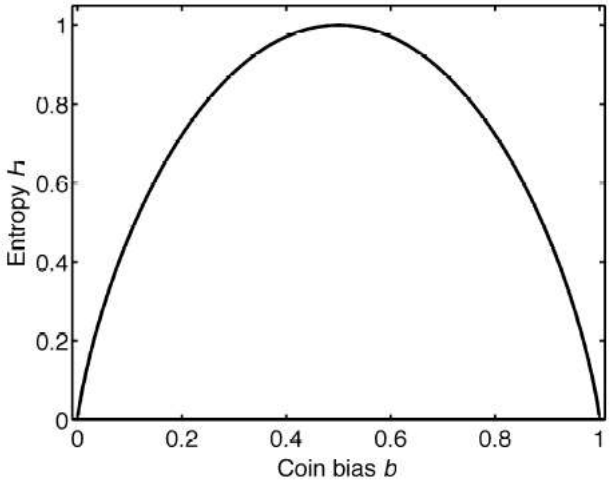
The Entropy of a Fair Coin. The average amount of surprise about the possible outcomes of a coin ip can be found as follows. If a coin is fair or unbiased then $p(x_h) = p(x_t) = 0{:}5$

then the Shannon information gained when a head or a tail is observed is log $1{=}0{:}5 = 1$ bit, so the average Shannon information gained after each coin ip is also 1 bit. Because entropy is dened as average Shannon information, the entropy of a fair coin is $H(x) = 1$ bit.

The Entropy of an Unfair (Biased) Coin. If a coin is biased such that the probability of a head is $p(x_h) = 0{:}9$ then it is easy to predict the result of each coin ip (i.e. with 90% accuracy if we predict a head for each ip). If the outcome is a head then the amount of Shannon information gained is $\log(1{=}0{:}9) = 0{:}15$ bits. But if the outcome is a tail then



(a)



(b)

Figure 3: a) Shannon information as surprise. Values of x that are less probable have larger values of surprise, defined as $\log_2(1/p(x))$ bits. b) Graph of entropy $H(x)$ versus coin bias (probability $p(x_h)$ of a head). The entropy of a coin is the average amount of surprise or Shannon information in the distribution of possible outcomes (i.e. heads and tails).

the amount of Shannon information gained is $\log(1/0.1) = 3.32$ bits. Notice that more information is associated with the more surprising outcome. Given that the proportion of flips that yield a head is $p(x_h)$, and that the proportion of flips that yield a tail is $p(x_t)$ (where $p(x_h) + p(x_t) = 1$), the average surprise is

$$H(x) = p(x_h) \log \frac{1}{p(x_h)} + p(x_t) \log \frac{1}{p(x_t)}, \tag{2}$$

which comes to $H(x) = 0.469$ bits, as in Figure 3b. If we define a tail as $x_1 = x_t$ and a head as $x_2 = x_h$ then Equation 2 can be written as

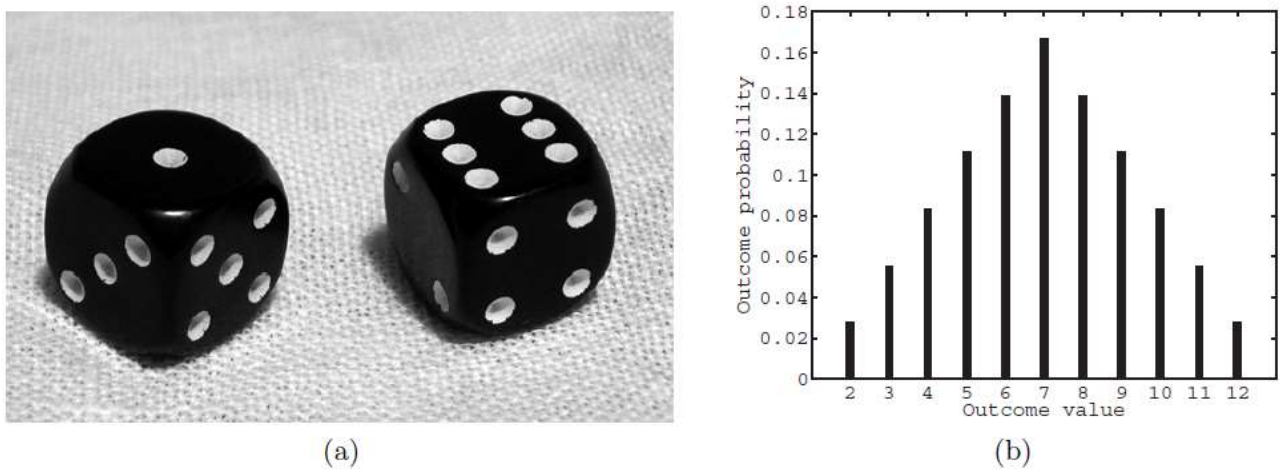
$$H(x) = \sum_{i=1}^{2} p(x_i) \log \frac{1}{p(x_i)} \text{bits}. \tag{3}$$

More generally, a random variable x with a probability distribution $p(x) = \{p(x_1), \ldots, p(x_m)\}$ has an entropy of

$$H(x) = \sum_{i=1}^{m} p(x_i) \log \frac{1}{p(x_i)} \text{ bits}. \tag{4}$$

The reason this denition matters is because Shannon's source coding theorem (see Section 7) guarantees that each value of the variable x can be represented with an average of (just over) H(x) binary digits. However, if the values of consecutive values of a random variable are not independent then each value is more predictable, and therefore less surprising, which reduces the information-carrying capability (i.e. entropy) of the variable. This is why it is important to specify whether or not consecutive variable values are independent.

Interpreting Entropy. If $H(x) = 1$ bit then the variable x could be used to represent $m = 2^{H(x)}$ or 2 equiprobable values. Similarly, if $H(x) = 0.469$ bits then the variable x



(a) (b)

Figure 4: (a) A pair of dice. (b) Histogram of dice outcome values.

could be used to represent m = $2_{0:469}$ or 1.38 equiprobable values; as if we had a die with 1.38 sides. At rst sight, this seems like an odd statement. Nevertheless, translating entropy into an equivalent number of equiprobable values serves as an intuitive guide for the amount of information represented by a variable.

Dicing With Entropy. Throwing a pair of 6-sided dice yields an outcome in the form of an ordered pair of numbers, and there are a total of 36 equiprobable outcomes, as shown in Table 1. If we dene an outcome value as the sum of this pair of numbers then there are m = 11 possible outcome values A_x = f2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12g, represented by the symbols x_1; : : : ; x_{11}. These outcome values occur with the frequencies shown in Figure 4b

and Table 1. Dividing the frequency of each outcome value by 36 yields the probability P of each outcome value. Using Equation 4, we can use these 11 probabilities to nd the entropy

$$H(x) = p(x_1)\log\frac{1}{p(x_1)} + p(x_2)\log\frac{1}{p(x_2)} + \cdots + p(x_{11})\log\frac{1}{p(x_{11})}$$
$$= 3.27 \text{ bits.}$$

Using the interpretation described above, a variable with an entropy of 3.27 bits can represent $2^{3.27} = 9.65$ equiprobable values.

Entropy and Uncertainty. Entropy is a measure of *uncertainty*. When our uncertainty is reduced, we gain information, so information and entropy are two sides of the same coin. However, information has a rather subtle interpretation, which can easily lead to confusion.

Average information shares the same definition as entropy, but whether we call a given quantity information or entropy depends on whether it is being given to us or taken away.

Symbol	Sum	Outcome	Frequency	P	Surprisal
x_1	2	1:1	1	0.03	5.17
x_2	3	1:2, 2:1	2	0.06	4.17
x_3	4	1:3, 3:1, 2:2	3	0.08	3.59
x_4	5	2:3, 3:2, 1:4, 4:1	4	0.11	3.17
x_5	6	2:4, 4:2, 1:5, 5:1, 3:3	5	0.14	2.85
x_6	7	3:4, 4:3, 2:5, 5:2, 1:6, 6:1	6	0.17	2.59
x_7	8	3:5, 5:3, 2:6, 6:2, 4:4	5	0.14	2.85
x_8	9	3:6, 6:3, 4:5, 5:4	4	0.11	3.17
x_9	10	4:6, 6:4, 5:5	3	0.08	3.59
x_{10}	11	5:6, 6:5	2	0.06	4.17
x_{11}	12	6:6	1	0.03	5.17

Table 1: A pair of dice have 36 possible outcomes.
Sum: outcome value, total number of dots for a given throw of the dice.
Outcome: ordered pair of dice numbers that could generate each symbol.
Freq: number of dierent outcomes that could generate each outcome value.
P: the probability that the pair of dice yield a given outcome value (freq/36).
Surprisal: P log(1=P) bits.

For example, if a variable has high entropy then our initial uncertainty about the value of that variable is large and is, by denition, exactly equal to its entropy. If we are told the value of that variable then, on average, we have been given information equal to the uncertainty (entropy) we had about its value. Thus, receiving an amount of information is equivalent to having exactly the same amount of entropy (uncertainty) taken away.

Shannon's Information Theory

Claude Shannon: ~~A~~ The Mathematical Theory of Communication

Bell System Technical Journal, 1948

- Shannon's measure of information is the number of bits to represent the amount of uncertainty (randomness) in a data source, and is defined as entropy

$$H = -\sum_{i=1}^{n} p_i \log(p_i)$$

Where there are n symbols 1, 2, … n, each with probability of occurrence of p_i

Example

Determine the entropy of the mobile phone screen

if its resolution is 320/240 and each pixel can reflect one of 4096 colors

decision
assume that the colors of the pixels are equally probable and mutually independent, then the entropy of one pixel is

$H_п = log_2 4096 = 12$ бит

total pixels 320*240=76800

and the entropy of the whole screen

$H_э = 76800*12 = 921600$ бит

if system A has n states A1, A2,…An and the probabilities of these states are respectively p1,p2,…pn; p1+p2+…+pn=1, then the entropy of system A is called the quantity

$$H(A) = -(p1\ log_2\ p1 + p2\ log_2\ p2 + \ldots + pn\ log\ pn)$$

$$\mathrm{H}(A) = -\sum_{i=1}^{n} pi\ log_2\ pi$$

Unit of entropy

One bit is the entropy of the simplest physical system, which can only be in the bottom of two states, and these states are equally probable

Example

Let system A have two states A1 and A2 with probabilities B1 and B2, then the entropy of such a system is:

$$H(A) = -(0{,}5 log_2 0{,}5 + 0{,}5 log_2 0{,}5) = 1$$

ЗНАЧЕНИЕ ФУНКЦИИ $-P\log_2 P$

Таблица

P	$-P\log_2 P$	P	$-P\log_2 P$	P	$-P\log_2 P$	P	$-P\log_2 P$
0.00	0.0000	0.26	0.5053	0.52	0.4906	0.78	0.2796
0.01	0.0664	0.27	0.5100	0.53	0.4854	0.79	0.2678
0.02	0.1129	0.28	0.5142	0.54	0.4800	0.80	0.2575
0.03	0.1517	0.29	0.5179	0.55	0.4744	0.81	0.2462
0.04	0.1857	0.30	0.5211	0.56	0.4684	0.82	0.2348
0.05	0.2161	0.31	0.5238	0.57	0.4623	0.83	0.2231
0.06	0.2435	0.32	0.5260	0.58	0.4558	0.84	0.2113
0.07	0.2686	0.33	0.5278	0.59	0.4491	0.85	0.1993
0.08	0.2915	0.34	0.5292	0.60	0.4422	0.86	0.1871
0.09	0.3127	0.35	0.5301	0.61	0.4350	0.87	0.1748
0.10	0.3322	0.36	0.5306	0.62	0.4276	0.88	0.1623
0.11	0.3503	0.37	0.5307	0.63	0.4199	0.89	0.1496
0.12	0.3671	0.38	0.5304	0.64	0.4121	0.90	0.1368
0.13	0.3826	0.39	0.5298	0.65	0.4040	0.91	0.1238
0.14	0.3971	0.40	0.5288	0.66	0.3957	0.92	0.1107
0.15	0.4105	0.41	0.5274	0.67	0.3871	0.93	0.0978
0.16	0.4230	0.42	0.5856	0.68	0.3784	0.94	0.0839
0.17	0.4346	0.43	0.5236	0.69	0.3694	0.95	0.0703
0.18	0.4453	0.44	0.5211	0.70	0.3602	0.06	0.0565
0.19	0.4552	0.45	0.5181	0.71	0.3508	0.97	0.0426
0.20	0.4644	0.46	0.5153	0.72	0.3412	0.98	0.0286
0.21	0.4728	0.47	0.5120	0.73	0.3314	0.99	0.0140
0.22	0.4806	0.48	0.5083	0.74	0.3215	1.00	0.0000
0.23	0.4877	0.49	0.5043	0.75	0.3113		
0.24	0.4941	0.50	0.5000	0.76	0.3009		
0.25	0.500	0.51	0.4954	0.77	0.2903		

entropy properties

1. entropy is always non-negative

since the probability values are expressed in quantities not exceeding 1, but their logarithms are negative numbers or 0

2. if pi = 1 (and everything else Pj = 0, j=1,…(n-1)) then H(A)=0
3. H(A)=Hmax, for p1=p2=…=pn=1/n
4. H(AB)=H(A)+H(B)

reconsider your answers