Basic Notions of Information Theory

1 Entropy, Relative Entropy, and Mutual Information

First we introduce some notions that are used in the forthcoming coding theory. Many of these notions (Shannon entropy, Shannon capacity) were named after Claude E. Shannon, American mathematician, engineer, computer scientist (born 100 years ago and died 15 years ago), whose book on the mathematical theory of communication was a breakthrough in 1949.

The entropy will be used to measure the average surprise when observing a random variable. If an event A of probability p happens we are less surprised when p is large, and more surprised if it is small. The measure of our surprise has to satisfy the following:

- our surprise over the sure event is 0;
- the surprise is a continuous, strictly decreasing function of p;
- our surprise over observing independent events is added together.

It can be proved that the function $-\log p$ does this job, where in the sequel under log the log base 2 function is understood (it will be convenient for binary coding).

If we have an observation for the rv X, then we are looking for the average surprise over its possible values.

Definition 1 *The entropy of the rv X is defined by*

$$H(X) = \mathbb{E}(-\log P(X))$$

if X has a discrete distribution with pmf $P(x)$, and

$$H(X) = \mathbb{E}(-\log f(X))$$

if X has an absolutely continuous distribution with pdf $f(x)$, provided the expectation exist. We also use the $0 \log 0 = 0 \log \frac{0}{0} = 0$ conventions.

We will mainly consider discrete rv's taking on finitely many values. We say that X has a finite alphabet if the range of it is the finite set $A = \{a_1, a_2, \dots a_{|A|}\}$. With the notation $P(a) = \mathbb{P}(X = a)$:

$$H(X) = -\sum_{a \in A} P(a) \log P(a),$$

which is nonnegative, finite, and – under fixed A – is the largest for the $\mathbb{P}$ which is the discrete uniform distribution over A, i.e., $P(a) = \frac{1}{|A|}$, $\forall a \in A$. Also $H(X) \leq \log|A|$ and $H(X) = 0$ if and only if X is constant with probability 1.

So $H(X)$ is the average amount of surprise or uncertainty received when the value of X is observed. Since $H(X)$ depends only on the distribution $\mathbb{P}$ of X, sometimes the notation $H(\mathbb{P})$ will be used.

The notion of entropy naturally extends to random vectors (multivariate distributions). Consider the bivariate, finite alphabet case. Say, X takes on values in the finite alphabet $A = \{a_1, \ldots, a_n\}$, Y takes on values in the finite alphabet $B = \{b_1, \ldots, b_m\}$, and let $R(a_i, b_j) = \text{Prob}(X = a_i, Y = b_j)$ $(i = 1, \ldots, n;\ j = 1, \ldots, m)$ be the pmf of their joint distribution. Let $\mathbb{P}$ and $\mathbb{Q}$ denote the marginal distributions with pmf's $P(a_i) = \sum_{j=1}^{m} R(a_i, b_j)$ for $i = 1, \ldots n$ and $Q(b_j) = \sum_{i=1}^{n} R(a_i, b_j)$ for $j = 1, \ldots, m$. (Note that when we take an i.i.d. sample $(X_1, Y_1), \ldots, (X_N, Y_N)$ from this bivariate distribution, the counts form an $n \times m$ contingency table.) The entropy of the random vector (X, Y) is

$$H(X, Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} R(a_i, b_j) \log R(a_i, b_j).$$

It is symmetric in X and Y.

Definition 2 *The conditional entropy of X conditioned on Y is the entropy of the conditional distribution of X conditioned on Y, and denoted by $H(X|Y)$.*

Remark 1 *If X and Y take on values in the finite alphabets A and B, then with the above notation,*

$$H(X|Y) = -\sum_{j=1}^{m}\sum_{i=1}^{n} R(a_i, b_j) \log \frac{R(a_i, b_j)}{Q(b_j)}.$$

This formula follows by the rule of conditional probabilities, akin to the following statement.

Proposition 1

$$H(X, Y) = H(Y) + H(X|Y) = H(X) + H(Y|X).$$

This statement means that the joint entropy of two rv's is the entropy of the one plus the entropy of the other conditioned on the one. The subsequent theorem states that the entropy (uncertainty) of a rv is decreased if another rv is observed, and it is the same if the two rv's are independent.

Theorem 1

$$H(X|Y) \leq H(X)$$

with equality if and only if X and Y are independent.

For the proof we need the following lemma.

Lemma 1 ***(Log-sum inequality.)*** *Let $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$ be non-negative numbers such that $\sum_{i=1}^n p_i = \sum_{i=1}^n q_i = 1$. Then*

$$-\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i$$

with equality if and only if $p_i = q_i$ $(i = 1, \ldots, n)$.

Proof of Theorem 1. Denoting by R the joint, whereas by P and Q the marginal pmf's:

$$\begin{aligned} H(X) + H(Y) &= -\sum_{i=1}^{n} P(a_i) \log P(a_i) - \sum_{j=1}^{m} Q(b_j) \log Q(b_j) \\ &= -\sum_{i=1}^{n} \sum_{j=1}^{m} R(a_i, b_j)[\log P(a_i) + \log Q(b_j)] \\ &= -\sum_{i=1}^{n} \sum_{j=1}^{m} R(a_i, b_j) \log(P(a_i)Q(b_j)) \\ &\geq -\sum_{i=1}^{n} \sum_{j=1}^{m} R(a_i, b_j) \log R(a_i, b_j) = H(X, Y), \end{aligned}$$

where the inequality follows by Lemma 1 (with nm terms and probabilities $R(a_i, b_j)$ versus $P(a_i)Q(b_j)$ both summing to 1, corresponding to the $\mathbb{R}$ and $\mathbb{P} \times \mathbb{Q}$ distributions, respectively). The inequality is attained with equality if and only if $R(a_i, b_j) = P(a_i)Q(b_j)$, $\forall i, j$, i.e., when X and Y are independent. This, together with Proposition 1 implies the statement. $\square$

Corollary 1

$$H(X) + H(Y) \geq H(X, Y)$$

with equality if and only if X and Y are independent.

It means that the joint uncertainty of two rv's cannot exceed the sum of their individual uncertainties; further, the uncertentainties are added together for independent rv's, and only for those.

Corollary 1 makes rise the following definition to measure the difference between the two sides of the inequality there.

Definition 3 *The mutual information between X and Y is*

$$I(X \wedge Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y).$$

By Theorem 1 and Corollary 1, $I(X \wedge Y) \geq 0$ with equality if and only if X and Y are independent; further, $I(X \wedge X) = H(X)$. Anyway, $I(X \wedge Y)$ is symmetric in X and Y.

Likewise, the mutual information of the random vector $\mathbf{X} = (X_1, \ldots, X_N)$ is

$$I(\mathbf{X}) = \sum_{i=1}^{N} H(X_i) - H(X_1, \ldots, X_N).$$

It is nonnegative and 0 if and only if $X_1, \ldots, X_N$ are independent.

Remark 2 *The joint entropy of the i.i.d. sample $X_1, \ldots, X_N$ is*

$$H(X_1, \ldots, X_N) = \sum_{i=1}^{N} H(X_i) = N \cdot H(X_1)$$

provided $H(X_1)$ exists. In case of discrete X_1 taking on values in a finite alphabet it holds, otherwise certain regularity conditions are needed. The additivity of $H(X_1, \ldots, X_N)$ also resembles to that of the Fisher information; however, the entropy rather measures the uncertainty, while the Fisher information the information content of the sample.

Remark 3 *The Independent Component Analysis (ICA) is a generalization of the Principal Component Analysis (PCA) which is mainly applicable to multivariate Gaussian random vectors (in this case the principal component transformation results in uncorrelated rv's, which are also independent). Given the p-dimensional random vector $\mathbf{X}$, ICA looks for an orthogonal transformation ($p \times p$ orthogonal matrix) $\mathbf{A}$ such that the mutual information of the components of the random vector $\mathbf{Y} = \mathbf{AX}$ is minimized. Since*

$$I(\mathbf{Y}) = \sum_{i=1}^{p} H(Y_i) - H(\mathbf{Y}) = \sum_{i=1}^{p} H(Y_i) - H(\mathbf{X}) - \log|\det(\mathbf{A})| = \sum_{i=1}^{p} H(Y_i) - H(\mathbf{X})$$

and $H(\mathbf{X})$ is fixed, the entropy of Y_i's should be minimized. Note that, among among the absolutely continuous distributions with given variance the Gaussian distribution has the largest entropy. Therefore we are looking for the orthogonal transformation $\mathbf{A}$ which maximizes the departure of the distribution of Y_i's from the Gaussian. There are algorithms to do so (based on a sample and using empirical values of the resulting entropies). ICA makes sense if

$\mathbf{X}$ *follows a non-Gaussian multivariate distribution, where independence is much stronger than pairwise uncorrelatedness. If* $\mathbf{X}$ *were multivariate Gaussian, then the usual PCA would result in a multivariate Gaussian* $\mathbf{Y}$ *with independent components and* $I(\mathbf{Y})$ *would be zero, it would not make sense to decrease it.*

Now we introduce a measure for the deviation of two distributions over the same finite alphabet.

Definition 4 *The information divergence (briefly, I-divergence) or, equivalently, Kullbach–Leibler distance of the distributions* $\mathbb{P}$ *and* $\mathbb{Q}$ *over the same finite alphabets* A *is*

$$D(\mathbb{P}|\mathbb{Q}) = \sum_{a\in A} P(a) \log \frac{P(a)}{Q(a)}.$$

It is sometimes called the relative entropy of $\mathbb{P}$ *with respect to* $\mathbb{Q}$*.*

By Lemma 1, $D(\mathbb{P}|\mathbb{Q}) \geq 0$ with equality if and only if $\mathbb{P} = \mathbb{Q}$. Note that here the role of $\mathbb{P}$ and $\mathbb{Q}$ is not symmetric. Observe that

$$D(\mathbb{P}|\mathbb{Q}) = \mathbb{E}_{\mathbb{P}} \log \frac{P(X)}{Q(X)}.$$

Using this fact, the notion of I-divergence naturally extends to absolutely continuous distributions $\mathbb{P}$ and $\mathbb{Q}$ with pdf's $f(x)$ and $g(x)$, respectively. In this case $D(\mathbb{P}|\mathbb{Q}) \geq 0$ follows by the Jensen inequality, using that $-\log x$ is a convex function:

$$\begin{aligned}\mathbb{E}_{\mathbb{P}} \log \frac{f(X)}{g(X)} &= \mathbb{E}_{\mathbb{P}}[-\log \frac{g(X)}{f(X)}] \geq -\log \mathbb{E}_{\mathbb{P}} \frac{g(X)}{f(X)} \\ &= -\log \int \frac{g(x)}{f(x)} f(x)\, dx = -\log 1 = 0.\end{aligned}$$

Remark 4 *An easy calculation shows that*

$$I(X \wedge Y) = D(\mathbb{R}|\mathbb{P} \times \mathbb{Q}),$$

that is the mutual information of X *and* Y *is the I-divergence between their joint distribution* $\mathbb{R}$ *and the distribution obtained as the Cartesian product of their marginal distributions* $\mathbb{P}$ *and* $\mathbb{Q}$*, which corresponds to the independent case. This is another proof that the mutual information of two rv's is 0 if and only if they are independent.*

2 Coding Theory and Entropy

We want to send a sequence of possible values (from a finite alphabet A) of X from a source to a destination through a binary channel. First we have to encode all possible values of X into a sequence of 0's and 1's. In order the destination be able to decode it, we have to use so-called *prefix coding*: no encoded sequence can be obtained from a shorter encoded sequence by adding more terms to the shorter. We will show that it is possible, and if the channel is noiseless, we also want to minimize the number of bits (binary digits) that need be sent through the channel. If $A = \{a_1, \ldots, a_n\}$, then $L(a_i)$ denotes the length of the codeword assigned to a_i $(i = 1, \ldots, n)$.

For example, mannequins are arriving to a fashion sow one by one, wearing red, white, green, or blue dresses. A reporter wants to broadcast the consecutive colors through a binary channel. A possible prefix coding of the four colors is:

$$\begin{aligned} a_1 &\Leftrightarrow 00 \\ a_2 &\Leftrightarrow 01 \\ a_3 &\Leftrightarrow 10 \\ a_4 &\Leftrightarrow 11 \end{aligned} \tag{1}$$

here all the lengths are 2. Another prefix coding is

$$\begin{aligned} a_1 &\Leftrightarrow 0 \\ a_2 &\Leftrightarrow 10 \\ a_3 &\Leftrightarrow 110 \\ a_4 &\Leftrightarrow 111 \end{aligned} \tag{2}$$

here the lengths are 1, 2, or 3. However, the following coding is not prefix:

$$\begin{aligned} a_1 &\Leftrightarrow 0 \\ a_2 &\Leftrightarrow 1 \\ a_3 &\Leftrightarrow 00 \\ a_4 &\Leftrightarrow 01 \end{aligned}$$

let us forget it.

Exercise 1 *Try to encode, then decode a sequence of realizations $x_1, \ldots, x_N$ (dress colors of N consecuitive mannequins) with $N = 7$ and coding systems (1) and (2).*

Assume that the distribution of the colors is:

$$P(a_1) = \frac{1}{2}, \quad P(a_2) = \frac{1}{4}, \quad P(a_3) = \frac{1}{8}, \quad P(a_4) = \frac{1}{8}. \tag{3}$$

Then the coding (1) would expect to send $\frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 2 + \frac{1}{8} \cdot 2 = 2$ bits; whereas, the coding (2) would expect to send $\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = 1.75$ bits per signal, on average. Hence, for the above distribution, the coding (2) is more efficient than the coding (1). This raises the following question: What is the maximum efficiency achievable by a coding scheme? It will depend on $H(X)$, i.e. the entropy $H(\mathbb{P})$ of the distribution of X. The forthcoming coding theorem, due to Shannon, answers this question.

Lemma 2 ***(Kraft inequality.)*** *Let X take on its possible values in the finite alphabet $A = \{a_1, \ldots, a_n\}$. Then $L(a_1), \ldots, L(a_n)$ are the lengths of some prefix code if and only if they satisfy the so-called Kraft inequality:*

$$\sum_{i=1}^{n} \left(\frac{1}{2}\right)^{L(a_i)} \leq 1.$$

Proof of Lemma 2. Let m_j denote the multiplicity of the length j among the lengths $L(a_1), \ldots, L(a_n)$, $j = 1, 2, \ldots$, but of course there are finitely many different lengths (maximum n).

Since there are at most 2 binary codes of length 1, $m_1 \leq 2$. Furthermore, as no binary sequence is allowed to be the extension of any other, $m_2 \leq 2^2 - 2m_1$. Indeed, there are at most 2^2 binary codes of length 2, but $2m_1$ of them are not prefix. With he same reasoning we must have

$$m_k \leq 2^k - m_1 2^{k-1} - m_2 2^{k-2} - \cdots - m_{k-1} 2, \qquad k = 1, \ldots$$

It is easy to see that these conditions are not only necessary but are also sufficient to have a prefix coding with the above lengths. Rewriting the above inequality as

$$m_k + m_{k-1} 2 \cdots + m_2 2^{k-2} + m_1 2^{k-1} \leq 2^k$$

and dividing by 2^k yields the necessary and sufficient condition

$$\sum_{j=1}^{k} m_j \left(\frac{1}{2}\right)^j \leq 1 \qquad \forall k \in \mathbb{N}.$$

However, as $\sum_{j=1}^{k} m_j \left(\frac{1}{2}\right)^j$ is increasing in k, it follows, that the above inequality is true if and only if

$$\sum_{j=1}^{\infty} m_j \left(\frac{1}{2}\right)^j \leq 1.$$

But by the definition of m_j, it follows that

$$\sum_{j=1}^{\infty} m_j \left(\frac{1}{2}\right)^j = \sum_{i=1}^{n} \left(\frac{1}{2}\right)^{L(a_i)}$$

which finishes the proof of the lemma. □

Theorem 2 ***(Noiseless coding theorem.)*** *Let $\mathbb{P}$ be probability distribution on the finite alphabet A. Then each prefix code has expected length*

$$\mathbb{E}(L) = \sum_{a \in A} P(a)L(a) \geq H(\mathbb{P}).$$

Furthermore, there is a prefix code with length function $L(a) = \lceil -\log P(a) \rceil$, the expected length of which satisfies

$$\mathbb{E}(L) < H(\mathbb{P}) + 1.$$

Proof. Since

$$\sum_{i=1}^{n} \frac{2^{-L(a_i)}}{\sum_{j=1}^{n} 2^{-L(a_j)}} = 1,$$

by Lemma 1 (log-sum inequality) it follows that

$$\begin{aligned} H(\mathbb{P}) &= -\sum_{i=1}^{n} P(a_i) \log P(a_i) \leq -\sum_{i=1}^{n} P(a_i) \log \left(\frac{2^{-L(a_i)}}{\sum_{j=1}^{n} 2^{-L(a_j)}} \right) \\ &= \sum_{i=1}^{n} P(a_i)L(a_i) + \sum_{i=1}^{n} P(a_i) \log \left(\sum_{j=1}^{n} 2^{-L(a_j)} \right) \\ &= \sum_{i=1}^{n} P(a_i)L(a_i) + \log \left(\sum_{j=1}^{n} 2^{-L(a_j)} \right) \leq \sum_{i=1}^{n} P(a_i)L(a_i), \end{aligned}$$

where in the last inequality we used Lemma 2 (Kraft inequality). According to this, $\log \left(\sum_{j=1}^{n} 2^{-L(a_j)} \right) \leq \log 1 = 0$.

Now we will prove the other part of the theorem: show that it is possible to devise a code such that the average number of bits is within 1 of $H(\mathbb{P})$. With the choice $L(a_i) = \lceil -\log P(a_i) \rceil$ $(i = 1, \ldots, n)$:

$$-\log P(a_i) \leq L(a_i) < -\log P(a_i) + 1, \qquad (i = 1, \ldots, n). \tag{4}$$

Due to the first inequality,

$$\sum_{i=1}^{n} 2^{-L(a_i)} \leq \sum_{i=1}^{n} 2^{\log P(a_i)} = \sum_{i=1}^{n} P(a_i) = 1,$$

and hence, the lengths satisfy the Kraft inequality. Therefore, we can associate sequences of bits of length $L(a_i)$ to a_i $(i = 1, \dots, n)$. Finally, the inequalities (4) imply that

$$H(\mathbb{P}) \le \sum_{i=1}^{n} P(a_i)L(a_i) < H(\mathbb{P}) + 1.$$

This finishes the proof. □

Exercise 2 *Revisiting the mannequin problem: prove that the coding (2) is the best possible for the distribution (3).*

Exercise 3 *Which of the codings (1), (2) is the more effective if the distribution of the colors is uniform?*

Exercise 4 *The next exercise is from the lecture notes of Péter Major, based on his lectures hold at the University of Szeged on Information Theory. I give a short review only. If we are not experts in football, we have to fill in 3^n TOTO-tickets to surely get a full-hit ($n = 13$). Assume that someone who knows more about football, can forecast the 3 possibilities with probabilities p_1, p_2, p_3 which differ from the uniform $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ law. Then we can have a full-hit with high probability with filling in about $2^{nH(\mathbb{P})}$ TOTO-tickets, where*

$$H(\mathbb{P}) = -\sum_{i=1}^{3} p_i \log p_i.$$

We can do it by selecting so-called typical sequences which contain about np_1 1's, np_2 2's, and np_3 x's (mode of these signals by filling in a ticket, or it also follows by the weak law of the large numbers). The probability if such a sequence is $\prod_{i=1}^{3} p_i^{np_i}$. Since the sum of the probabilities of the typical sequences is almost 1, approximately we can fill in $\frac{1}{\prod_{i=1}^{3} p_i^{np_i}}$ tickets with them. Easy calculation shows that this number is just $2^{nH(\mathbb{P})}$.

When $\mathbb{P}$ is the uniform law, then $H(\mathbb{P}) = \log 3 \approx 1.585$ and

$$2^{nH(\mathbb{P})} = 2^{13 \log 3} = 3^{13}$$

the number of bets of a non-expert. When $\mathbb{P}$ has pmf $\frac{1}{2}, \frac{1}{3}, \frac{1}{6}$, then $H(\mathbb{P}) \approx 1.459$ and

$$2^{nH(\mathbb{P})} = 2^{13 \cdot 1.459} = (3^{13})^{0.92} < 3^{13},$$

so the number of bets to be done is smaller for an expert, with high probability. More accurate results can be obtained by large deviations.

Note that in the noiseless setup the sink and destination can best collaborate, since there are not disturbances in the channel. Shannon also proved a noisy coding theorem, and there are several other variants too.