5.1.3 Conditional entropy and mutual information

Consider a joint chance event (X, Y) with probability $p(x, y)$ that $X = x$ and $Y = y$. The marginal distributions of X and Y are given by $p(x) = \sum_y p(x, y)$ and $p(y) = \sum_x p(x, y)$. The conditional probability of $X = x$ given $Y = y$ is defined by $p(x|y) = p(x, y)/p(y)$. Likewise, $p(y|x) = p(x, y)/p(x)$. [1]

The entropy of the joint ensemble (X, Y) is given by (this is not a new definition)

$$H(X, Y) = \sum_{x,y} -p(x, y) \log p(x, y).$$

The marginal entropies are given by

$$H(X) = \sum_{x,y} -p(x, y) \log \left(\sum_y p(x, y) \right),$$

$$H(Y) = \sum_{x,y} -p(x, y) \log \left(\sum_x p(x, y) \right).$$

The *conditional entropy* of X given Y is defined by

$$H(X|Y) = \sum_{x,y} -p(x, y) \log p(x|y).$$

Likewise,

$$H(Y|X) = \sum_{x,y} -p(x, y) \log p(y|x).$$

It is easily verified that we have the identity

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

The conditional entropy is always upperbounded by the unconditional entropy, $H(X|Y) \leq H(X)$. This may be proved as follows.

$$\begin{aligned}
H(X|Y) - H(X) &= \log(e) \left[\sum_{x,y} p(x, y) \ln p(x)/p(x|y) \right] \\
&\leq \log(e) \left[\sum_{x,y} p(x, y) \left(p(x)/p(x|y) - 1 \right) \right] \\
&= \log(e) \left[\sum_{x,y} p(x, y) p(x)/p(x|y) - \sum_{x,y} p(x, y) \right] \\
&= 0,
\end{aligned}$$

[1] To prevent confusion, we may use the full notation $p_X(x)$, $p_Y(y)$, $p_{X,Y}(x, y)$, $p_{X|Y}(x|y)$, $p_{Y|X}(y|x)$.

where in the last line we used $p(x,y)p(x)/p(x|y) = p(x)p(y)$. Note that equality holds iff $p(x,y) = p(x)p(y)$, i.e., X and Y are independent.

$H(X|Y)$ may be interpreted as the average uncertainty remaining about X when one is informed of Y. The inequality $H(X|Y) \leq H(Y)$ shows that knowing Y on average cannot increase the uncertainty about X.

The *mutual information* between X and Y is defined as

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

It is easy to verify that all three definitions above are equivalent. An immediate property is symmetry, $I(Y;X) = I(X;Y)$. Another property is nonnegativity, $I(X;Y) \geq 0$, with equality iff X and Y are independent, which has already been proved above.

These definitions generalize to arbitrary number of random variables as follows. Let $(X_1, \ldots, X_n)$ be a joint ensemble of random variables with a joint probability distribution $p(x_1, \ldots, x_n)$. One may write virtually unlimited number of identities involving these random variables, some of which are as follows.

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2)$$

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1, \ldots, X_{n-1})$$

$$I(X_1; X_2, X_3) = H(X_1) - H(X_1|X_2, X_3)$$

A sequence of random variables $X_1, \ldots, X_n$ is said to form a Markov chain if

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_n|x_{n-1}) = p(x_1)\prod_{i=2}^{n} p(x_i|x_{i-1}).$$

We write $X_1 \to X_2 \to \cdots \to X_n$ to denote a Markov chain.

Two important properties relating to Markov chains are as follows.

1. Strong subadditivity. $H(X,Y,Z) + H(Y) \leq H(X,Y) + H(Y,Z)$ with equality iff $X \to Y \to Z$ forms a Markov chain.

2. Data processing. Let $X \to Y \to Z$ be a Markov chain. Then, $I(X;Z) \leq I(X;Y)$. Thus, given X and Y, no data-processing on Y alone will generate a Z that conveys more information about X than Y does.

The proofs are left for homework exercises.

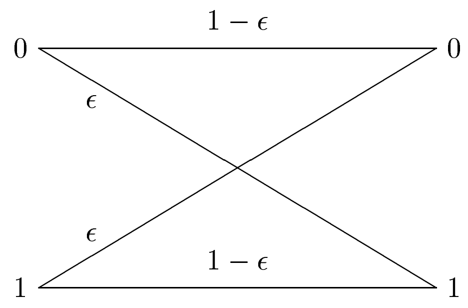

Figure 5.1: Binary symmetric channel.

5.1.4 Noisy channel coding theorem

The mutual information $I(X;Y)$ gains an operational significance when we consider X as the input Y as the output of a noisy communication channel. A simple example of a communication channel is a binary symmetric channel (BSC) as shown in Fig. 5.1. Here, X and Y take the values 0 and 1 and the channel is characterized by the *transition probabilities* $p(y|x)$, which are shown in the figure. Thus, when x is sent over a BSC, the output y equals x with probability $1-\epsilon$, and $y \neq x$ with probability ϵ. The parameter ϵ is called the crossover or error probability. One may assume that $0 \leq \epsilon \leq 1/2$ by relabeling the output letters if necessary. The case $\epsilon = 0$ corresponds to a noiseless channel and $\epsilon = 1/2$ to a useless channel.

More generally, a discrete memoryless channel (DMC) is defined as a system with an input alphabet $\mathcal{X}$ an output alphabet $\mathcal{Y}$ and a set of transition probabilities $\{p(y|x);\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$. The channel is memoryless in the sense that if a sequence of input letters $x_1, \ldots, x_N$ is transmitted across the channel, the sequence of output letters $y_1, \ldots, y_N$ is received with probability

$$p(y_1, \ldots, y_N | x_1, \ldots, x_N) = \prod_{n=1}^{N} p(y_n|x_n).$$

Shannon considered the transmission of information across such noisy communication channels, using a system as shown in Fig. 5.2. The source generates a message m, which is chosen at random uniformly over the set $\{1, \ldots, M\}$ of all possible messages. The message m

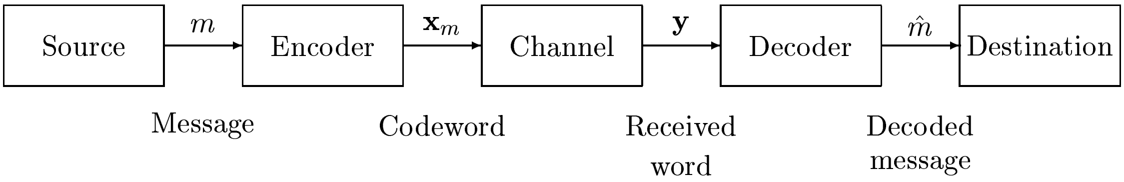

Figure 5.2: Block diagram of a communication system.

is encoded into a codeword $\mathbf{x}_m$, which consists of a block of N channel input letters, $\mathbf{x}_m = (x_{m,1}, \ldots, x_{m,N})$. The choice of codewords and the block length N are design variables. The codeword $\mathbf{x}_m$ is then transmitted over the noisy channel and a channel output block $\mathbf{y} = (y_1, \ldots, y_N)$ is received with probability

$$p(\mathbf{y}|\mathbf{x}_m) = \prod_{n=1}^{N} p(y_n|x_{m,n}),$$

where the transition probabilities are assumed given and not subject to design. The decoder then estimates which codeword was sent and produces an estimate $\hat{m}$ of the source output. The decision by the decoder may be described in terms of "decoding regions" $\mathcal{D}_1, \ldots, \mathcal{D}_M$ which are disjoint sets whose union equals the "observation space" $\mathcal{Y}^N$. Thus, the decoder sets $\hat{m} = i$ iff the channel output (observation) $\mathbf{y} \in \mathcal{D}_i$. The decoder design problem is one of partitioning the observation space into such decoding sets.

The goal in the design of the code book and the decoder is three-fold: (i) to maximize the rate of information transmission per channel use, as measured by

$$R = \frac{1}{N} \log_2 M \ \text{ bits};$$

(ii) to minimize the probability of decoding error,

$$P_e = \text{Prob}[\hat{m} \neq m];$$

and (iii) to minimize system complexity. Typically, a communication engineer is specified a desired rate R, such as 9,600 bits/sec, and a reliability level, such as $P_e \leq 10^{-5}$, and the task is to design the system with as low a complexity as possible so as to meet these goals.

Shannon (1948) was able to show that associated with every channel there is a *capacity* above which reliable communication is impossible, regardless of complexity. The capacity of a DMC is defined as

$$C = \max_{p(x)} I(X;Y) = \max_{p(x)} \sum_{x,y} p(x)p(y|x) \log \left(\frac{p(y|x)}{\sum_{x'} p(x')p(y|x')} \right).$$

The maximization here is over all possible probability assignments on the channel input alphabet.

For example, for the BSC, the capacity is obtained by the uniform distribution $p(0) = p(1) = 1/2$ and is given by

$$C = 1 - H(\epsilon) = 1 + \epsilon \log(\epsilon) + (1-\epsilon)\log(1-\epsilon).$$

A precise statement of Shannon's noisy channel coding theorem is as follows.

Proposition 1 *Consider a DMC with capacity C. (i) Any rate R below channel capacity is achievable in the sense that, given any $\delta > 0$, no matter how small, one can design an encoder-decoder pair such that the code rate is $\geq R$ and $P_e \leq \delta$. (ii) Conversely, rates above the channel capacity are unachievable. Given any rate $R > C$, there exists a $\delta > 0$ such that for all encoder-decoder pairs with rate $\geq R$, the probability of error is bounded away from zero by $P_e \geq \delta$.*

We will sketch a proof of this theorem for the special case of a BSC. First, we prove the converse statement. Consider a code with M codewords and block length N. When a certain codeword $\mathbf{x}_m$ is transmitted, the received word $\mathbf{y}$ typically differs from $\mathbf{x_m}$ in $\approx N\epsilon$ coordinates, where ϵ is the crossover probability. To state this more precisely, let $d(\mathbf{x_m}, \mathbf{y})$ denote the number of positions in which $\mathbf{x_m}$ and $\mathbf{y}$ differ; this is called the Hamming distance. Thus, $d(\mathbf{x_m}, \mathbf{y})$ equals the number of errors that occurred in the channel and by the law of large numbers

$$P[d(\mathbf{x_m}, \mathbf{y}) > N(\epsilon + \delta)] \leq o(N),$$

for any fixed $\delta > 0$, where $o(N)$ is a quantity that goes to zero as $N \to \infty$. In other words, with overwhelming probability the received word $\mathbf{y}$ lies within a sphere of radius $N\epsilon$ around the transmitted point. For reliable communication, the decoding region $\mathcal{D}_m$ must contain substantially all points in this sphere. To estimate the number of points in the sphere, let

$B(\mathbf{x}_m, r) = \{\mathbf{y} : d(\mathbf{x_m}, \mathbf{y}) \leq Nr\}$. This is the sphere around $\mathbf{x}_m$ of radius Nr. We have the estimate

$$|B(\mathbf{x}_m, \epsilon)| = \sum_{k=0}^{N\epsilon} \begin{pmatrix} N \\ k \end{pmatrix} \geq \begin{pmatrix} N \\ N\epsilon \end{pmatrix}.$$

By the Stirling approximation

$$n! = n^n e^{-n+o(n)},$$

it is easy to show that

$$\begin{pmatrix} N \\ k \end{pmatrix} = 2^{N[H(k/N)+o(N)]}.$$

Using this, we obtain

$$|B(\mathbf{x}_m, \epsilon)| \geq 2^{N[H(\epsilon)+o(N)]}.$$

Since the decoding sets $\{\mathcal{D}_m\}$ must be disjoint and since each set must contain $\geq 2^{N[H(\epsilon)+o(N)]}$ points for reliable communication, we have the following bound on the number of codewords,

$$M 2^{N[H(\epsilon)+o(N)]} \leq 2^N,$$

where 2^N is the total number of points in the observation space. Equivalently,

$$M \leq 2^{N[1-H(\epsilon)-o(N)]}.$$

Taking logarithms of both sides and dividing through by N and dropping asymptotically vanishing terms, we obtain the requirement that

$$R = \frac{1}{N} \log M \leq 1 - H(\epsilon)$$

for reliable communication. Recalling that the capacity of the BSC is given by $C = 1 - H(\epsilon)$, we see that $R < C$ is required for reliable communication over the BSC.

We now sketch the proof for the achievability part. It is interesting that no one knows how to compactly describe near optimal codes. Shannon gave a brilliant proof based on what is called "random-coding." The idea is to show that a code whose codewords are selected at random has a high probability of being a good code. So, let us consider such a code where each codeword is selected independently of each other and from the uniform distribution over the channel input blocks of length N, namely, the set $\{0,1\}^N$ of binary sequences. Let the decoding sets be

$$\mathcal{D}_m = \{\mathbf{y} : \ d(\mathbf{x}_m, \mathbf{y}) < d(\mathbf{x}_{m'}, \mathbf{y}), \text{ for all } m' \neq m\},$$

i.e, the rule is to decode to the codeword nearest to the received word in Hamming distance.[2]

With probability $1/M$, the source produces message m, which gets encoded into a codeword $\mathbf{x}_m$, which in turn gets sent over the channel, and a channel output word $\mathbf{y}$ is received. With probability $1 - o(N)$, we have $d(\mathbf{x}_m, \mathbf{y}) \leq N(\epsilon + \delta)$, where $\delta > 0$ is a fixed parameter independent of N. Thus, no decoding error will occur if the codewords for all other messages happen to be at greater distances from $\mathbf{y}$. The probability that codeword $\mathbf{x}_{m'}$, $m' \neq m$, satisfies $d(\mathbf{x}_{m'}, \mathbf{y}) \leq N(\epsilon + \delta)$ is given by

$$\frac{|B(\mathbf{y}, \epsilon + \delta)|}{2^N} \approx 2^{-N[1-H(\epsilon+\delta)]},$$

in accordance with the assumption that codewords are selected at random. By the union bound, the probability that *any* codeword other than m lies at distance $\leq N(\epsilon + \delta)$ from $\mathbf{y}$ is upper bounded by

$$(M-1)2^{-N[1-H(\epsilon+\delta)]} = 2^{N[R-1+H(\epsilon+\delta)]}.$$

Since $\delta > 0$ can be chosen arbitrarily small, we conclude that, as long as $R < 1 - H(\epsilon) = C$, the average probability of decoding error for a code chosen at random will be exponentially small in N for all N large enough. This establishes the achievability and completes the proof.

[2]This rule minimizes the probability of decoding error for the BSC, but this does not have to be proved here.