ITCT Lecture 4.2:
Channel Capacity and the Binary Symmetric Channel

2.1 Maximization of Mutual Information and Channel Capacity

Each time the transmitter sends a symbol, it is said to use the channel.

The channel capacity is the maximum average amount of information that can be sent per channel use.

why is this not the same as the mutual information ？

The mutual information is a function of the probability distribution of X.

By changing P_X ,we get different results of $I(X;Y)$.

For a fixed transition probability matrix, a change in P_X also results in a different output symbol distribution Q_Y.

The maximum mutual information achieved for a given transition probability matrix is the channel capacity.

$$C_x = \max_{P_X} I(X;Y)$$

C_x having units of bits per channel use.

An analytical closed-form solution to find C_x is difficult to achieve for an arbitrary channel. Fortunately, an efficient numerical technique for finding C_x was derived in 1972 by Blahut and Arimoto. The algorithm is based on the fact that C_x can be both upper and lower bounded by some simple functions of P_X. The algorithm recursively updates P_X based on these bounds.

- S. Arimoto, “An algorithm for calculating the capacity of an arbitrary DMC,” *IEEE Trans. Inform. Theory*, vol. 18, pp. 14-20, 1972.
- R. E. Blahut, “Computation of channel capacity and rate distortion functions,” *IEEE Trans. Inform. Theory*, vol. 18, pp. 460-473, 1972.

Arimoto-Blahut Algorithm

Let $| X |= M$, $| Y |= N$, and let $F = [f_0, f_1, \cdots, f_{M-1}]$. Let $\in$ be some small positive number. Let *j* and *k* be indices having ranges $j \in [0,1,\cdots,M-1], k \in [0,1,\cdots,N-1]$.

Initialize P_X with element values $P_j = 1/M$, and initialize $Q_Y = P_{Y|X} \cdot P_X$.

Repeat until stopping point is reached：

$$f_j = \exp\{\sum_k \left[P_{k|j} \log\left(\frac{P_{k|j}}{q_k} \right) \right] \text{ for } j \in [0,1,\cdots,M-1]$$

$$g = F \cdot P_j$$

$$I_L = \log_2(g)$$

$$I_U = \log_2(\max(f_j))$$

```
If $I_U - I_L < \in$ then
    $c_X = I_L$
stop
else
    $P_j = f_j P_j / g$ for $j = 0,1,\cdots,M-1$
    $Q_Y = P_{Y|X} \cdot P_X$
End if
End Repeat
```

Upon termination, this algorithm provides an estimate of the channel capacity accurate to within the stopping factor and the input probability distribution that achieves this capacity.

The channel capacity proves to be a sensitive function of the transition probability matrix $P_{Y|X}$ but, in many cases, a fairly weak function of P_X.

Example 2.1

Using the Arimoto-Blahut algorithm, find the channel capacity, the input and output probability distributions that achieve the channel capacity, and the mutual information given a uniform P_X for channels with the following transition probability matrices：

$$\text{a)}\ P_{Y|X} = \begin{pmatrix} .98 & .05 \\ .02 & .95 \end{pmatrix} \quad \text{b)}\ P_{Y|X} = \begin{pmatrix} .8 & .05 \\ .2 & .95 \end{pmatrix} \quad \text{c)}\ P_{Y|X} = \begin{pmatrix} .8 & .1 \\ .2 & .9 \end{pmatrix}$$

$$\text{d)}\ P_{Y|X} = \begin{pmatrix} .6 & .01 \\ .4 & .99 \end{pmatrix} \quad \text{e)}\ P_{Y|X} = \begin{pmatrix} .8 & .3 \\ .2 & .7 \end{pmatrix} \quad \text{f)}\ P_{Y|X} = \begin{pmatrix} .8 & .05 \\ .15 & .15 \\ .05 & .8 \end{pmatrix}$$

Solutions： Using the algorithm given above with $\in = 10^{-7}$,the following computer solutions are obtained：

a) $C_x = .78585, \;\; P_X = [.51289 \quad .48711]^T, \;\; Q_Y = [.52698 \quad .47302]^T, \;\; I(X;Y) = .78543$

b) $C_x = .48130, \;\; P_X = [.46761 \quad .53239]^T, \;\; Q_Y = [.4007 \quad .5993]^T, \;\; I(X;Y) = .47955$

c) $C_x = .39775, \;\; P_X = [.4824 \quad .5176]^T, \;\; Q_Y = [.4377 \quad .5623]^T, \;\; I(X;Y) = .39731$

d) $C_x = .36877, \;\; P_X = [.4238 \quad .5762]^T, \;\; Q_Y = [.26 \quad .74]^T, \;\; I(X;Y) = .36145$

e) $C_x = .191238, \;\; P_X = [.510 \quad .490]^T, \;\; Q_Y = [.555 \quad .445]^T, \;\; I(X;Y) = .191165$

f) $C_x = I(C;Y) = .57566, \;\; P_X = [.5 \quad .5]^T, \;\; Q_Y = [.425 \quad .15 \quad .425]^T$

Example 2.1 illustrates a wide range of channel capacities in cases (a) through (e). Comparison of C_x with the accompanying value of $I(X;Y)$ demonstrates, however, that the maximum mutual information is actually quite close to the uniform-P_X case. The cases considered in this example therefore show relatively small sensitivity to P_X, since the percent change in P_X from the uniform distribution to that which maximizes mutual information is less than the percent change in mutual information that it produces. On the other hand, values of C_x vary significantly from one case to another. This indicates the strong effect has $P_{Y|X}$ on determining C_x.

Case (f) is interesting since, for this transition probability matrix, the uniform input distribution produces the maximum mutual information. This case is an example of a *symmetric channel*. Note that the columns of its transition probability matrix are permutations of each other. Likewise, the top and bottom rows of $P_{Y|X}$ are permutations of each other. The center row, which is *not* a permutation of the other rows, corresponds to output symbol y_1, which, as we saw in example 1.4, makes no contribution to the mutual information.

2.2 Symmetric Channels

Symmetric channels play an important role in communication systems and many such systems attempt, by design, to achieve a symmetric channel function.

The reason for the importance of the symmetric channel is that when such a channel is possible, it frequently has greater channel capacity than an otherwise equivalent nonsymmetric channel would have. The effect of perturbing the channel of Example 2.1(f) to a non-symmetric case is illustrated in the following example.

Example 2.2

Repeat Example 2.1 for a channel with transition probability matrix

$$P_{Y|X} = \begin{pmatrix} .79 & .05 \\ .16 & .15 \\ .05 & .8 \end{pmatrix}$$

Solution： Using the channel capacity algorithm, we get

$$C_x = .571215, P_X = [.50095 \quad .49905]^T, Q_Y = [.4207 \quad .1550 \quad .4243]^T,$$
$$I(X;Y) = .5712$$

The channel capacity for this example is roughly 1% less than the symmetric case, although the actual changes made to the transition probability matrix were very slight.

Example 2.3

Quadrature phase-shift keying (QPSK) is a modulation method that produces a symmetric channel. For QPSK, $|X|=|Y|=4.$ Using the Arimoto-Blahut algorithm, determine the channel capacity and the input symbol probability distribution that achieves this capacity if the transition probability matrix is

$$P_{Y|X} = \begin{pmatrix} .95 & .024 & .024 & .002 \\ .024 & .95 & .002 & .024 \\ .024 & .002 & .95 & .024 \\ .002 & .024 & .024 & .95 \end{pmatrix}$$

Solution $:C_x = 1.653488, \quad P_X = [.25 \quad .25 \quad .25 \quad .25]^T.$

Because symmetric channels play a very important role in practical communication system, it is worthwhile to examine some of their special properties. We begin with the channel in Example 2.3. Notice that the capacity for this channel is achieved when P_X is uniformly distributed. This is always the case for a symmetric channel as we shall now see.

First, notice that every column of $P_{Y|X}$ is a permutation of the first column. Also notice that every row of $P_{Y|X}$ is a permutation of the first row. When $P_{Y|X}$ is a square matrix, this permutation property of the columns and the rows is sufficient condition for a uniformly distributed input alphabet to achieve the maximum mutual information. Indeed, the permutation condition is what gives rise to the term “symmetric channel’.

A symmetric channel of considerable importance, both theoretically and practically, is the binary symmetric channel (BSC) for which

$$P_{Y|X} = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \qquad (2.1)$$

The parameter p is known as the crossover probability, and is equal to the probability that the demodulator/detector will make a hard-decision decoding error. The BSC is the model for essentially all binary-pulse transmission systems of practical importance.

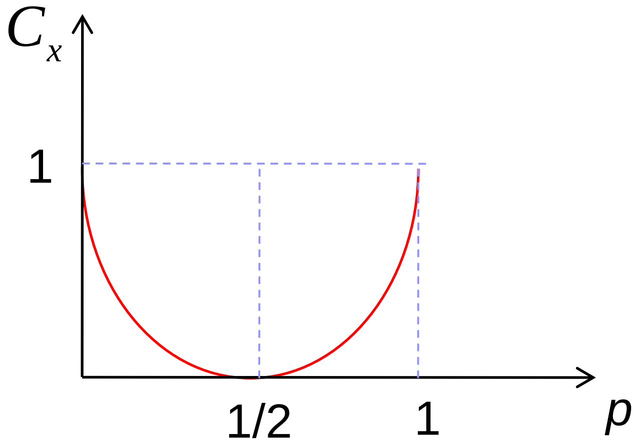
$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y \mid X) \\
&= H(Y) - \sum p(x) H(Y \mid X = x) \\
&= H(Y) - [(1-p)\log(1-p) + p \log p] \\
&= H(Y) - H(p) \\
&\le 1 - H(p) \qquad \text{Since Y is a binary random variable}
\end{aligned}
$$

$$
\begin{aligned}
C_x &= \max_{p(x)}(I(X;Y)) \\
&= (1-p)\log_2(2(1-p)) + p\log_2(2p) \\
&= 1 + (1-p)\log_2(1-p) + p\log_2(p) \qquad (2.2)
\end{aligned}
$$

Because the BSC occurs frequently in theory and in practice, Equation (2.2) is often abbreviated as

$$C_x = 1 - H(p) \qquad (2.3)$$

where the notation $H(p)$ arises from the fact that the terms involving p in Equation (2.2).



Examination of Equation (2.2) reveals some interesting properties of the capacity of the BSC. We observe that the capacity is bounded by the range $0 \le C_x \le 1$. The upper bound is achieved only if $p = 0$ or $p = 1$. The $p = 0$ case is not surprising, since this corresponds to a channel which does not make errors (known as a “noiseless channel”). The $p = 1$ case is initially starting because this corresponds to a channel which always makes errors. However, if we know that the channel output is always wrong, we can easily set things right by decoding the opposite of what the channel output symbol is.

The zero capacity case occurs when $P = 0.5$. In this case, the channel output symbol is as likely to be correct as it is to be incorrect. Under this condition, the information loss in the channel is total. The capacity equation for the BSC is a concave-upward function possessing a single minimum at $P = 0.5$.

We note from Equation (2.3) that, except for the $p = 0$ and $p = 1$ cases, the capacity of the BSC is always less than the source entropy. If we try to send information through the channel using the maximum amount of information per symbol, some of this information will be lost, and decoding errors at the receiver will result. However, we will later see that, if we add sufficient redundancy to the transmitted data stream, it is possible to reduce the probability of lost information to an arbitrarily low level.

III. Block Coding and Shannon’s Second Theorem

3.1 Equivocation

In the previous section, we saw that there is a maximum amount of information per channel use that can be supported by a channel. Any attempt to exceed this channel capacity will result in information being lost during transmission. That is,

$$I(X;Y) = H(X) - H(X|Y)$$

and so

$$C_x = \max_{P_X}(H(X) - H(X|Y)) \tag{3.1}$$

The conditional entropy $H(X\,|\,Y)$ corresponds to our uncertainty about what the input to the channel was, given our observation of the channel output. It is a measure of the information lost during the transmission. This conditional entropy is often called the equivocation for this reason. From its definition, it is clear that equivocation is a non-negative function. What is less clear, but nonetheless true, is

$$H(X\,|\,Y) \le H(X) \qquad (3.2)$$

Example 3.1 Derive Equation (3.2)

Solution：$H(X)-H(X\,|\,Y)=\sum_{x\in X}\sum_{y\in Y}p_{x,y}\log_2\left(\frac{p_{x,y}}{p_x q_y}\right)$

Since $\ln(1/x)\geq 1-x$ for $x>0$, we can re-write this as

$$H(X)-H(X\,|\,Y)\geq\sum_{x\in X}\sum_{y\in Y}\frac{p_{x,y}}{\ln(2)}\left(1-\frac{p_x q_y}{p_{x,y}}\right)$$

Since $\sum_{x\in X}\sum_{y\in Y}p_{x,y}=1,\quad \sum_{x\in X}\sum_{y\in Y}p_x q_y=\sum_{x\in X}p_x\sum_{y\in Y}q_y=1,$

we have $H(X)-H(X\,|\,Y)\geq 0$

The equivocation is zero if and only if the transition probabilities $p_{y|x}$ are either zero or one for all pairs $(y\in Y, x\in X)$

3.2 Entropy Rate and the Channel-Coding Theorem

The entropy of a block of *n* symbols：

$$H(X_0, X_1, \cdots, X_{n-1}) \leq nH(X)$$

with equality if and only if *X* is a memoryless source. In transmitting a block of *n* symbols, we use the channel *n* times. Recall that channel capacity has units of bits per channel use, and refers to an average amount of information per channel use. Since $H(X_0, X_1, \cdots, X_{n-1})$ is the average information contained in the *n*-symbol block, it follows that the average information per channel use would be

$$R \triangleq \frac{H(X_0, X_1, \cdots, X_{n-1})}{n} \leq H(X) \qquad (3.3)$$

As it happens, *R* in Equation (3.3) is an unbiased estimate of the average bits per channel use, rather than exactly the expected number of bits per channel use. Formally, the true average bits per channel use is achieved in the limit

$$R = \lim_{n \to \infty} \frac{H(X_0, X_1, \cdots, X_{n-1})}{n} \text{£} H(X) \qquad (3.4)$$

R is therefore called the entropy rate.

Now, $R \leq H(X)$,with equality if and only if all the symbols x_t are statistically independent. Suppose they are not; suppose, in our transmission of the block, we deliberately introduce redundant symbols. Then $R < H(X)$. Taking this further, suppose we introduce a sufficient number of redundant symbols in the block so that $R \leq C_x$.Does this mean that transmission without information loss, i.e., with zero equivocation, is possible？

Remarkably enough, the answer to this question is “yes”! What is the implication of doing so？The implication is：If we signal without information loss, then it is possible to send information through the noisy channel with an arbitrarily low probability of error (even though individual symbols may be decoded incorrectly!). The process of adding redundancy to a block of transmitted symbols is called channel coding. Does there exist a channel code that will accomplish this purpose？The answer to this question is given by Shannon’s second theorem：

Theorem： Suppose $R < C_x$,where C_x is the capacity of a memoryless channel. Then for any $\in > 0$,there exists a block length n and a code of block length n and rate R whose probability of block decoding error p_e satisfies $p_e \le \in$ when the code is used on this channel.

Shannon’s second theorem (also called Shannon’s main theorem) tells us that it is possible to transmit information over a noisy channel with arbitrarily small probability of error. This is such a remarkable statement that it deserves close examination.

The theorem says that if the entropy rate R in a block of n symbols is smaller than the channel capacity, then we can make the probability of error arbitrarily small. What error are we talking about？

Now suppose we send a block of n bits in which $k < n$ of these bits are statistically independent “information” bits and $R = n - k$ are redundant “parity” bits computed from the k information bits according to some “coding rule”. The entropy of the block will then be k bits and the average information in bits per channel use will be $R = k/n$. If this entropy rate is less than the channel capacity, Shannon’s theorem says we can make the probability of error in recovering our original k information bits arbitrarily small. The channel will make errors within our block of n bits, but the redundancy built into the block will be sufficient to correct these errors and recover the k bits of information we transmitted.

Shannon’s theorem does not say that we can do this for just any block length *n* we might care to choose. The theorem says there exists a block length *n* for which there is a code of rate *R*. the required size of the block length *n* depends on the upper bound $\in$ we pick for our error probability.

Actually, Shannon’s theorem implies very strongly that the block length *n* is going to be very large if *R* is to approach C_x to within an arbitrarily small distance with an arbitrarily small probability of error.

The complexity and expense of an error-correcting channel code are believed to grow rapidly as R approaches the channel capacity and the probability of a block decoding error is made arbitrarily small. It is believed by many that beyond a particular rate called the cutoff rate, R_0, it is prohibitively expensive to use the channel. In the case of the binary symmetric channel, this rate is given by

$$R_0 = -\log_2\left(0.5 + \sqrt{p(1-p)}\right)$$

The belief that R_0 is some kind of "sound barrier" for practical error correcting codes stems from the fact that for certain kinds of decoding methods the complexity of the decoder grows extremely rapidly as R exceeds R_0.