6

Mutual information and channel capacity

6.1 Introduction

We go on to take a closer look at a typical problem in communications: how to send information reliably over a noisy communication channel. A communication channel can be thought of as the medium through which the message signal propagates from the transmit end (the source) to the receive end (the destination). A channel is said to be noisy if the data read at the channel output is not necessarily the same as the input (due to, e.g., perturbation caused by the ambient noise). Consider for example that Alice writes down a "7" on a paper with a small font size, and uses a fax machine to transfer this page to Bob. Due to limited resolution of electronic cable data transfer, Bob sees a "distorted 7" on the faxed page and could decode it incorrectly as, say, "9" (see Figure 6.1).

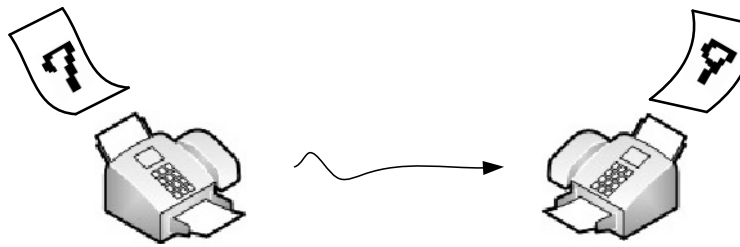

Figure 6.1 Cable data transfer as a channel.

In this example the route of data transfer through the cable acts as the channel, which is noisy since it distorts the input alphabet and, in turn, leads to possibly incorrect message decoding at the destination. It is still probable that Bob reads the message correctly as "7." The higher the probability of correct message decoding is, the more reliable the communication will be.

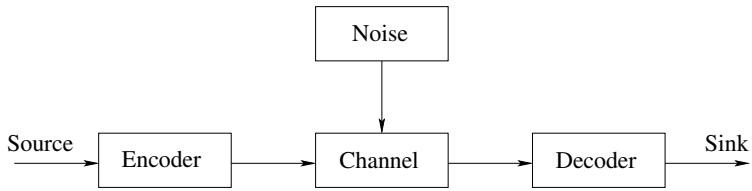

Figure 6.2 Signaling model.

The general block diagram of a typical communication system is depicted in Figure 6.2.

For a probabilistic message source, we are now able to quantify the amount of its information content in terms of the entropy defined in Chapter 5. We implicitly assume that the message has been compressed in order to remove the inherent redundancy, if any; this can be done via data compression as introduced in Chapter 4 (see also the discussion in Appendix 7.7). To combat the detrimental effect induced by the channel, the source message is further encoded with certain channel coding schemes, like the Hamming code introduced in Chapter 3. The encoded data stream is then sent over the channel. Message decoding is performed at the receiver based on the channel output. We examine each of the following problems.

- How should we measure the amount of information that can get through the channel, and what is the maximal amount?
- How can we use the channel to convey information reliably?

Note that if the channel is noiseless, i.e. the input is always reproduced at the output without errors, the answers to the aforementioned problems are simple: the maximal amount of information that can be conveyed over the channel equals the source entropy, and this can be done without any data protection mechanisms such as channel coding. If the channel is noisy, the answers turn out to be rather nontrivial. Let us begin the discussions with the mathematical model of a noisy communication channel.

6.2 The channel

Recall that in the example depicted in Figure 6.1, the input letter "7" can be either correctly decoded or mis-recognized as some other letter. The uncertainty in source symbol recovery naturally suggests a probabilistic characterization of the input–output relation of a noisy channel; such a mathematical channel

model is needed in order to pin down various intrinsic properties of a channel, e.g. how much information can go through a channel.

Below is the formal definition for a channel.

Definition 6.1 (Channel) A *channel* $(\mathcal{X}, P_{Y|X}(y_j|x_i), \mathcal{Y})$ is given by

(1) an input alphabet $\mathcal{X} \triangleq \{x_1, \ldots, x_s\}$, where s denotes the number of input letters;
(2) an output alphabet $\mathcal{Y} \triangleq \{y_1, \ldots, y_t\}$, where t denotes the number of output letters; and
(3) a conditional probability distribution $P_{Y|X}(y_j|x_i)$, which specifies the probability of observing $Y = y_j$ at the output given that $X = x_i$ is sent, $1 \le i \le s$, $1 \le j \le t$.

Hence a channel with input $X \in \mathcal{X}$ and output $Y \in \mathcal{Y}$ is entirely specified by a set of conditional probabilities $P_{Y|X}(y_j|x_i)$. The size of the input and output alphabets, namely s and t, need not be the same. A schematic description of the channel is shown Figure 6.3.

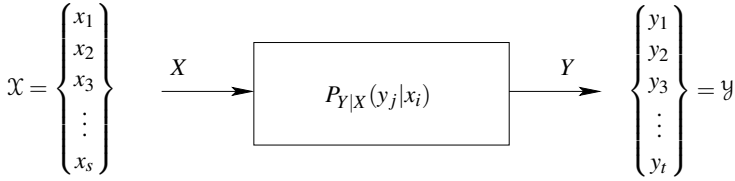

Figure 6.3 Channel model.

In this model the channel is completely described by the matrix of conditional probabilities, the so-called *channel transition matrix*:

$$\begin{pmatrix} P_{Y|X}(y_1|x_1) & P_{Y|X}(y_2|x_1) & \ldots & P_{Y|X}(y_t|x_1) \\ P_{Y|X}(y_1|x_2) & P_{Y|X}(y_2|x_2) & \ldots & P_{Y|X}(y_t|x_2) \\ \vdots & \vdots & \ddots & \vdots \\ P_{Y|X}(y_1|x_s) & P_{Y|X}(y_2|x_s) & \ldots & P_{Y|X}(y_t|x_s) \end{pmatrix}. \tag{6.1}$$

The channel transition matrix has the following properties.

(1) The entries on the ith row consist of the probabilities of observing output letters $y_1, \ldots, y_t$ given that the ith input symbol x_i is sent.
(2) The entries on the jth column consist of the probabilities of observing the jth output letter y_j given, respectively, the ith input symbols x_i are sent, $i = 1, \ldots, s$.

(3) The sum of the entries in a row is always 1, i.e.

$$\sum_{j=1}^{t} P_{Y|X}(y_j|x_i) = 1. \tag{6.2}$$

This merely means that for each input x_i we are certain that something will come out, and that the $P_{Y|X}(y_j|x_i)$ give the distribution of these probabilities.

(4) If $P_X(x_i)$ is the probability of the input symbol x_i, then

$$\sum_{i=1}^{s}\sum_{j=1}^{t} P_{Y|X}(y_j|x_i)P_X(x_i) = 1, \tag{6.3}$$

meaning that when something is put into the system, then certainly something comes out.

The probabilities $P_{Y|X}(y_j|x_i)$, $1 \le i \le s$, $1 \le j \le t$, characterize the channel completely. We assume that the channel is *stationary*, i.e. the probabilities do not change with time. We note that X is not a source but is an information-carrying channel input, which is typically a stream of encoded data (see Figure 6.2; see Chapters 3 and 7 for more details).

6.3 The channel relationships

At the transmit end we have s possible input symbols $\{x_1, \ldots, x_s\}$. If the ith symbol x_i is selected and sent over the channel, the probability of observing the jth channel output letter y_j is given by the conditional probability $P_{Y|X}(y_j|x_i)$. This means that the probability that the input–output pair (x_i, y_j) simultaneously occurs, i.e. the joint probability of $X = x_i$ and $Y = y_j$, is given by

$$P_{X,Y}(x_i, y_j) \triangleq P_{Y|X}(y_j|x_i)P_X(x_i). \tag{6.4}$$

Let us go one step further by asking the question of how to determine the probability that the jth letter y_j will occur at the channel output, hereafter denoted by $P_Y(y_j)$. A simple argument, taking into account that each input symbol occurs with probability $P_X(x_i)$, yields

$$P_Y(y_j) = P_{Y|X}(y_j|x_1)P_X(x_1) + \cdots + P_{Y|X}(y_j|x_s)P_X(x_s) \tag{6.5}$$

$$= \sum_{i=1}^{s} P_{Y|X}(y_j|x_i)P_X(x_i), \qquad 1 \le j \le t. \tag{6.6}$$

The above "channel equation" characterizes the input–output relation of a channel. Note that in terms of the joint probability $P_{X,Y}(x_i, y_j)$ in (6.4), we can

rewrite (6.6) in a more compact form:

$$P_Y(y_j) = \sum_{i=1}^{s} P_{X,Y}(x_i, y_j), \qquad 1 \le j \le t. \tag{6.7}$$

Now take a further look at (6.4), which relates the probability of a joint occurrence of the symbol pair (x_i, y_j) with the input distribution via the *forward conditional probability* $P_{Y|X}(y_j|x_i)$ (starting from the input front with x_i given and expressing the probability that y_j is the resultant output). We can alternatively write $P_{X,Y}(x_i, y_j)$ as

$$P_{X,Y}(x_i, y_j) = P_{X|Y}(x_i|y_j)P_Y(y_j), \tag{6.8}$$

which evaluates the joint probability $P_{X,Y}(x_i, y_j)$ based on the output distribution and the *backward conditional probability* $P_{X|Y}(x_i|y_j)$ (given that y_j is received, the probability that x_i is sent). Equating (6.4) with (6.8) yields

$$P_{X|Y}(x_i|y_j) = \frac{P_{Y|X}(y_j|x_i)P_X(x_i)}{P_Y(y_j)}, \tag{6.9}$$

which is the well known *Bayes' Theorem* on conditional probabilities [BT02].

In the Bayes' formula (6.9) we can write $P_Y(y_j)$ in the denominator in terms of (6.6) to get the equivalent expression

$$P_{X|Y}(x_i|y_j) = \frac{P_{Y|X}(y_j|x_i)P_X(x_i)}{\sum_{i'=1}^{s} P_{Y|X}(y_j|x_{i'})P_X(x_{i'})}. \tag{6.10}$$

Summing (6.10) over all the x_i clearly gives

$$\sum_{i=1}^{s} P_{X|Y}(x_i|y_j) = \sum_{i=1}^{s} \frac{P_{Y|X}(y_j|x_i)P_X(x_i)}{\sum_{i'=1}^{s} P_{Y|X}(y_j|x_{i'})P_X(x_{i'})} \tag{6.11}$$

$$= \frac{\sum_{i=1}^{s} P_{Y|X}(y_j|x_i)P_X(x_i)}{\sum_{i'=1}^{s} P_{Y|X}(y_j|x_{i'})P_X(x_{i'})} \tag{6.12}$$

$$= 1, \tag{6.13}$$

which means that, given output y_j, some x_i was certainly put into the channel.

6.4 The binary symmetric channel

A simple special case of a channel is the *binary channel*, which has two input symbols, 0 and 1, and two output symbols, 0 and 1; a schematic description is depicted in Figure 6.4.

The binary channel is said to be *symmetric* if

$$P_{Y|X}(0|0) = P_{Y|X}(1|1), \qquad P_{Y|X}(0|1) = P_{Y|X}(1|0). \tag{6.14}$$

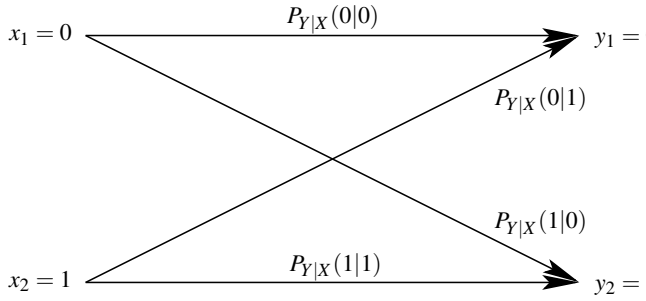

Figure 6.4 The binary channel.

Usually we abbreviate *binary symmetric channel* to *BSC*.

Let the probabilities of the input symbols be

$$P_X(0) = \delta, \tag{6.15}$$
$$P_X(1) = 1 - \delta, \tag{6.16}$$

and let the BSC probabilities be

$$P_{Y|X}(0|0) = P_{Y|X}(1|1) = 1 - \varepsilon, \tag{6.17}$$
$$P_{Y|X}(1|0) = P_{Y|X}(0|1) = \varepsilon. \tag{6.18}$$

The channel matrix is therefore

$$\begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix} \tag{6.19}$$

and the channel relationships (6.6) become

$$P_Y(0) = (1-\varepsilon)\delta + \varepsilon(1-\delta), \tag{6.20}$$
$$P_Y(1) = \varepsilon\delta + (1-\varepsilon)(1-\delta). \tag{6.21}$$

Note that these equations can be simply checked by computing their sum:

$$P_Y(0) + P_Y(1) = (1-\varepsilon+\varepsilon)\delta + (1-\varepsilon+\varepsilon)(1-\delta) = \delta + 1 - \delta = 1. \tag{6.22}$$

Given that we know what symbol we received, what are the probabilities for the various symbols that might have been sent?

We first compute the two denominators in Equation (6.10):

$$\sum_{i=1}^{2} P_{Y|X}(y_1|x_i)P_X(x_i) = (1-\varepsilon)\delta + \varepsilon(1-\delta), \tag{6.23}$$

$$\sum_{i=1}^{2} P_{Y|X}(y_2|x_i)P_X(x_i) = \varepsilon\delta + (1-\varepsilon)(1-\delta), \tag{6.24}$$

which of course are the same as (6.20) and (6.21). We then have

$$P_{X|Y}(0|0) = \frac{(1-\varepsilon)\delta}{(1-\varepsilon)\delta + \varepsilon(1-\delta)}, \tag{6.25}$$

$$P_{X|Y}(1|0) = \frac{\varepsilon(1-\delta)}{(1-\varepsilon)\delta + \varepsilon(1-\delta)}, \tag{6.26}$$

$$P_{X|Y}(0|1) = \frac{\varepsilon\delta}{\varepsilon\delta + (1-\varepsilon)(1-\delta)}, \tag{6.27}$$

$$P_{X|Y}(1|1) = \frac{(1-\varepsilon)(1-\delta)}{\varepsilon\delta + (1-\varepsilon)(1-\delta)}. \tag{6.28}$$

Note that this involves the choice of the probabilities of the channel input.

In the special case of equally likely input symbols ($\delta = 1/2$) we have the very simple equations

$$P_{X|Y}(0|0) = P_{X|Y}(1|1) = 1-\varepsilon, \tag{6.29}$$

$$P_{X|Y}(1|0) = P_{X|Y}(0|1) = \varepsilon. \tag{6.30}$$

As a more peculiar example, suppose that $1-\varepsilon = 9/10$ and $\varepsilon = 1/10$ for the BSC, but suppose also that the probability of the input $x = 0$ being sent is $\delta = 19/20$ and $x = 1$ being sent is $1-\delta = 1/20$. We then have

$$P_{X|Y}(0|0) = \frac{171}{172}, \tag{6.31}$$

$$P_{X|Y}(1|0) = \frac{1}{172}, \tag{6.32}$$

$$P_{X|Y}(0|1) = \frac{19}{28}, \tag{6.33}$$

$$P_{X|Y}(1|1) = \frac{9}{28}. \tag{6.34}$$

Thus if we receive $y = 0$, it is more likely that $x = 0$ was sent because $171/172$ is much larger than $1/172$. If, however, $y = 1$ is received, we still have $19/28 > 9/28$, and hence $x = 0$ has a higher probability of being the one that has been sent. Therefore, $x = 0$ is always claimed regardless of the symbol received. As a result, if a stream of n binary bits is generated according to the probability law $P_X(0) = \delta = 19/20$, then there are about $n(1-\delta) = n/20$ 1s in the input sequence that will be decoded incorrectly as 0s at the channel output. Hence,

the above transmission scheme does not use the channel properly, since irrespective of n it will incur an average decoding error of about $1/20$, significantly away from zero.

This situation arises whenever both following conditions are valid:

$$P_{X|Y}(0|0) > P_{X|Y}(1|0), \tag{6.35}$$

$$P_{X|Y}(0|1) > P_{X|Y}(1|1). \tag{6.36}$$

From (6.25)–(6.28) we see that for the binary symmetric channel these conditions are

$$(1-\varepsilon)\delta > \varepsilon(1-\delta), \tag{6.37}$$

$$\varepsilon\delta > (1-\varepsilon)(1-\delta), \tag{6.38}$$

or equivalently

$$\delta > \varepsilon, \tag{6.39}$$

$$\delta > 1-\varepsilon; \tag{6.40}$$

i.e., the bias in the choice of input symbols is greater than the bias of the channel. This discussion shows that it is possible, for given values of the channel transition probabilities, to come up with values for the channel input probabilities that do not make much sense in practice. As will be shown below, we can improve this if we can learn more about the fundamental characteristics of the channel and then use the channel properly through a better assignment of the input distribution.[1] To this end, we leverage the entropy defined in Chapter 5 to define the notion of "capability of a channel for conveying information" in a precise fashion.

6.5 System entropies

We can regard the action of a channel as "transferring" the probabilistic information-carrying message X into the output Y by following the conditional probability law $P_{Y|X}(y_j|x_i)$. Both the input and output ends of the channel are thus uncertain in nature: we know neither exactly which input symbol will be selected nor which output letter will be certainly seen at the output (rather, only probabilistic characterizations of various input–output events in terms of $P_X(\cdot)$ and $P_{Y|X}(\cdot|\cdot)$ are available). One immediate question to ask is: how much

[1] Note that, whereas the source is assumed to be given to us and therefore cannot be modified, we can freely choose the channel input probabilities by properly designing the channel encoder (see Figure 6.2).

"aggregate" information, or amount of uncertainty, is contained in the overall channel system? From Chapter 5, we know that the average amount of uncertainty of the input is quantified by the entropy as

$$H(X) = \sum_{i=1}^{s} P_X(x_i) \log_2 \left(\frac{1}{P_X(x_i)} \right). \tag{6.41}$$

We have shown that $H(X) \geq 0$, and $H(X) = 0$ if the input is certain; also, $H(X)$ is maximized when all x_i are equally likely.[2] We can also likewise define the entropy of the output as

$$H(Y) = \sum_{j=1}^{t} P_Y(y_j) \log_2 \left(\frac{1}{P_Y(y_j)} \right), \tag{6.42}$$

which, as expected, measures the uncertainty of the channel output. If we look at both the input and the output, the probability of the event that $X = x_i$ and $Y = y_j$ simultaneously occur is given by the joint probability $P_{X,Y}(x_i, y_j)$ (see (6.4)). Analogous to the entropy of X (or Y), we have the following definition of the entropy when both X and Y are simultaneously taken into account.

Definition 6.2 The *joint entropy* of X and Y, defined as

$$H(X,Y) \triangleq \sum_{i=1}^{s} \sum_{j=1}^{t} P_{X,Y}(x_i, y_j) \log_2 \left(\frac{1}{P_{X,Y}(x_i, y_j)} \right), \tag{6.43}$$

measures the total amount of uncertainty contained in the channel input and output, hence the overall channel system.

One might immediately ask about the relation between $H(X,Y)$ and the individual entropies, in particular whether $H(X,Y)$ just equals the sum of $H(X)$ and $H(Y)$. This is in general not true, unless X and Y are *statistically independent*, meaning that what comes out does not depend on what goes in. More precisely, independence among X and Y is characterized by [BT02]

$$P_{X,Y}(x_i, y_j) = P_X(x_i) P_Y(y_j). \tag{6.44}$$

Based on (6.43) and (6.44), we have the following proposition.

Proposition 6.3 *If X and Y are statistically independent, then*

$$H(X,Y) = H(X) + H(Y). \tag{6.45}$$

[2] As noted in Lemma 5.11, $H(X) \leq \log_2 s$ bits, with equality if $P_X(x_i) = 1/s$ for all i.

Proof With (6.44), we have

$$H(X,Y) = \sum_{i=1}^{s}\sum_{j=1}^{t} P_X(x_i)P_Y(y_j)\log_2\left(\frac{1}{P_X(x_i)P_Y(y_j)}\right) \tag{6.46}$$

$$= \sum_{i=1}^{s}\sum_{j=1}^{t} P_X(x_i)P_Y(y_j)\left(\log_2\left(\frac{1}{P_X(x_i)}\right)+\log_2\left(\frac{1}{P_Y(y_j)}\right)\right) \tag{6.47}$$

$$= \sum_{i=1}^{s}\sum_{j=1}^{t} P_X(x_i)P_Y(y_j)\log_2\left(\frac{1}{P_X(x_i)}\right) + \sum_{i=1}^{s}\sum_{j=1}^{t} P_X(x_i)P_Y(y_j)\log_2\left(\frac{1}{P_Y(y_j)}\right) \tag{6.48}$$

$$= \sum_{j=1}^{t} P_Y(y_j)\underbrace{\sum_{i=1}^{s} P_X(x_i)\log_2\left(\frac{1}{P_X(x_i)}\right)}_{H(X)} + \sum_{i=1}^{s} P_X(x_i)\underbrace{\sum_{j=1}^{t} P_Y(y_j)\log_2\left(\frac{1}{P_Y(y_j)}\right)}_{H(Y)} \tag{6.49}$$

$$= H(X)+H(Y), \tag{6.50}$$

where the last equality follows since

$$\sum_{j=1}^{t} P_Y(y_j) = \sum_{i=1}^{s} P_X(x_i) = 1. \tag{6.51}$$

□

In Proposition 6.4 we derive the relation that links the joint entropy $H(X,Y)$ with the individual $H(X)$ (or $H(Y)$) when X and Y are dependent, which is typically true since the channel output depends at least partly on the channel input (otherwise no information can be conveyed through the channel).

Proposition 6.4 (Chain rule) *The following result holds:*

$$H(X,Y) = H(X)+H(Y|X), \tag{6.52}$$

where

$$H(Y|X) \triangleq \sum_{i=1}^{s}\sum_{j=1}^{t} P_{X,Y}(x_i,y_j)\log_2\left(\frac{1}{P_{Y|X}(y_j|x_i)}\right) \tag{6.53}$$

is the conditional entropy *associated with Y given X.*

Proof By means of the relation $P_{X,Y}(x_i,y_j) = P_{Y|X}(y_j|x_i)P_X(x_i)$ (see (6.4)) it

follows that

$$H(X,Y) = \sum_{i=1}^{s}\sum_{j=1}^{t} P_{X,Y}(x_i,y_j)\log_2\left(\frac{1}{P_{Y|X}(y_j|x_i)P_X(x_i)}\right) \tag{6.54}$$

$$= \sum_{i=1}^{s}\sum_{j=1}^{t} P_{X,Y}(x_i,y_j)\log_2\left(\frac{1}{P_X(x_i)}\right) + \sum_{i=1}^{s}\sum_{j=1}^{t} P_{X,Y}(x_i,y_j)\log_2\left(\frac{1}{P_{Y|X}(y_j|x_i)}\right) \tag{6.55}$$

$$= \sum_{i=1}^{s}\log_2\left(\frac{1}{P_X(x_i)}\right)\underbrace{\sum_{j=1}^{t} P_{X,Y}(x_i,y_j)}_{P_X(x_i)} + \sum_{i=1}^{s}\sum_{j=1}^{t} P_{X,Y}(x_i,y_j)\log_2\left(\frac{1}{P_{Y|X}(y_j|x_i)}\right) \tag{6.56}$$

$$= \sum_{i=1}^{s} P_X(x_i)\log_2\left(\frac{1}{P_X(x_i)}\right) + \sum_{i=1}^{s}\sum_{j=1}^{t} P_{X,Y}(x_i,y_j)\log_2\left(\frac{1}{P_{Y|X}(y_j|x_i)}\right) \tag{6.57}$$

$$= H(X) + H(Y|X). \tag{6.58}$$

□

The joint entropy $H(X,Y)$ is thus the sum of the input entropy $H(X)$ and the conditional entropy $H(Y|X)$, which measures the uncertainty remaining in Y, given that X is known. Note that if X and Y are independent, i.e. one can infer nothing about Y even if X is already known, we have $H(Y|X) = H(Y)$ and Proposition 6.4 reduces to Proposition 6.3.

Another interpretation of $H(Y|X)$ is that it represents how much must be added to the input entropy to obtain the joint entropy; in this regard, $H(Y|X)$ is called the *equivocation* of the channel. We can again use (6.4) to rewrite $H(Y|X)$ as

$$H(Y|X) = \sum_{i=1}^{s}\sum_{j=1}^{t} P_{Y|X}(y_j|x_i)P_X(x_i)\log_2\left(\frac{1}{P_{Y|X}(y_j|x_i)}\right) \tag{6.59}$$

$$= \sum_{i=1}^{s} P_X(x_i)H(Y|x_i), \tag{6.60}$$

where

$$H(Y|x_i) \triangleq \sum_{j=1}^{t} P_{Y|X}(y_j|x_i)\log_2\left(\frac{1}{P_{Y|X}(y_j|x_i)}\right) \tag{6.61}$$

is the conditional entropy of Y given a particular $X = x_i$. Finally, we remark that, starting from the alternative expression for $P_{X,Y}(x_i, y_j)$ given in (6.8), $H(X,Y)$ can be accordingly expressed as

$$H(X,Y) = H(Y) + H(X|Y). \tag{6.62}$$

Exercise 6.5 *Let (X,Y) have the joint distribution given in Table 6.1. Compute $H(X)$, $H(X,Y)$, and $H(X|Y)$.* ◊

Table 6.1 *A joint distribution of (X,Y)*

		X			
		1	2	3	4
Y	1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
	2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
	3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
	4	$\frac{1}{4}$	0	0	0

Exercise 6.6 *Verify that $H(Y|X) = H(Y)$ if X and Y are independent.* ◊

6.6 Mutual information

Consider again the transmission system shown in Figure 6.3. We wish to determine how much information about the input can be gained based on some particular received output letter $Y = y_j$; this is the first step toward quantifying the amount of information that can get through the channel.

At the transmit side, the probability that the ith input symbol x_i occurs is $P_X(x_i)$, which is called the *a priori*[3] *probability* of x_i. Upon receiving $Y = y_j$, one can try to infer which symbol probably has been sent based on the information carried by y_j. In particular, given y_j is received, the probability that x_i has been sent is given by the backward conditional probability $P_{X|Y}(x_i|y_j)$, which is commonly termed the *a posteriori*[4] *probability* of x_i. The change of probability (from a priori to a posteriori) is closely related to how much information one can learn about x_i from the reception of y_j. Specifically, the difference between the uncertainty before and after receiving y_j measures the

[3] From the Latin, meaning "from what comes first" or "before."
[4] From the Latin, meaning "from what comes after" or "afterwards."

gain in information due to the reception of y_j. Such an information gain is called the *mutual information* and is naturally defined to be

$$\underbrace{I(x_i;y_j)}_{\substack{\text{information gain}\\ \text{or uncertainty loss}\\ \text{after receiving } y_j}} \triangleq \underbrace{\log_2\left(\frac{1}{P_X(x_i)}\right)}_{\substack{\text{uncertainty}\\ \text{before receiving } y_j}} - \underbrace{\log_2\left(\frac{1}{P_{X|Y}(x_i|y_j)}\right)}_{\substack{\text{uncertainty}\\ \text{after receiving } y_j}} \tag{6.63}$$

$$= \log_2\left(\frac{P_{X|Y}(x_i|y_j)}{P_X(x_i)}\right). \tag{6.64}$$

Note that if the two events $X = x_i$ and $Y = y_j$ are independent, thereby

$$P_{X|Y}(x_i|y_j) = P_X(x_i), \tag{6.65}$$

we have $I(x_i;y_j) = 0$, i.e. no information about x_i is gained once y_j is received.

For the noiseless channel, thus $y_j = x_i$, we have $P_{X|Y}(x_i|y_j) = 1$ since, based on what is received, we are completely certain about which input symbol has been sent. In this case, the mutual information attains the maximum value $\log_2(1/P_X(x_i))$; this means that all information about x_i is conveyed without any loss over the channel.

Since

$$P_{X|Y}(x_i|y_j)P_Y(y_j) = P_{X,Y}(x_i,y_j) = P_{Y|X}(y_j|x_i)P_X(x_i), \tag{6.66}$$

we have

$$I(x_i;y_j) = \log_2\left(\frac{P_{X,Y}(x_i,y_j)}{P_X(x_i)P_Y(y_j)}\right) = I(y_j;x_i). \tag{6.67}$$

Hence, we see that x_i provides the same amount of information about y_j as y_j does about x_i. This is why $I(x_i;y_j)$ has been coined "*mutual* information."

We have now characterized the mutual information with respect to a particular input–output event. Owing to the random nature of the source and channel output, the mutual information should be averaged with respect to both the input and output in order to account for the true statistical behavior of the channel. This motivates the following definition.

Definition 6.7 The *system mutual information*, or *average mutual information*, is defined as

$$I(X;Y) \triangleq \sum_{i=1}^{s}\sum_{j=1}^{t} P_{X,Y}(x_i,y_j)I(x_i;y_j) \tag{6.68}$$

$$= \sum_{i=1}^{s}\sum_{j=1}^{t} P_{X,Y}(x_i,y_j)\log_2\left(\frac{P_{X,Y}(x_i,y_j)}{P_X(x_i)P_Y(y_j)}\right). \tag{6.69}$$

The average mutual information $I(X;Y)$ has the following properties (the proofs are left as exercises).

Lemma 6.8 *The system mutual information has the following properties:*

(1) $I(X;Y) \geq 0$;
(2) $I(X;Y) = 0$ *if, and only if,* X *and* Y *are independent;*
(3) $I(X;Y) = I(Y;X)$.

Exercise 6.9 *Prove Lemma 6.8.*

Hint: For the first and second property use the IT Inequality (Lemma 5.10). You may proceed similarly to the proof of Lemma 5.11. The third property can be proven based on the formula of the mutual information, i.e. (6.69). ◊

Starting from (6.68), we have

$$I(X;Y) = \sum_{i=1}^{s}\sum_{j=1}^{t} P_{X,Y}(x_i,y_j)I(x_i;y_j) \tag{6.70}$$

$$= \sum_{i=1}^{s}\sum_{j=1}^{t} P_{X|Y}(x_i|y_j)P_Y(y_j)I(x_i;y_j) \tag{6.71}$$

$$= \sum_{j=1}^{t} P_Y(y_j)I(X;y_j), \tag{6.72}$$

where

$$I(X;y_j) \triangleq \sum_{i=1}^{s} P_{X|Y}(x_i|y_j)I(x_i;y_j) \tag{6.73}$$

measures the information about the entire input X provided by the reception of the particular y_j. In an analogous way we can obtain

$$I(X;Y) = \sum_{i=1}^{s}\sum_{j=1}^{t} P_{X,Y}(x_i,y_j)I(x_i;y_j) \tag{6.74}$$

$$= \sum_{i=1}^{s}\sum_{j=1}^{t} P_{Y|X}(y_j|x_i)P_X(x_i)I(x_i;y_j) \tag{6.75}$$

$$= \sum_{i=1}^{s} P_X(x_i)I(x_i;Y), \tag{6.76}$$

where

$$I(x_i;Y) \triangleq \sum_{j=1}^{t} P_{Y|X}(y_j|x_i)I(x_i;y_j) \tag{6.77}$$

represents the information about the output Y given that we know the input letter x_i is sent.

Let us end this section by specifying the relation between the average mutual information $I(X;Y)$ and various information quantities introduced thus far, e.g. input entropy $H(X)$, output entropy $H(Y)$, joint entropy $H(X,Y)$, and conditional entropies $H(X|Y)$ and $H(Y|X)$. To proceed, let us use (6.69) to express $I(X;Y)$ as

$$I(X;Y) = \sum_{i=1}^{s}\sum_{j=1}^{t} P_{X,Y}(x_i,y_j)\log_2\left(\frac{P_{X,Y}(x_i,y_j)}{P_X(x_i)P_Y(y_j)}\right) \tag{6.78}$$

$$= \sum_{i=1}^{s}\sum_{j=1}^{t} P_{X,Y}(x_i,y_j)\Big(\log_2 P_{X,Y}(x_i,y_j) - \log_2 P_X(x_i) - \log_2 P_Y(y_j)\Big) \tag{6.79}$$

$$= -\sum_{i=1}^{s}\sum_{j=1}^{t} P_{X,Y}(x_i,y_j)\log_2\left(\frac{1}{P_{X,Y}(x_i,y_j)}\right) + \sum_{i=1}^{s} P_X(x_i)\log_2\left(\frac{1}{P_X(x_i)}\right) + \sum_{j=1}^{t} P_Y(y_j)\log_2\left(\frac{1}{P_Y(y_j)}\right) \tag{6.80}$$

$$= H(X)+H(Y)-H(X,Y) \geq 0. \tag{6.81}$$

Since

$$H(X,Y) = H(X)+H(Y|X) \tag{6.82}$$

$$= H(Y)+H(X|Y), \tag{6.83}$$

we also have

$$I(X;Y) = H(X)-H(X|Y) \tag{6.84}$$

$$= H(Y)-H(Y|X). \tag{6.85}$$

We have the following corollary (the proof is left as an exercise).

Corollary 6.10 (Conditioning reduces entropy) *The following inequalities hold:*

$$(a) \qquad 0 \leq H(X|Y) \leq H(X), \tag{6.86}$$

$$0 \leq H(Y|X) \leq H(Y); \tag{6.87}$$

$$(b) \qquad H(X,Y) \leq H(X)+H(Y). \tag{6.88}$$

Part (a) of Corollary 6.10 asserts that conditioning cannot increase entropy, whereas Part (b) shows that the joint entropy $H(X,Y)$ is maximized when X and Y are independent.

Exercise 6.11 *Prove Corollary 6.10.*
Hint: Use known properties of $I(X;Y)$ and $H(X)$. ◊

To summarize: the average mutual information is given by

$$I(X;Y) = \begin{cases} H(X) + H(Y) - H(X,Y), \\ H(X) - H(X|Y), \\ H(Y) - H(Y|X); \end{cases} \tag{6.89}$$

the equivocation is given by

$$H(X|Y) = H(X) - I(X;Y), \tag{6.90}$$

$$H(Y|X) = H(Y) - I(X;Y); \tag{6.91}$$

the joint entropy is given by

$$H(X,Y) = \begin{cases} H(X) + H(Y) - I(X;Y), \\ H(X) + H(Y|X), \\ H(Y) + H(X|Y). \end{cases} \tag{6.92}$$

A schematic description of the relations between various information quantities is given by the Venn diagram in Figure 6.5.

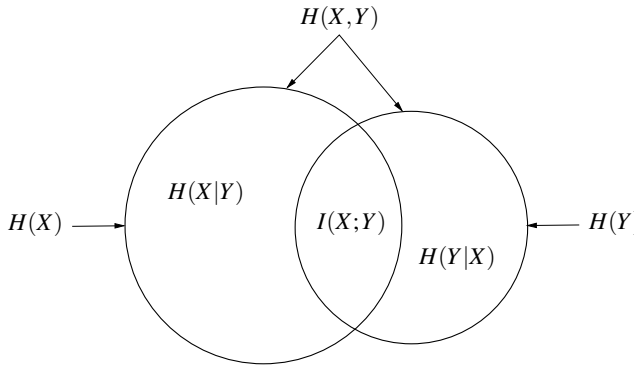

Figure 6.5 Relation between entropy, conditional entropy, and mutual information.

6.7 Definition of channel capacity

Given the conditional probabilities $P_{Y|X}(y_j|x_i)$, which define a channel, what is the maximum amount of information we can send through the channel? This is the main question attacked in the rest of this chapter.

The mutual information connects the two ends of the channel together. It is defined by (6.84) as

$$I(X;Y) = H(X) - H(X|Y), \tag{6.93}$$

where the entropy $H(X)$ is the uncertainty of the channel input *before* the reception of Y, and $H(X|Y)$ is the uncertainty that remains *after* the reception of Y. Thus $I(X;Y)$ is the change in the uncertainty. An alternative expression for $I(X;Y)$ is

$$I(X;Y) = \sum_{i=1}^{s}\sum_{j=1}^{t} P_X(x_i)P_{Y|X}(y_j|x_i)\log_2\left(\frac{P_{Y|X}(y_j|x_i)}{\sum_{i'=1}^{s} P_X(x_{i'})P_{Y|X}(y_j|x_{i'})}\right). \tag{6.94}$$

This formula involves the input symbol frequencies $P_X(x_i)$; in particular, for a given channel law $P_{Y|X}(y_j|x_i)$, $I(X;Y)$ depends completely on $P_X(x_i)$. We saw in the example of a binary symmetric channel (Section 6.4) how a poor match of $P_X(x_i)$ to the channel can ruin a channel. Indeed, we know that if the probability of one symbol is $P_X(x_i) = 1$, then all the others must be zero and the constant signal contains no information.

How can we best choose the $P_X(x_i)$ to get the most through the channel, and what is that amount?

Definition 6.12 (Capacity) For a given channel, the *channel capacity*, denoted by C, is defined to be the maximal achievable system mutual information $I(X;Y)$ among all possible input distributions $P_X(\cdot)$:

$$\mathrm{C} \triangleq \max_{P_X(\cdot)} I(X;Y). \tag{6.95}$$

Finding a closed-form solution to the channel capacity is in general difficult, except for some simple channels, e.g. the binary symmetric channel defined in Section 6.4 (see also Section 6.8 below).

We would like to point out that even though this definition of capacity is intuitively quite pleasing, at this stage it is a mere mathematical quantity, i.e. a number that is the result of a maximization problem. However, we will see in Section 6.11 that it really is the capacity of a channel in the sense that it is only possible to transmit signals reliably (i.e. with very small error probability) through the channel as long as the transmission rate is below the capacity.

6.8 Capacity of the binary symmetric channel

Consider again the binary symmetric channel (BSC), with the probability of transmission error equal to ε, as depicted in Figure 6.6.

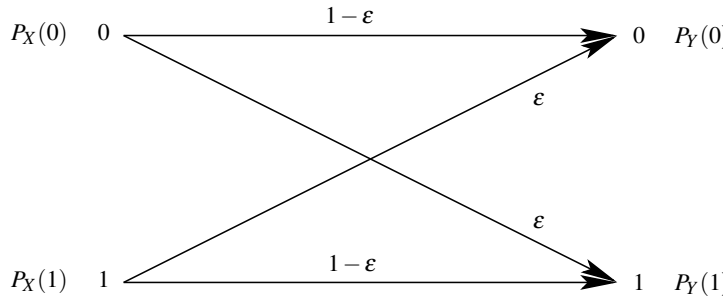

Figure 6.6 Binary symmetric channel (BSC).

The channel matrix is given by

$$\begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}. \tag{6.96}$$

Exercise 6.13 *For the BSC, show that*

$$H(Y|X=0) = H(Y|X=1) = H_{\mathrm{b}}(\varepsilon), \tag{6.97}$$

where $H_{\mathrm{b}}(\cdot)$ is the binary entropy function defined in (5.24). ◊

We start from the definition of mutual information (6.85) to obtain the following set of relations:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) && (6.98)\\ &= H(Y) - \sum_{x\in\mathcal{X}} P_X(x) H(Y|X=x) && (6.99)\\ &= H(Y) - \big(P_X(0)H(Y|X=0) + P_X(1)H(Y|X=1)\big) && (6.100)\\ &= H(Y) - H_{\mathrm{b}}(\varepsilon) && (6.101)\\ &\le 1 - H_{\mathrm{b}}(\varepsilon) \text{ bits}, && (6.102) \end{aligned}$$

where (6.102) follows since Y is a binary random variable (see Lemma 5.11). Since equality in (6.102) is attained if Y is uniform, which will hold if input X is uniform, we conclude that the capacity of the BSC is given by

$$\mathrm{C} = 1 - H_{\mathrm{b}}(\varepsilon) \text{ bits}, \tag{6.103}$$

and that the achieving input distribution is $P_X(0) = P_X(1) = 1/2$. Alternatively, we can find the capacity of a BSC by starting from $P_X(0) = \delta = 1 - P_X(1)$ and

expressing $I(X;Y)$ as

$$I(X;Y) = H(Y) - H(Y|X) \tag{6.104}$$

$$\begin{aligned} &= -\big(\delta(1-\varepsilon)+(1-\delta)\varepsilon\big)\log_2\big(\delta(1-\varepsilon)+(1-\delta)\varepsilon\big) \\ &\quad -\big(\delta\varepsilon+(1-\delta)(1-\varepsilon)\big)\log_2\big(\delta\varepsilon+(1-\delta)(1-\varepsilon)\big) \\ &\quad +(1-\varepsilon)\log_2(1-\varepsilon)+\varepsilon\log_2\varepsilon. \end{aligned} \tag{6.105}$$

If we now maximize the above quantity over $\delta \in [0,1]$, we find that the optimal δ is $\delta = 1/2$, which immediately yields (6.103).

Figure 6.7 depicts the mutual information in (6.105) versus δ with respect to three different choices of the error probability ε. As can be seen from the figure, the peak value of each curve is indeed attained by $\delta = 1/2$.

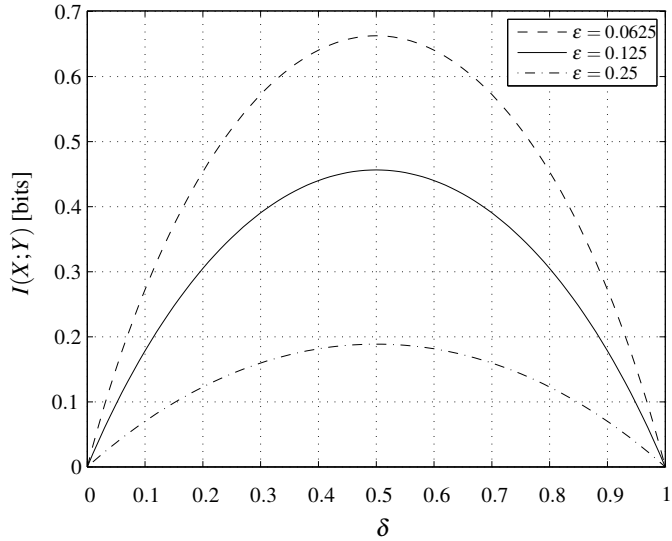

Figure 6.7 Mutual information over the binary symmetric channel (BSC): (6.105) as a function of δ, for various values of ε.

Figure 6.8 plots the capacity C in (6.103) versus the cross-over probability ε. We see from the figure that C attains the maximum value 1 bit when $\varepsilon = 0$ or $\varepsilon = 1$, and attains the minimal value 0 when $\varepsilon = 1/2$.

When $\varepsilon = 0$, it is easy to see that C = 1 bit is the maximum rate at which information can be communicated through the channel reliably. This can be achieved simply by transmitting uncoded bits through the channel, and no decoding is necessary because the bits are received unchanged. When $\varepsilon = 1$ the

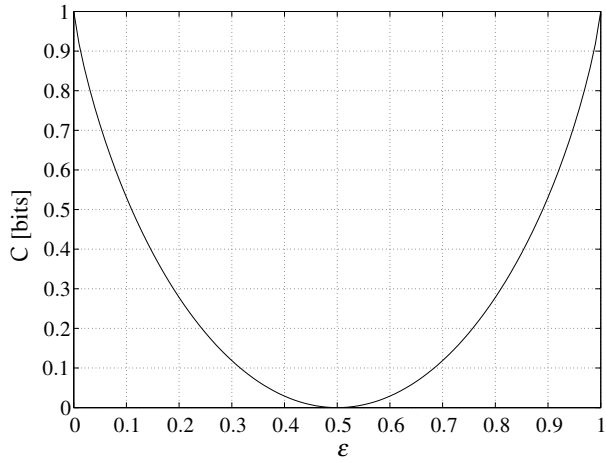

Figure 6.8 Capacity of the binary symmetric channel (BSC).

same can be achieved with the additional decoding step which complements all the received bits. By doing so, the bits transmitted through the channel can be recovered without error. Thus from a communications point of view, for binary channels, a channel which never makes an error and a channel which always makes an error are equally good.

When $\varepsilon = 1/2$, the channel output is independent of the channel input. Therefore, no information can possibly be communicated through the channel.

6.9 Uniformly dispersive channel

Recall that the channel transition matrix for the BSC is

$$\begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}, \tag{6.106}$$

in which the second row is a permutation of the first row. In fact, the BSC belongs to the class of *uniformly dispersive channels*.

Definition 6.14 A channel is said to be *uniformly dispersive* if the set

$$\mathcal{A}(x) \triangleq \{P_{Y|X}(y_1|x), \ldots, P_{Y|X}(y_t|x)\} \tag{6.107}$$

is identical for each input symbol x. Hence a uniformly dispersive channel has

a channel matrix

$$\begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1t} \\ P_{21} & P_{22} & \cdots & P_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ P_{s1} & P_{s2} & \cdots & P_{st} \end{pmatrix} \tag{6.108}$$

such that each row is a permutation of the first row.

According to the definition, for a uniformly dispersive channel the entropy of the output conditioned on a particular input alphabet x being sent, namely

$$H(Y|X=x) = \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x) \log_2 \left(\frac{1}{P_{Y|X}(y|x)} \right), \tag{6.109}$$

is thus identical for all x. By means of (6.109), the mutual information between the channel input and output reads

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) && (6.110)\\ &= H(Y) - \sum_{x\in\mathcal{X}} P_X(x) H(Y|X=x) && (6.111)\\ &= H(Y) - H(Y|X=x) \underbrace{\sum_{x\in\mathcal{X}} P_X(x)}_{=1} && (6.112)\\ &= H(Y) - H(Y|X=x). && (6.113) \end{aligned}$$

Equations (6.110)–(6.113) should be reminiscent of (6.98)–(6.101). This is no surprise since the BSC is uniformly dispersive.

Recall also from (6.102) that the capacity of the BSC is attained with a uniform output, which can be achieved by a uniform input. At a first glance one might expect that a similar argument can be directly applied to the uniformly dispersive case to find the capacity of an arbitrary uniformly dispersive channel. However, for a general uniformly dispersive channel, the uniform input does not necessarily result in the uniform output, and neither can the capacity necessarily be achieved with the uniform output. For example, consider the binary erasure channel (BEC) depicted in Figure 6.9. In this channel, the input alphabet is $\mathcal{X} = \{0,1\}$, while the output alphabet is $\mathcal{Y} = \{0,1,?\}$. With probability γ, the erasure symbol ? is produced at the output, which means that the input bit is lost; otherwise the input bit is reproduced at the output without error. The parameter γ is thus called the *erasure probability*. The BEC has the channel transition matrix

$$\begin{pmatrix} 1-\gamma & \gamma & 0 \\ 0 & \gamma & 1-\gamma \end{pmatrix} \tag{6.114}$$

and is thus, by definition, uniformly dispersive (the second row is a permutation of the first row). However, with the input distribution $P_X(0) = P_X(1) = 1/2$, the output is, in general, not uniform ($P_Y(?) = \gamma$ and $P_Y(0) = P_Y(1) = (1-\gamma)/2$). Despite this, the uniform input remains as the capacity-achieving input distribution for the BEC.

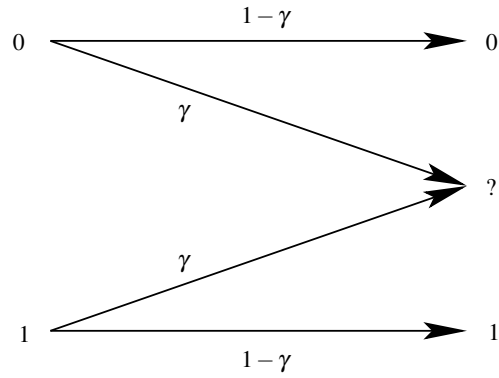

Figure 6.9 Binary erasure channel (BEC).

Exercise 6.15 *Based on (6.113), show that the capacity of the BEC is*

$$\mathrm{C} = 1 - \gamma \; \textit{bits}, \tag{6.115}$$

which is attained with $P_X(0) = P_X(1) = 1/2$. The result is intuitively reasonable: since a proportion γ of the bits are lost in the channel, we can recover (at most) a proportion $(1-\gamma)$ of the bits, and hence the capacity is $(1-\gamma)$. $\Diamond$

6.10 Characterization of the capacity-achieving input distribution

Even though it is, in general, difficult to find the closed-form capacity formula and the associated capacity-achieving input distribution, it is nonetheless possible to specify some underlying properties of the optimal $P_X(\cdot)$. The following theorem, which is stated without proof, provides one such characterization in terms of $I(x;Y)$, i.e. the information gain about Y given that $X = x$ is sent (see (6.77)).

Theorem 6.16 (Karush–Kuhn–Tucker (KKT) conditions) *An input distribution $P_X(\cdot)$ achieves the channel capacity* C *if, and only if,*

$$I(x;Y)\begin{cases} = \mathrm{C} & \text{for all } x \text{ with } P_X(x) > 0; \\ \leq \mathrm{C} & \text{for all } x \text{ with } P_X(x) = 0. \end{cases} \tag{6.116}$$

Remark 6.17 The KKT conditions were originally named after Harold W. Kuhn and Albert W. Tucker, who first published the conditions in 1951 [KT51]. Later, however, it was discovered that the necessary conditions for this problem had already been stated by William Karush in his master's thesis [Kar39].

The assertion of Theorem 6.16 is rather intuitive: if $P_X(\cdot)$ is the capacity-achieving input distribution and $P_X(x) > 0$, i.e. the particular letter x will be used with a nonvanishing probability to convey information over the channel, then the contribution of the mutual information due to this x must attain the capacity; otherwise there will exist another $P_{X'}(\cdot)$ capable of achieving the capacity by just disregarding this x (thus, $P_{X'}(x) = 0$) and using more often input letters other than x.

Theorem 6.16 can also be exploited for finding the capacity of some channels. Consider again the BSC case; the capacity should satisfy one of the following three cases:

$$\mathrm{C} = I(0;Y) = I(1;Y) \quad \text{for } P_X(0) > 0 \text{ and } P_X(1) > 0 \tag{6.117}$$

or

$$\mathrm{C} = I(0;Y) \geq I(1;Y) \quad \text{for } P_X(0) = 1 \text{ and } P_X(1) = 0 \tag{6.118}$$

or

$$\mathrm{C} = I(1;Y) \geq I(0;Y) \quad \text{for } P_X(0) = 0 \text{ and } P_X(1) = 1. \tag{6.119}$$

Since (6.118) and (6.119) only yield uninteresting zero capacity, it remains to verify whether or not (6.117) can give a positive capacity. By rearrangement (6.117) implies

$$\begin{aligned} \mathrm{C} &= I(0;Y) && (6.120)\\ &= \sum_{y=0}^{1} P_{Y|X}(y|0) \log_2 \frac{P_{Y|X}(y|0)}{P_Y(y)} && (6.121)\\ &= -\sum_{y=0}^{1} P_{Y|X}(y|0) \log_2 P_Y(y) + \sum_{y=0}^{1} P_{Y|X}(y|0) \log_2 P_{Y|X}(y|0) && (6.122)\\ &= -(1-\varepsilon)\log_2 P_Y(0) - \varepsilon \log_2 P_Y(1) - H_\mathrm{b}(\varepsilon) && (6.123) \end{aligned}$$

and

$$\begin{aligned}
\mathrm{C} &= I(1;Y) && (6.124)\\
&= \sum_{y=0}^{1} P_{Y|X}(y|1)\log_2\frac{P_{Y|X}(y|1)}{P_Y(y)} && (6.125)\\
&= -\sum_{y=0}^{1} P_{Y|X}(y|1)\log_2 P_Y(y) + \sum_{y=0}^{1} P_{Y|X}(y|1)\log_2 P_{Y|X}(y|1) && (6.126)\\
&= -\varepsilon\log_2 P_Y(0) - (1-\varepsilon)\log_2 P_Y(1) - H_\mathrm{b}(\varepsilon), && (6.127)
\end{aligned}$$

which yields

$$\begin{aligned}
&-(1-\varepsilon)\cdot\log_2 P_Y(0) - \varepsilon\cdot\log_2 P_Y(1) - H_\mathrm{b}(\varepsilon)\\
&\quad = -\varepsilon\cdot\log_2 P_Y(0) - (1-\varepsilon)\cdot\log_2 P_Y(1) - H_\mathrm{b}(\varepsilon). && (6.128)
\end{aligned}$$

This can only be satisfied if $P_Y(0) = P_Y(1)$ $(= 1/2)$. Thus

$$\begin{aligned}
\mathrm{C} &= -\varepsilon\cdot\log_2\left(\frac{1}{2}\right) - (1-\varepsilon)\cdot\log_2\left(\frac{1}{2}\right) - H_\mathrm{b}(\varepsilon) && (6.129)\\
&= 1 - H_\mathrm{b}(\varepsilon)\ \text{bits}. && (6.130)
\end{aligned}$$

Exercise 6.18 *Repeat the above arguments to derive the capacity of the BEC.* ◊

6.11 Shannon's Channel Coding Theorem

The channel capacity measures the amount of information that can be carried over the channel; in fact, it characterizes the maximal amount of transmission rate for reliable communication. Prior to the mid 1940s people believed that transmitted data subject to noise corruption can never be perfectly recovered unless the transmission rate approaches zero [Gal01]. Shannon's landmark work [Sha48] in 1948 disproved this thinking and established the well known *Channel Coding Theorem*: as long as the transmission rate in the same units as the channel capacity, e.g. information bits per channel use, is below (but can be arbitrarily close to) the channel capacity, the error can be made *smaller than any given number* (which we term *arbitrarily small*) by some properly designed coding scheme.

In what follows are some definitions that are required to state the theorem formally; detailed mathematical proofs can be found in [CT06] and [Gal68].

Definition 6.19 An (M, n) *coding scheme* for the channel $(\mathcal{X}, P_{Y|X}(y|x), \mathcal{Y})$ consists of the following.

(1) A message set $\{1,2,\ldots,\mathrm{M}\}$.

(2) An encoding function $\phi\colon \{1,2,\ldots,\mathrm{M}\} \to \mathcal{X}^n$, which is a rule that associates message m with a channel input sequence of length n, called the mth *codeword* and denoted by $\mathbf{x}^n(m)$. The set of all codewords

$$\{\mathbf{x}^n(1), \mathbf{x}^n(2), \ldots, \mathbf{x}^n(\mathrm{M})\}$$

is called the *codebook* (or simply the *code*).

(3) A decoding function $\psi\colon \mathcal{Y}^n \to \{1,2,\ldots,\mathrm{M}\}$, which is a deterministic rule that assigns a guess to each possible received vector.

Definition 6.20 (Rate) The *rate* R of an (M,n) coding scheme is defined to be

$$\mathrm{R} \triangleq \frac{\log_2 \mathrm{M}}{n} \quad \text{bits per transmission.} \tag{6.131}$$

In (6.131), $\log_2 \mathrm{M}$ describes the number of digits needed to write the numbers $0,\ldots,\mathrm{M}-1$ in binary form. For example, for $\mathrm{M}=8$ we need three binary digits (or bits): $000,\ldots,111$. The denominator n tells how many times the channel is used for the total transmission of a codeword (recall that n is the codeword length). Hence the rate describes how many bits are transmitted on average in each channel use.

Definition 6.21 Let

$$\lambda_m \triangleq \Pr[\,\psi(\mathbf{Y}^n) \neq m \,|\, \mathbf{X}^n = \mathbf{x}^n(m)] \tag{6.132}$$

be the conditional probability that the receiver makes a wrong guess given that the mth codeword is sent. The *average error probability* $\lambda^{(n)}$ for an (M,n) coding scheme is defined as

$$\lambda^{(n)} \triangleq \frac{1}{\mathrm{M}} \sum_{m=1}^{\mathrm{M}} \lambda_m. \tag{6.133}$$

Now we are ready for the famous *Channel Coding Theorem* due to Shannon.

Theorem 6.22 (Shannon's Channel Coding Theorem)
For a discrete-time information channel, it is possible to transmit messages with an arbitrarily small error probability (i.e. we have so-called reliable communication*), if the communication rate* R *is below the channel capacity* C*. Specifically, for every rate* $\mathrm{R} < \mathrm{C}$*, there exists a sequence of* $(2^{n\mathrm{R}}, n)$ *coding schemes*[5] *with average error probability* $\lambda^{(n)} \to 0$ *as* $n \to \infty$*.*
Conversely, any sequence of $(2^{n\mathrm{R}}, n)$ *coding schemes with* $\lambda^{(n)} \to 0$ *must have* $\mathrm{R} \leq \mathrm{C}$*. Hence, any attempt of transmitting at a rate larger than capacity will for sure fail in the sense that the average error probability is strictly larger than zero.*

Take the BSC for example. If the cross-over probability is $\varepsilon = 0.1$, the resulting capacity is $\mathrm{C} = 0.531$ bits per channel use. Hence reliable communication is only possible for coding schemes with a rate smaller than 0.531 bits per channel use.

Although the theorem shows that there exist good coding schemes with arbitrarily small error probability for long blocklength n, it does not provide a way of constructing the best coding schemes. Actually, the only knowledge we can infer from the theorem is perhaps "a good code favors a large blocklength." Ever since Shannon's original findings, researchers have tried to develop practical coding schemes that are easy to encode and decode; the Hamming code we discussed in Chapter 3 is the simplest of a class of algebraic error-correcting codes that can correct one error in a block of bits. Many other techniques have also been proposed to construct error-correcting codes, among which the *turbo code* – to be discussed in Chapter 7 – has come close to achieving the so-called *Shannon limit* for channels contaminated by Gaussian noise.

6.12 Some historical background

In his landmark paper [Sha48], Shannon only used H, R, and C to denote entropy, rate, and capacity, respectively. The first to use I for information were

[5] In Theorem 6.22, $2^{n\mathrm{R}}$ is a convenient expression for the code size and should be understood as either the smallest integer no less than its value or the largest integer no greater than its value. Researchers tend to drop the ceiling or flooring function applying to it, because the ratio of $2^{n\mathrm{R}}$, against the integer it is understood to be, will be very close to unity as n is large. Since the theorem actually deals with very large codeword lengths n (note that $\lambda^{(n)}$ approaches zero only when n is very large), the slack use of $2^{n\mathrm{R}}$ as an integer is somehow justified in concept.

Philip M. Woodward and Ian L. Davies in [WD52]. This paper is a very good read and gives an astoundingly clear overview of the fundamentals of information theory only four years after the theory had been established by Shannon. The authors give a slightly different interpretation of Shannon's theory and redevelop it using two additivity axioms. However, they did not yet use the name "mutual information." The name only starts to appear between 1954 and 1956. In 1954, Mark Pinsker published a paper in Russian [Pin54] with the title "Mutual information between a pair of stationary Gaussian random processes." However, depending on the translation, the title also might read "The quantity of information about a Gaussian random stationary process, contained in a second process connected with it in a stationary manner." Shannon certainly used the term "mutual information" in a paper about the zero-error capacity in 1956 [Sha56].

By the way, Woodward is also a main pioneer in modern radar theory. He had the insight to apply probability theory and statistics to the problem of recovering data from noisy samples. Besides this, he is a huge clock fan and made many contributions to horology; in particular, he built the world's most precise mechanical clock, the *Clock W5*, inventing a completely new mechanism.[6]

6.13 Further reading

Full discussions of the mutual information, channel capacity, and Shannon's Channel Coding Theorem in terms of probability theory can be found in many textbooks, see, e.g., [CT06] and [Gal68]. A unified discussion of the capacity results of the uniformly dispersive channel is given in [Mas96]. Further generalizations of uniformly dispersive channels are *quasi-symmetric channels*, discussed in [CA05, Chap. 4], and *T-symmetric channels*, described in [RG04]. The proof of Theorem 6.16 is closely related to the subject of constrained optimization, which is a standard technique for finding the channel capacity; see, e.g., [Gal68] and [CT06]. In addition to the Source Coding Theorem (Theorem 5.28 introduced in Chapter 5) and the Channel Coding Theorem (Theorem 6.22), Shannon's third landmark contribution is the development of the so-called *rate distortion theory*, which describes how to represent a continuous source with good fidelity using only a finite number of "representation levels"; for more details, please also refer to [CT06] and [Gal68].

[6] To see some amazing videos of this, search on http://www.youtube.com for "clock W5."

References

[BT02] Dimitri P. Bertsekas and John N. Tsitsiklis, *Introduction to Probability*. Athena Scientific, Belmont, MA, 2002.

[CA05] Po-Ning Chen and Fady Alajaji, *Lecture Notes on Information Theory,* vol. 1, Department of Electrical Engineering, National Chiao Tung University, Hsinchu, Taiwan, and Department of Mathematics & Statistics, Queen's University, Kingston, Canada, August 2005. Available: http://shannon.cm.nctu.edu.tw/it/itvol12004.pdf

[CT06] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, 2nd edn. John Wiley & Sons, Hoboken, NJ, 2006.

[Gal68] Robert G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, New York, 1968.

[Gal01] Robert G. Gallager, "Claude E. Shannon: a retrospective on his life, work, and impact," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2681–2695, November 2001.

[Kar39] William Karush, "Minima of functions of several variables with inequalities as side constraints," Master's thesis, Department of Mathematics, University of Chicago, Chicago, IL, 1939.

[KT51] Harold W. Kuhn and Albert W. Tucker, "Nonlinear programming," in *Proceedings of Second Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, ed. University of California Press, Berkeley, CA, 1951, pp. 481–492.

[Mas96] James L. Massey, *Applied Digital Information Theory I and II,* Lecture notes, Signal and Information Processing Laboratory, ETH Zurich, 1995/1996. Available: http://www.isiweb.ee.ethz.ch/archive/massey_scr/

[Pin54] Mark S. Pinsker, "Mutual information between a pair of stationary Gaussian random processes," (in Russian), *Doklady Akademii Nauk SSSR*, vol. 99, no. 2, pp. 213–216, 1954, also known as "The quantity of information about a Gaussian random stationary process, contained in a second process connected with it in a stationary manner."

[RG04] Mohammad Rezaeian and Alex Grant, "Computation of total capacity for discrete memoryless multiple-access channels," *IEEE Transactions on Information Theory*, vol. 50, no. 11, pp. 2779–2784, November 2004.

[Sha48] Claude E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, July and October 1948. Available: http://moser.cm.nctu.edu.tw/nctu/doc/shannon1948.pdf

[Sha56] Claude E. Shannon, "The zero error capacity of a noisy channel," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 8–19, September 1956.

[WD52] Philip M. Woodward and Ian L. Davies, "Information theory and inverse probability in telecommunication," *Proceedings of the IEE*, vol. 99, no. 58, pp. 37–44, March 1952.