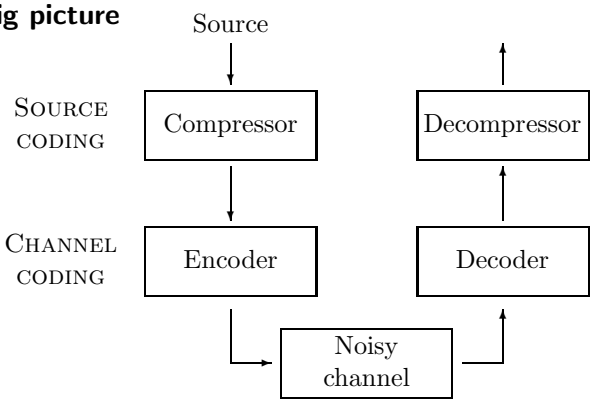
9

Communication over a Noisy Channel

▶ 9.1 The big picture



In Chapters 4–6, we discussed source coding with block codes, symbol codes and stream codes. We implicitly assumed that the channel from the compressor to the decompressor was noise-free. Real channels are noisy. We will now spend two chapters on the subject of noisy-channel coding – the fundamental possibilities and limitations of error-free communication through a noisy channel. The aim of channel coding is to make the noisy channel behave like a noiseless channel. We will assume that the data to be transmitted has been through a good compressor, so the bit stream has no obvious redundancy. The channel code, which makes the transmission, will put back redundancy of a special sort, designed to make the noisy received signal decodeable.

Suppose we transmit 1000 bits per second with $p_0 = p_1 = 1/2$ over a noisy channel that flips bits with probability $f = 0.1$. What is the rate of transmission of information? We might guess that the rate is 900 bits per second by subtracting the expected number of errors per second. But this is not correct, because the recipient does not know where the errors occurred. Consider the case where the noise is so great that the received symbols are independent of the transmitted symbols. This corresponds to a noise level of $f = 0.5$, since half of the received symbols are correct due to chance alone. But when $f = 0.5$, no information is transmitted at all.

Given what we have learnt about entropy, it seems reasonable that a measure of the information transmitted is given by the mutual information between the source and the received signal, that is, the entropy of the source minus the conditional entropy of the source given the received signal.

We will now review the definition of conditional entropy and mutual information. Then we will examine whether it is possible to use such a noisy channel to communicate *reliably*. We will show that for any channel Q there is a non-zero rate, the capacity $C(Q)$, up to which information can be sent

with arbitrarily small probability of error.

▶ 9.2 Review of probability and information

As an example, we take the joint distribution XY from exercise 8.6 (p.140). The marginal distributions $P(x)$ and $P(y)$ are shown in the margins.

$P(x,y)$		x				$P(y)$
		1	2	3	4	
	1	1/8	1/16	1/32	1/32	1/4
y	2	1/16	1/8	1/32	1/32	1/4
	3	1/16	1/16	1/16	1/16	1/4
	4	1/4	0	0	0	1/4
	$P(x)$	1/2	1/4	1/8	1/8	

The joint entropy is $H(X,Y) = 27/8$ bits. The marginal entropies are $H(X) = 7/4$ bits and $H(Y) = 2$ bits.

We can compute the conditional distribution of x for each value of y, and the entropy of each of those conditional distributions:

$P(x \mid y)$		x				$H(X \mid y)$/bits
		1	2	3	4	
	1	1/2	1/4	1/8	1/8	7/4
y	2	1/4	1/2	1/8	1/8	7/4
	3	1/4	1/4	1/4	1/4	2
	4	1	0	0	0	0
						$H(X \mid Y) = 11/8$

Note that whereas $H(X \mid y=4) = 0$ is less than $H(X)$, $H(X \mid y=3)$ is greater than $H(X)$. So in some cases, learning y can *increase* our uncertainty about x. Note also that although $P(x \mid y=2)$ is a different distribution from $P(x)$, the conditional entropy $H(X \mid y=2)$ is equal to $H(X)$. So learning that y is 2 changes our knowledge about x but does not reduce the uncertainty of x, as measured by the entropy. On average though, learning y does convey information about x, since $H(X \mid Y) < H(X)$.

One may also evaluate $H(Y|X) = 13/8$ bits. The mutual information is $I(X;Y) = H(X) - H(X \mid Y) = 3/8$ bits.

▶ 9.3 Noisy channels

A discrete memoryless channel Q is characterized by an input alphabet $\mathcal{A}_X$, an output alphabet $\mathcal{A}_Y$, and a set of conditional probability distributions $P(y \mid x)$, one for each $x \in \mathcal{A}_X$.

These *transition probabilities* may be written in a matrix

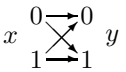
$$Q_{j|i} = P(y=b_j \mid x=a_i). \tag{9.1}$$

I usually orient this matrix with the output variable j indexing the rows and the input variable i indexing the columns, so that each column of $\mathbf{Q}$ is a probability vector. With this convention, we can obtain the probability of the output, $\mathbf{p}_Y$, from a probability distribution over the input, $\mathbf{p}_X$, by right-multiplication:

$$\mathbf{p}_Y = \mathbf{Q}\mathbf{p}_X. \tag{9.2}$$

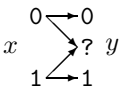
Some useful model channels are:

Binary symmetric channel. $\mathcal{A}_X = \{0, 1\}$. $\mathcal{A}_Y = \{0, 1\}$.

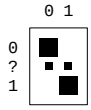


$$\begin{array}{lcllcl} P(y=0\,|\,x=0) &=& 1-f; & P(y=0\,|\,x=1) &=& f; \\ P(y=1\,|\,x=0) &=& f; & P(y=1\,|\,x=1) &=& 1-f. \end{array}$$

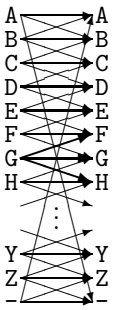
Binary erasure channel. $\mathcal{A}_X = \{0, 1\}$. $\mathcal{A}_Y = \{0, ?, 1\}$.



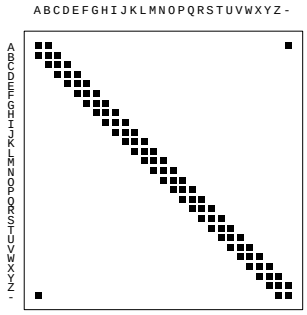
$$\begin{array}{lcllcl} P(y=0\,|\,x=0) &=& 1-f; & P(y=0\,|\,x=1) &=& 0; \\ P(y=?\,|\,x=0) &=& f; & P(y=?\,|\,x=1) &=& f; \\ P(y=1\,|\,x=0) &=& 0; & P(y=1\,|\,x=1) &=& 1-f. \end{array}$$



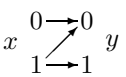
Noisy typewriter. $\mathcal{A}_X = \mathcal{A}_Y =$ the 27 letters {A, B, . . . , Z, -}. The letters are arranged in a circle, and when the typist attempts to type B, what comes out is either A, B or C, with probability 1/3 each; when the input is C, the output is B, C or D; and so forth, with the final letter '-' adjacent to the first letter A.



$$\begin{array}{lcl} \vdots & & \\ P(y=\mathtt{F}\,|\,x=\mathtt{G}) &=& 1/3; \\ P(y=\mathtt{G}\,|\,x=\mathtt{G}) &=& 1/3; \\ P(y=\mathtt{H}\,|\,x=\mathtt{G}) &=& 1/3; \\ \vdots & & \end{array}$$



Z channel. $\mathcal{A}_X = \{0, 1\}$. $\mathcal{A}_Y = \{0, 1\}$.



$$\begin{array}{lcllcl} P(y=0\,|\,x=0) &=& 1; & P(y=0\,|\,x=1) &=& f; \\ P(y=1\,|\,x=0) &=& 0; & P(y=1\,|\,x=1) &=& 1-f. \end{array}$$

▶ 9.4 Inferring the input given the output

If we assume that the input x to a channel comes from an ensemble X, then we obtain a joint ensemble XY in which the random variables x and y have the joint distribution:

$$P(x, y) = P(y\,|\,x)P(x). \tag{9.3}$$

Now if we receive a particular symbol y, what was the input symbol x? We typically won't know for certain. We can write down the posterior distribution of the input using Bayes' theorem:

$$P(x\,|\,y) = \frac{P(y\,|\,x)P(x)}{P(y)} = \frac{P(y\,|\,x)P(x)}{\sum_{x'} P(y\,|\,x')P(x')}. \tag{9.4}$$

Example 9.1. Consider a binary symmetric channel with probability of error $f = 0.15$. Let the input ensemble be $\mathcal{P}_X : \{p_0 = 0.9, p_1 = 0.1\}$. Assume we observe $y = 1$.

$$\begin{array}{rcl} P(x=1\,|\,y=1) &=& \dfrac{P(y=1\,|\,x=1)P(x=1)}{\sum_{x'} P(y\,|\,x')P(x')} \\ &=& \dfrac{0.85 \times 0.1}{0.85 \times 0.1 + 0.15 \times 0.9} \\ &=& \dfrac{0.085}{0.22} \;=\; 0.39. \end{array} \tag{9.5}$$

Thus '$x=1$' is still less probable than '$x=0$', although it is not as improbable as it was before.

Exercise 9.2.[1, p.157] Now assume we observe $y=0$. Compute the probability of $x=1$ given $y=0$.

Example 9.3. Consider a Z channel with probability of error $f=0.15$. Let the input ensemble be $\mathcal{P}_X : \{p_0=0.9, p_1=0.1\}$. Assume we observe $y=1$.

$$\begin{aligned} P(x=1\,|\,y=1) &= \frac{0.85\times 0.1}{0.85\times 0.1 + 0\times 0.9} \\ &= \frac{0.085}{0.085} = 1.0. \end{aligned} \tag{9.6}$$

So given the output $y=1$ we become certain of the input.

Exercise 9.4.[1, p.157] Alternatively, assume we observe $y=0$. Compute $P(x=1\,|\,y=0)$.

▶ 9.5 Information conveyed by a channel

We now consider how much information can be communicated through a channel. In operational terms, we are interested in finding ways of using the channel such that all the bits that are communicated are recovered with negligible probability of error. In mathematical terms, assuming a particular input ensemble X, we can measure how much information the output conveys about the input by the mutual information:

$$I(X;Y) \equiv H(X) - H(X\,|\,Y) = H(Y) - H(Y|X). \tag{9.7}$$

Our aim is to establish the connection between these two ideas. Let us evaluate $I(X;Y)$ for some of the channels above.

Hint for computing mutual information

We will tend to think of $I(X;Y)$ as $H(X) - H(X\,|\,Y)$, i.e., how much the uncertainty of the input X is reduced when we look at the output Y. But for computational purposes it is often handy to evaluate $H(Y)-H(Y|X)$ instead.

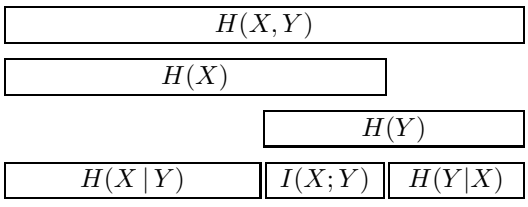


Figure 9.1. The relationship between joint information, marginal entropy, conditional entropy and mutual entropy. This figure is important, so I'm showing it twice.

Example 9.5. Consider the binary symmetric channel again, with $f=0.15$ and $\mathcal{P}_X : \{p_0=0.9, p_1=0.1\}$. We already evaluated the marginal probabilities $P(y)$ implicitly above: $P(y=0) = 0.78$; $P(y=1) = 0.22$. The mutual information is:

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

What is $H(Y|X)$? It is defined to be the weighted sum over x of $H(Y \mid x)$; but $H(Y \mid x)$ is the same for each value of x: $H(Y \mid x=0)$ is $H_2(0.15)$, and $H(Y \mid x=1)$ is $H_2(0.15)$. So

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H_2(0.22) - H_2(0.15) \\ &= 0.76 - 0.61 \;=\; 0.15 \text{ bits.} \end{aligned} \tag{9.8}$$

This may be contrasted with the entropy of the source $H(X) = H_2(0.1) = 0.47$ bits.

Note: here we have used the binary entropy function $H_2(p) \equiv H(p, 1-p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{(1-p)}$.

Throughout this book, log means $\log_2$.

Example 9.6. And now the Z channel, with $\mathcal{P}_X$ as above. $P(y=1) = 0.085$.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H_2(0.085) - [0.9H_2(0) + 0.1H_2(0.15)] \\ &= 0.42 - (0.1 \times 0.61) = 0.36 \text{ bits.} \end{aligned} \tag{9.9}$$

The entropy of the source, as above, is $H(X) = 0.47$ bits. Notice that the mutual information $I(X;Y)$ for the Z channel is bigger than the mutual information for the binary symmetric channel with the same f. The Z channel is a more reliable channel.

Exercise 9.7.[1, p.157] Compute the mutual information between X and Y for the binary symmetric channel with $f = 0.15$ when the input distribution is $\mathcal{P}_X = \{p_0 = 0.5, p_1 = 0.5\}$.

Exercise 9.8.[2, p.157] Compute the mutual information between X and Y for the Z channel with $f = 0.15$ when the input distribution is $\mathcal{P}_X$: $\{p_0 = 0.5, p_1 = 0.5\}$.

Maximizing the mutual information

We have observed in the above examples that the mutual information between the input and the output depends on the chosen input ensemble.

Let us assume that we wish to maximize the mutual information conveyed by the channel by choosing the best possible input ensemble. We define the *capacity* of the channel to be its maximum mutual information.

The capacity of a channel Q is:

$$C(Q) = \max_{\mathcal{P}_X} I(X;Y). \tag{9.10}$$

The distribution $\mathcal{P}_X$ that achieves the maximum is called the *optimal input distribution*, denoted by $\mathcal{P}_X^*$. [There may be multiple optimal input distributions achieving the same value of $I(X;Y)$.]

In Chapter 10 we will show that the capacity does indeed measure the maximum amount of error-free information that can be transmitted over the channel per unit time.

Example 9.9. Consider the binary symmetric channel with $f = 0.15$. Above, we considered $\mathcal{P}_X = \{p_0 = 0.9, p_1 = 0.1\}$, and found $I(X;Y) = 0.15$ bits.

How much better can we do? By symmetry, the optimal input distribution is $\{0.5, 0.5\}$ and the capacity is

$$C(Q_{\rm BSC}) = H_2(0.5) - H_2(0.15) = 1.0 - 0.61 = 0.39\,\text{bits}. \quad (9.11)$$

We'll justify the symmetry argument later. If there's any doubt about the symmetry argument, we can always resort to explicit maximization of the mutual information $I(X;Y)$,

$$I(X;Y) = H_2((1-f)p_1 + (1-p_1)f) - H_2(f) \quad \text{(figure 9.2)}. \quad (9.12)$$

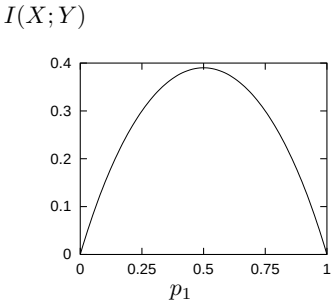


Figure 9.2. The mutual information $I(X;Y)$ for a binary symmetric channel with $f = 0.15$ as a function of the input distribution.

Example 9.10. The noisy typewriter. The optimal input distribution is a uniform distribution over x, and gives $C = \log_2 9$ bits.

Example 9.11. Consider the Z channel with $f = 0.15$. Identifying the optimal input distribution is not so straightforward. We evaluate $I(X;Y)$ explicitly for $\mathcal{P}_X = \{p_0, p_1\}$. First, we need to compute $P(y)$. The probability of $y = 1$ is easiest to write down:

$$P(y=1) = p_1(1-f). \quad (9.13)$$

Then the mutual information is:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H_2(p_1(1-f)) - (p_0 H_2(0) + p_1 H_2(f)) \\ &= H_2(p_1(1-f)) - p_1 H_2(f). \end{aligned} \quad (9.14)$$

This is a non-trivial function of p_1, shown in figure 9.3. It is maximized for $f = 0.15$ by $p_1^* = 0.445$. We find $C(Q_{\rm Z}) = 0.685$. Notice the optimal input distribution is not $\{0.5, 0.5\}$. We can communicate slightly more information by using input symbol 0 more frequently than 1.

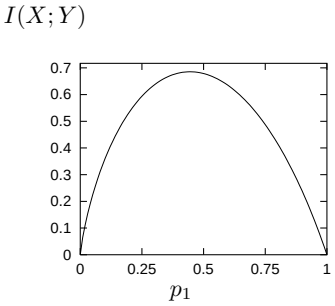


Figure 9.3. The mutual information $I(X;Y)$ for a Z channel with $f = 0.15$ as a function of the input distribution.

Exercise 9.12.[1, p.158] What is the capacity of the binary symmetric channel for general f?

Exercise 9.13.[2, p.158] Show that the capacity of the binary erasure channel with $f = 0.15$ is $C_{\rm BEC} = 0.85$. What is its capacity for general f? Comment.

▶ 9.6 The noisy-channel coding theorem

It seems plausible that the 'capacity' we have defined may be a measure of information conveyed by a channel; what is not obvious, and what we will prove in the next chapter, is that the capacity indeed measures the rate at which blocks of data can be communicated over the channel *with arbitrarily small probability of error.*

We make the following definitions.

An (N, K) block code for a channel Q is a list of $S = 2^K$ codewords

$$\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(2^K)}\}, \quad \mathbf{x}^{(s)} \in \mathcal{A}_X^N,$$

each of length N. Using this code we can encode a signal $s \in \{1, 2, 3, \ldots, 2^K\}$ as $\mathbf{x}^{(s)}$. [The number of codewords S is an integer, but the number of bits specified by choosing a codeword, $K \equiv \log_2 S$, is not necessarily an integer.]

The *rate* of the code is $R = K/N$ bits per channel use.

[We will use this definition of the rate for any channel, not only channels with binary inputs; note however that it is sometimes conventional to define the rate of a code for a channel with q input symbols to be $K/(N \log q)$.]

A decoder for an (N, K) block code is a mapping from the set of length-N strings of channel outputs, $\mathcal{A}_Y^N$, to a codeword label $\hat{s} \in \{0, 1, 2, \ldots, 2^K\}$.

The extra symbol $\hat{s} = 0$ can be used to indicate a 'failure'.

The probability of block error of a code and decoder, for a given channel, and for a given probability distribution over the encoded signal $P(s_{\text{in}})$, is:

$$p_{\text{B}} = \sum_{s_{\text{in}}} P(s_{\text{in}}) P(s_{\text{out}} \neq s_{\text{in}} \,|\, s_{\text{in}}). \tag{9.15}$$

The maximal probability of block error is

$$p_{\text{BM}} = \max_{s_{\text{in}}} P(s_{\text{out}} \neq s_{\text{in}} \,|\, s_{\text{in}}). \tag{9.16}$$

The optimal decoder for a channel code is the one that minimizes the probability of block error. It decodes an output $\mathbf{y}$ as the input s that has maximum posterior probability $P(s \,|\, \mathbf{y})$.

$$P(s \,|\, \mathbf{y}) = \frac{P(\mathbf{y} \,|\, s) P(s)}{\sum_{s'} P(\mathbf{y} \,|\, s') P(s')} \tag{9.17}$$

$$\hat{s}_{\text{optimal}} = \text{argmax}\, P(s \,|\, \mathbf{y}). \tag{9.18}$$

A uniform prior distribution on s is usually assumed, in which case the optimal decoder is also the *maximum likelihood decoder*, i.e., the decoder that maps an output $\mathbf{y}$ to the input s that has maximum *likelihood* $P(\mathbf{y} \,|\, s)$.

The probability of bit error p_{b} is defined assuming that the codeword number s is represented by a binary vector $\mathbf{s}$ of length K bits; it is the average probability that a bit of $\mathbf{s}_{\text{out}}$ is not equal to the corresponding bit of $\mathbf{s}_{\text{in}}$ (averaging over all K bits).

Shannon's noisy-channel coding theorem (part one). Associated with each discrete memoryless channel, there is a non-negative number C (called the channel capacity) with the following property. For any $\epsilon > 0$ and $R < C$, for large enough N, there exists a block code of length N and rate $\geq R$ and a decoding algorithm, such that the maximal probability of block error is $< \epsilon$.

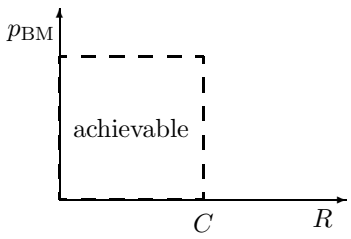


Figure 9.4. Portion of the R, p_{BM} plane asserted to be achievable by the first part of Shannon's noisy channel coding theorem.

Confirmation of the theorem for the noisy typewriter channel

In the case of the noisy typewriter, we can easily confirm the theorem, because we can create a completely error-free communication strategy using a block code of length $N = 1$: we use only the letters `B`, `E`, `H`, ..., `Z`, i.e., every third letter. These letters form a *non-confusable subset* of the input alphabet (see figure 9.5). Any output can be uniquely decoded. The number of inputs in the non-confusable subset is 9, so the error-free information rate of this system is $\log_2 9$ bits, which is equal to the capacity C, which we evaluated in example 9.10 (p.151).

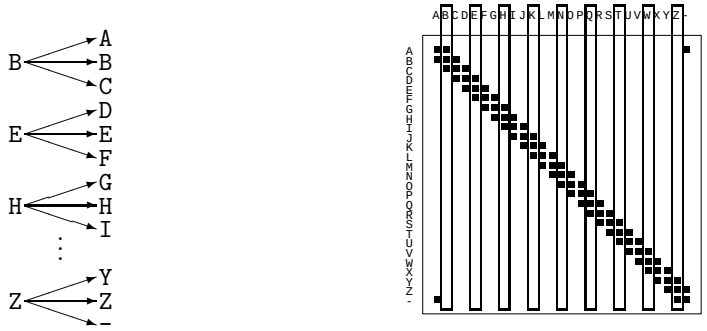


Figure 9.5. A non-confusable subset of inputs for the noisy typewriter.

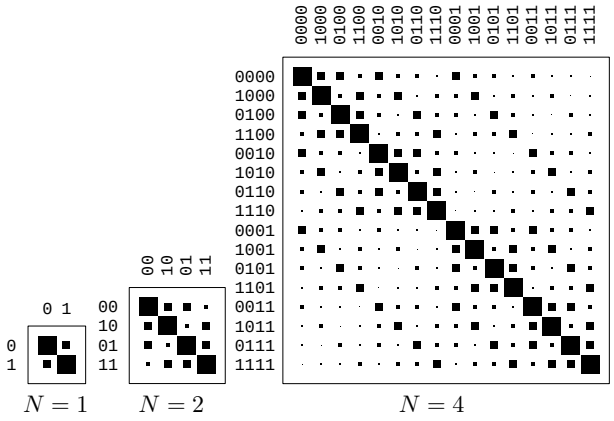


Figure 9.6. Extended channels obtained from a binary symmetric channel with transition probability 0.15.

How does this translate into the terms of the theorem? The following table explains.

The theorem	How it applies to the noisy typewriter
Associated with each discrete memoryless channel, there is a non-negative number C.	The capacity C is $\log_2 9$.
For any $\epsilon > 0$ and $R < C$, for large enough N,	No matter what ϵ and R are, we set the blocklength N to 1.
there exists a block code of length N and rate $\geq R$	The block code is $\{\mathtt{B}, \mathtt{E}, \ldots, \mathtt{Z}\}$. The value of K is given by $2^K = 9$, so $K = \log_2 9$, and this code has rate $\log_2 9$, which is greater than the requested value of R.
and a decoding algorithm,	The decoding algorithm maps the received letter to the nearest letter in the code;
such that the maximal probability of block error is $< \epsilon$.	the maximal probability of block error is zero, which is less than the given ϵ.

▶ 9.7 Intuitive preview of proof

Extended channels

To prove the theorem for any given channel, we consider the *extended channel* corresponding to N uses of the channel. The extended channel has $|\mathcal{A}_X|^N$ possible inputs $\mathbf{x}$ and $|\mathcal{A}_Y|^N$ possible outputs. Extended channels obtained from a binary symmetric channel and from a Z channel are shown in figures 9.6 and 9.7, with $N = 2$ and $N = 4$.

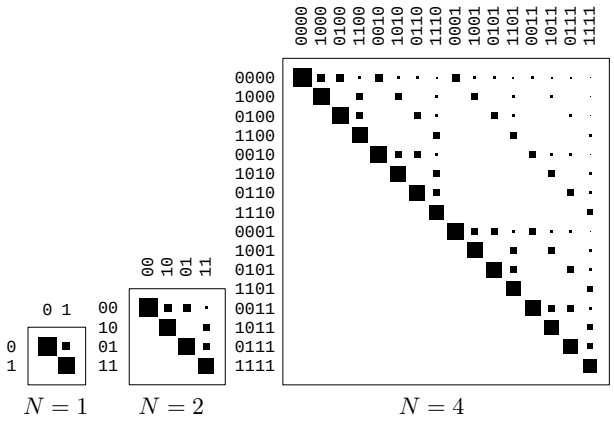


Figure 9.7. Extended channels obtained from a Z channel with transition probability 0.15. Each column corresponds to an input, and each row is a different output.

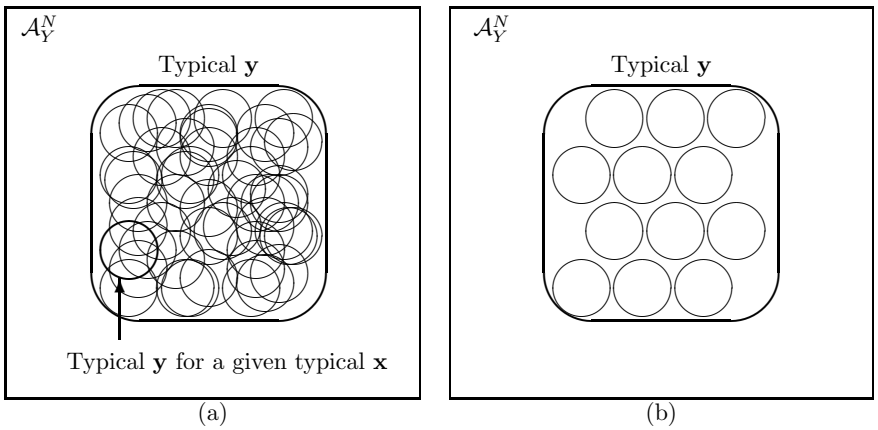


Figure 9.8. (a) Some typical outputs in $\mathcal{A}_Y^N$ corresponding to typical inputs $\mathbf{x}$. (b) A subset of the typical sets shown in (a) that do not overlap each other. This picture can be compared with the solution to the noisy typewriter in figure 9.5.

Exercise 9.14.[2, p.159] Find the transition probability matrices $\mathbf{Q}$ for the extended channel, with $N = 2$, derived from the binary erasure channel having erasure probability 0.15.

By selecting two columns of this transition probability matrix, we can define a rate-$1/2$ code for this channel with blocklength $N = 2$. What is the best choice of two columns? What is the decoding algorithm?

To prove the noisy-channel coding theorem, we make use of large blocklengths N. The intuitive idea is that, if N is large, *an extended channel looks a lot like the noisy typewriter.* Any particular input $\mathbf{x}$ is very likely to produce an output in a small subspace of the output alphabet – the typical output set, given that input. So we can find a non-confusable subset of the inputs that produce essentially disjoint output sequences. For a given N, let us consider a way of generating such a non-confusable subset of the inputs, and count up how many distinct inputs it contains.

Imagine making an input sequence $\mathbf{x}$ for the extended channel by drawing it from an ensemble X^N, where X is an arbitrary ensemble over the input alphabet. Recall the source coding theorem of Chapter 4, and consider the number of probable output sequences $\mathbf{y}$. The total number of typical output sequences $\mathbf{y}$ is $2^{NH(Y)}$, all having similar probability. For any particular typical input sequence $\mathbf{x}$, there are about $2^{NH(Y|X)}$ probable sequences. Some of these subsets of $\mathcal{A}_Y^N$ are depicted by circles in figure 9.8a.

We now imagine restricting ourselves to a subset of the typical inputs $\mathbf{x}$ such that the corresponding typical output sets do not overlap, as shown in figure 9.8b. We can then bound the number of non-confusable inputs by dividing the size of the typical $\mathbf{y}$ set, $2^{NH(Y)}$, by the size of each typical-$\mathbf{y}$-

given-typical-$\mathbf{x}$ set, $2^{NH(Y|X)}$. So the number of non-confusable inputs, if they are selected from the set of typical inputs $\mathbf{x} \sim X^N$, is $\leq 2^{NH(Y)-NH(Y|X)} = 2^{NI(X;Y)}$.

The maximum value of this bound is achieved if X is the ensemble that maximizes $I(X;Y)$, in which case the number of non-confusable inputs is $\leq 2^{NC}$. Thus asymptotically up to C bits per cycle, and no more, can be communicated with vanishing error probability. □

This sketch has not rigorously proved that reliable communication really is possible – that's our task for the next chapter.

▶ 9.8 Further exercises

Exercise 9.15.[3, p.159] Refer back to the computation of the capacity of the Z channel with $f = 0.15$.

(a) Why is p_1^* less than 0.5? One could argue that it is good to favour the 0 input, since it is transmitted without error – and also argue that it is good to favour the 1 input, since it often gives rise to the highly prized 1 output, which allows certain identification of the input! Try to make a convincing argument.

(b) In the case of general f, show that the optimal input distribution is

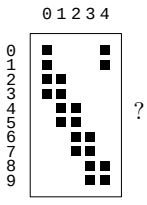
$$p_1^* = \frac{1/(1-f)}{1 + 2^{(H_2(f)/(1-f))}}. \tag{9.19}$$

(c) What happens to p_1^* if the noise level f is very close to 1?

Exercise 9.16.[2, p.159] Sketch graphs of the capacity of the Z channel, the binary symmetric channel and the binary erasure channel as a function of f.

▷ Exercise 9.17.[2] What is the capacity of the five-input, ten-output channel whose transition probability matrix is

$$\begin{bmatrix} 0.25 & 0 & 0 & 0 & 0.25 \\ 0.25 & 0 & 0 & 0 & 0.25 \\ 0.25 & 0.25 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.25 \\ 0 & 0 & 0 & 0.25 & 0.25 \end{bmatrix} \tag{9.20}$$



Exercise 9.18.[2, p.159] Consider a Gaussian channel with binary input $x \in \{-1, +1\}$ and *real* output alphabet $\mathcal{A}_Y$, with transition probability density

$$Q(y \,|\, x, \alpha, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-x\alpha)^2}{2\sigma^2}}, \tag{9.21}$$

where α is the signal amplitude.

(a) Compute the posterior probability of x given y, assuming that the two inputs are equiprobable. Put your answer in the form

$$P(x=1 \,|\, y, \alpha, \sigma) = \frac{1}{1 + e^{-a(y)}}. \tag{9.22}$$

Sketch the value of $P(x=1 \mid y, \alpha, \sigma)$ as a function of y.

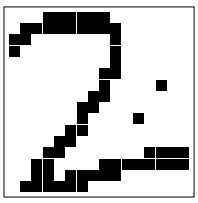
(b) Assume that a single bit is to be transmitted. What is the optimal decoder, and what is its probability of error? Express your answer in terms of the signal-to-noise ratio α^2/σ^2 and the error function (the cumulative probability function of the Gaussian distribution),

$$\Phi(z) \equiv \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, \mathrm{d}z. \tag{9.23}$$

[Note that this definition of the error function $\Phi(z)$ may not correspond to other people's.]

Pattern recognition as a noisy channel

We may think of many pattern recognition problems in terms of communication channels. Consider the case of recognizing handwritten digits (such as postcodes on envelopes). The author of the digit wishes to communicate a message from the set $\mathcal{A}_X = \{0, 1, 2, 3, \ldots, 9\}$; this selected message is the input to the channel. What comes out of the channel is a pattern of ink on paper. If the ink pattern is represented using 256 binary pixels, the channel Q has as its output a random variable $y \in \mathcal{A}_Y = \{0, 1\}^{256}$. An example of an element from this alphabet is shown in the margin.



Exercise 9.19.[2] Estimate how many patterns in $\mathcal{A}_Y$ are recognizable as the character '2'. [The aim of this problem is to try to demonstrate the existence of *as many patterns as possible* that are recognizable as 2s.]

Discuss how one might model the channel $P(y \mid x=2)$. Estimate the entropy of the probability distribution $P(y \mid x=2)$.

One strategy for doing pattern recognition is to create a model for $P(y \mid x)$ for each value of the input $x = \{0, 1, 2, 3, \ldots, 9\}$, then use Bayes' theorem to infer x given y.

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{\sum_{x'} P(y \mid x')P(x')}. \tag{9.24}$$

This strategy is known as *full probabilistic modelling* or *generative modelling*. This is essentially how current speech recognition systems work. In addition to the channel model, $P(y \mid x)$, one uses a prior probability distribution $P(x)$, which in the case of both character recognition and speech recognition is a language model that specifies the probability of the next character/word given the context and the known grammar and statistics of the language.

Figure 9.9. Some more 2s.

Random coding

Exercise 9.20.[2, p.160] Given twenty-four people in a room, what is the probability that there are at least two people present who have the same birthday (i.e., day and month of birth)? What is the expected number of pairs of people with the same birthday? Which of these two questions is easiest to solve? Which answer gives most insight? You may find it helpful to solve these problems and those that follow using notation such as A = number of days in year = 365 and S = number of people = 24.

▷ Exercise 9.21.[2] The birthday problem may be related to a coding scheme. Assume we wish to convey a message to an outsider identifying one of

the twenty-four people. We could simply communicate a number s from $\mathcal{A}_S = \{1, 2, \ldots, 24\}$, having agreed a mapping of people onto numbers; alternatively, we could convey a number from $\mathcal{A}_X = \{1, 2, \ldots, 365\}$, identifying the day of the year that is the selected person's birthday (with apologies to leapyearians). [The receiver is assumed to know all the people's birthdays.] What, roughly, is the probability of error of this communication scheme, assuming it is used for a single transmission? What is the capacity of the communication channel, and what is the rate of communication attempted by this scheme?

▷ Exercise 9.22.[2] Now imagine that there are K rooms in a building, each containing q people. (You might think of $K = 2$ and $q = 24$ as an example.) The aim is to communicate a selection of one person from each room by transmitting an ordered list of K days (from $\mathcal{A}_X$). Compare the probability of error of the following two schemes.

(a) As before, where each room transmits the birthday of the selected person.

(b) To each K-tuple of people, one drawn from each room, an ordered K-tuple of randomly selected days from $\mathcal{A}_X$ is assigned (this K-tuple has nothing to do with their birthdays). This enormous list of $S = q^K$ strings is known to the receiver. When the building has selected a particular person from each room, the ordered string of days corresponding to that K-tuple of people is transmitted.

What is the probability of error when $q = 364$ and $K = 1$? What is the probability of error when $q = 364$ and K is large, e.g. $K = 6000$?

▶ 9.9 Solutions

Solution to exercise 9.2 (p.149). If we assume we observe $y = 0$,

$$P(x=1 \mid y=0) = \frac{P(y=0 \mid x=1)P(x=1)}{\sum_{x'} P(y \mid x')P(x')} \tag{9.25}$$

$$= \frac{0.15 \times 0.1}{0.15 \times 0.1 + 0.85 \times 0.9} \tag{9.26}$$

$$= \frac{0.015}{0.78} = 0.019. \tag{9.27}$$

Solution to exercise 9.4 (p.149). If we observe $y = 0$,

$$P(x=1 \mid y=0) = \frac{0.15 \times 0.1}{0.15 \times 0.1 + 1.0 \times 0.9} \tag{9.28}$$

$$= \frac{0.015}{0.915} = 0.016. \tag{9.29}$$

Solution to exercise 9.7 (p.150). The probability that $y = 1$ is 0.5, so the mutual information is:

$$I(X;Y) = H(Y) - H(Y \mid X) \tag{9.30}$$

$$= H_2(0.5) - H_2(0.15) \tag{9.31}$$

$$= 1 - 0.61 = 0.39 \text{ bits.} \tag{9.32}$$

Solution to exercise 9.8 (p.150). We again compute the mutual information using $I(X;Y) = H(Y) - H(Y \mid X)$. The probability that $y = 0$ is 0.575, and

$H(Y \mid X) = \sum_x P(x) H(Y \mid x) = P(x=1) H(Y \mid x=1) + P(x=0) H(Y \mid x=0)$ so the mutual information is:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y \mid X) && (9.33)\\ &= H_2(0.575) - [0.5 \times H_2(0.15) + 0.5 \times 0] && (9.34)\\ &= 0.98 - 0.30 \;=\; 0.679 \text{ bits.} && (9.35) \end{aligned}$$

Solution to exercise 9.12 (p.151). By symmetry, the optimal input distribution is $\{0.5, 0.5\}$. Then the capacity is

$$\begin{aligned} C \;=\; I(X;Y) &= H(Y) - H(Y \mid X) && (9.36)\\ &= H_2(0.5) - H_2(f) && (9.37)\\ &= 1 - H_2(f). && (9.38) \end{aligned}$$

Would you like to find the optimal input distribution without invoking symmetry? We can do this by computing the mutual information in the general case where the input ensemble is $\{p_0, p_1\}$:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y \mid X) && (9.39)\\ &= H_2(p_0 f + p_1(1-f)) - H_2(f). && (9.40) \end{aligned}$$

The only p-dependence is in the first term $H_2(p_0 f + p_1(1-f))$, which is maximized by setting the argument to 0.5. This value is given by setting $p_0 = 1/2$.

Solution to exercise 9.13 (p.151). Answer 1. By symmetry, the optimal input distribution is $\{0.5, 0.5\}$. The capacity is most easily evaluated by writing the mutual information as $I(X;Y) = H(X) - H(X \mid Y)$. The conditional entropy $H(X \mid Y)$ is $\sum_y P(y) H(X \mid y)$; when y is known, x is uncertain only if $y = ?$, which occurs with probability $f/2 + f/2$, so the conditional entropy $H(X \mid Y)$ is $f H_2(0.5)$.

$$\begin{aligned} C \;=\; I(X;Y) &= H(X) - H(X \mid Y) && (9.41)\\ &= H_2(0.5) - f H_2(0.5) && (9.42)\\ &= 1 - f. && (9.43) \end{aligned}$$

The binary erasure channel fails a fraction f of the time. Its capacity is precisely $1 - f$, which is the fraction of the time that the channel is reliable. This result seems very reasonable, but it is far from obvious how to encode information so as to communicate *reliably* over this channel.

Answer 2. Alternatively, without invoking the symmetry assumed above, we can start from the input ensemble $\{p_0, p_1\}$. The probability that $y = ?$ is $p_0 f + p_1 f = f$, and when we receive $y = ?$, the posterior probability of x is the same as the prior probability, so:

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) && (9.44)\\ &= H_2(p_1) - f H_2(p_1) && (9.45)\\ &= (1-f) H_2(p_1). && (9.46) \end{aligned}$$

This mutual information achieves its maximum value of $(1-f)$ when $p_1 = 1/2$.

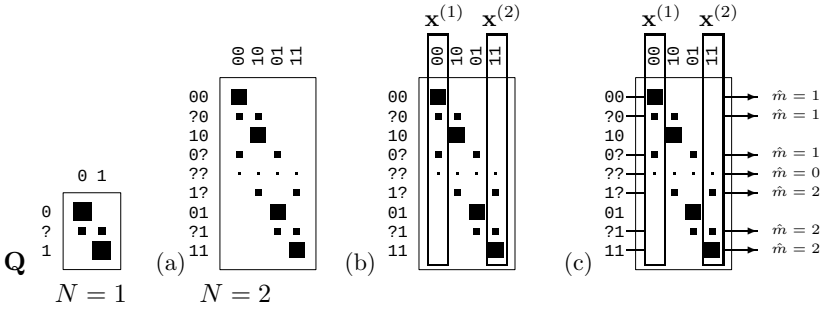


Figure 9.10. (a) The extended channel ($N = 2$) obtained from a binary erasure channel with erasure probability 0.15. (b) A block code consisting of the two codewords 00 and 11. (c) The optimal decoder for this code.

Solution to exercise 9.14 (p.153). The extended channel is shown in figure 9.10. The best code for this channel with $N = 2$ is obtained by choosing two columns that have minimal overlap, for example, columns 00 and 11. The decoding algorithm returns '00' if the extended channel output is among the top four and '11' if it's among the bottom four, and gives up if the output is '??'.

Solution to exercise 9.15 (p.155). In example 9.11 (p.151) we showed that the mutual information between input and output of the Z channel is

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y \mid X) \\ &= H_2(p_1(1-f)) - p_1 H_2(f). \end{aligned} \tag{9.47}$$

We differentiate this expression with respect to p_1, taking care not to confuse $\log_2$ with $\log_e$:

$$\frac{\mathrm{d}}{\mathrm{d}p_1} I(X;Y) = (1-f)\log_2 \frac{1-p_1(1-f)}{p_1(1-f)} - H_2(f). \tag{9.48}$$

Setting this derivative to zero and rearranging using skills developed in exercise 2.17 (p.36), we obtain:

$$p_1^*(1-f) = \frac{1}{1+2^{H_2(f)/(1-f)}}, \tag{9.49}$$

so the optimal input distribution is

$$p_1^* = \frac{1/(1-f)}{1+2^{(H_2(f)/(1-f))}}. \tag{9.50}$$

As the noise level f tends to 1, this expression tends to $1/e$ (as you can prove using L'Hôpital's rule).

For all values of f, p_1^* is smaller than 1/2. A rough intuition for why input 1 is used less than input 0 is that when input 1 is used, the noisy channel injects entropy into the received string; whereas when input 0 is used, the noise has zero entropy.

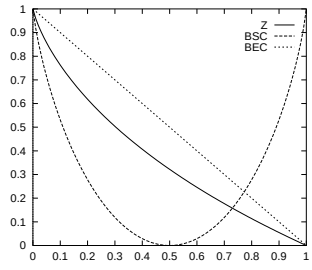


Figure 9.11. Capacities of the Z channel, binary symmetric channel, and binary erasure channel.

Solution to exercise 9.16 (p.155). The capacities of the three channels are shown in figure 9.11. For any $f < 0.5$, the BEC is the channel with highest capacity and the BSC the lowest.

Solution to exercise 9.18 (p.155). The logarithm of the posterior probability ratio, given y, is

$$a(y) = \ln \frac{P(x=1 \mid y, \alpha, \sigma)}{P(x=-1 \mid y, \alpha, \sigma)} = \ln \frac{Q(y \mid x=1, \alpha, \sigma)}{Q(y \mid x=-1, \alpha, \sigma)} = 2\frac{\alpha y}{\sigma^2}. \tag{9.51}$$

Using our skills picked up from exercise 2.17 (p.36), we rewrite this in the form

$$P(x=1\,|\,y,\alpha,\sigma)=\frac{1}{1+e^{-a(y)}}. \tag{9.52}$$

The optimal decoder selects the most probable hypothesis; this can be done simply by looking at the sign of $a(y)$. If $a(y)>0$ then decode as $\hat{x}=1$.

The probability of error is

$$p_{\rm b}=\int_{-\infty}^{0}{\rm d}y\,Q(y\,|\,x=1,\alpha,\sigma)=\int_{-\infty}^{-x\alpha}{\rm d}y\,\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{y^2}{2\sigma^2}}=\Phi\left(-\frac{x\alpha}{\sigma}\right). \tag{9.53}$$

Random coding

Solution to exercise 9.20 (p.156). The probability that $S=24$ people whose birthdays are drawn at random from $A=365$ days all have *distinct* birthdays is

$$\frac{A(A-1)(A-2)\ldots(A-S+1)}{A^S}. \tag{9.54}$$

The probability that two (or more) people share a birthday is one minus this quantity, which, for $S=24$ and $A=365$, is about 0.5. This exact way of answering the question is not very informative since it is not clear for what value of S the probability changes from being close to 0 to being close to 1.

The number of pairs is $S(S-1)/2$, and the probability that a particular pair shares a birthday is $1/A$, so the *expected number* of collisions is

$$\frac{S(S-1)}{2}\frac{1}{A}. \tag{9.55}$$

This answer is more instructive. The expected number of collisions is tiny if $S\ll\sqrt{A}$ and big if $S\gg\sqrt{A}$.

We can also approximate the probability that all birthdays are distinct, for small S, thus:

$$\begin{aligned}\frac{A(A-1)(A-2)\ldots(A-S+1)}{A^S} &= (1)(1-{}^1\!/_A)(1-{}^2\!/_A)\ldots(1-{}^{(S-1)}\!/_A)\\ &\simeq \exp(0)\exp(-1/A)\exp(-2/A)\ldots\exp(-(S-1)/A) \qquad (9.56)\\ &\simeq \exp\left(-\frac{1}{A}\sum_{i=1}^{S-1}i\right)=\exp\left(-\frac{S(S-1)/2}{A}\right). \qquad (9.57)\end{aligned}$$