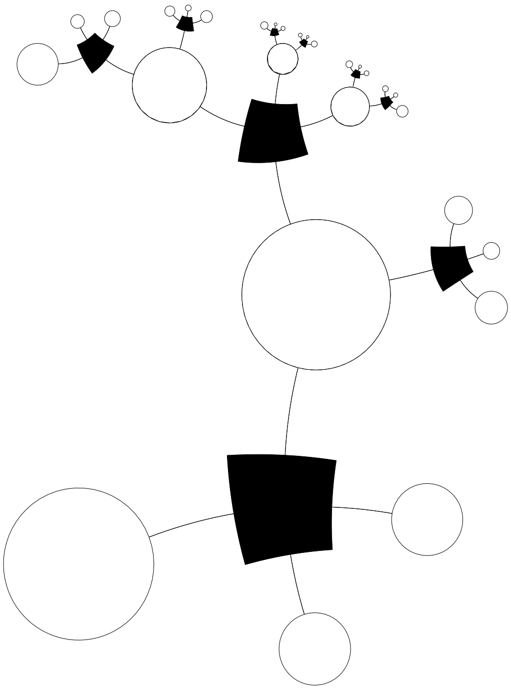
Part II

Noisy-Channel Coding



8

Correlated Random Variables

In the last three chapters on data compression we concentrated on random vectors $\mathbf{x}$ coming from an extremely simple probability distribution, namely the separable distribution in which each component x_n is independent of the others.

In this chapter, we consider *joint ensembles* in which the random variables are correlated. This material has two motivations. First, data from the real world have interesting correlations, so to do data compression well, we need to know how to work with models that include correlations. Second, a noisy channel with input x and output y defines a joint ensemble in which x and y are correlated – if they were independent, it would be impossible to communicate over the channel – so communication over noisy channels (the topic of chapters 9–11) is described in terms of the entropy of joint ensembles.

▶ 8.1 More about entropy

This section gives definitions and exercises to do with entropy, carrying on from section 2.4.

The joint entropy of X, Y is:

$$H(X,Y) = \sum_{xy \in \mathcal{A}_X\mathcal{A}_Y} P(x,y) \log \frac{1}{P(x,y)}. \tag{8.1}$$

Entropy is additive for independent random variables:

$$H(X,Y) = H(X) + H(Y) \ \text{ iff } \ P(x,y) = P(x)P(y). \tag{8.2}$$

The conditional entropy of X **given** $y\!=\!b_k$ is the entropy of the probability distribution $P(x \,|\, y\!=\!b_k)$.

$$H(X \,|\, y\!=\!b_k) \equiv \sum_{x \in \mathcal{A}_X} P(x \,|\, y\!=\!b_k) \log \frac{1}{P(x \,|\, y\!=\!b_k)}. \tag{8.3}$$

The conditional entropy of X **given** Y is the average, over y, of the conditional entropy of X given y.

$$\begin{aligned} H(X \,|\, Y) &\equiv \sum_{y \in \mathcal{A}_Y} P(y) \left[\sum_{x \in \mathcal{A}_X} P(x \,|\, y) \log \frac{1}{P(x \,|\, y)} \right] \\ &= \sum_{xy \in \mathcal{A}_X\mathcal{A}_Y} P(x,y) \log \frac{1}{P(x \,|\, y)}. \end{aligned} \tag{8.4}$$

This measures the average uncertainty that remains about x when y is known.

The marginal entropy of X is another name for the entropy of X, $H(X)$, used to contrast it with the conditional entropies listed above.

Chain rule for information content. From the product rule for probabilities, equation (2.6), we obtain:

$$\log \frac{1}{P(x,y)} \quad = \quad \log \frac{1}{P(x)} + \log \frac{1}{P(y \,|\, x)} \tag{8.5}$$

so

$$h(x,y) = h(x) + h(y \,|\, x). \tag{8.6}$$

In words, this says that the information content of x and y is the information content of x plus the information content of y given x.

Chain rule for entropy. The joint entropy, conditional entropy and marginal entropy are related by:

$$H(X,Y) = H(X) + H(Y \,|\, X) = H(Y) + H(X \,|\, Y). \tag{8.7}$$

In words, this says that the uncertainty of X and Y is the uncertainty of X plus the uncertainty of Y given X.

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

$$I(X;Y) \quad \equiv \quad H(X) - H(X \,|\, Y), \tag{8.8}$$

and satisfies $I(X;Y) = I(Y;X)$, and $I(X;Y) \geq 0$. It measures the average reduction in uncertainty about x that results from learning the value of y; **or vice versa**, the average amount of information that x conveys about y.

The conditional mutual information between X **and** Y **given** $z = c_k$ is the mutual information between the random variables X and Y in the joint ensemble $P(x, y \,|\, z = c_k)$,

$$I(X;Y \,|\, z = c_k) = H(X \,|\, z = c_k) - H(X \,|\, Y, z = c_k). \tag{8.9}$$

The conditional mutual information between X **and** Y **given** Z is the average over z of the above conditional mutual information.

$$I(X;Y \,|\, Z) = H(X \,|\, Z) - H(X \,|\, Y, Z). \tag{8.10}$$

No other 'three-term entropies' will be defined. For example, expressions such as $I(X;Y;Z)$ and $I(X \,|\, Y;Z)$ are illegal. But you may put conjunctions of arbitrary numbers of variables in each of the three spots in the expression $I(X;Y \,|\, Z)$ – for example, $I(A,B;C,D \,|\, E,F)$ is fine: it measures how much information on average c and d convey about a and b, assuming e and f are known.

Figure 8.1 shows how the total entropy $H(X,Y)$ of a joint ensemble can be broken down. **This figure is important.** $\ast$

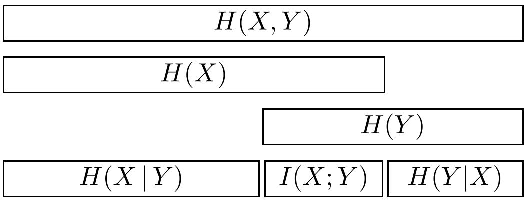


Figure 8.1. The relationship between joint information, marginal entropy, conditional entropy and mutual entropy.

▶ 8.2 Exercises

▷ Exercise 8.1.$^{[1]}$ Consider three independent random variables u, v, w with entropies H_u, H_v, H_w. Let $X \equiv (U,V)$ and $Y \equiv (V,W)$. What is $H(X,Y)$? What is $H(X\,|\,Y)$? What is $I(X;Y)$?

▷ Exercise 8.2.$^{[3,\text{ p.142}]}$ Referring to the definitions of conditional entropy (8.3–8.4), confirm (with an example) that it is possible for $H(X\,|\,y\!=\!b_k)$ to exceed $H(X)$, but that the average, $H(X\,|\,Y)$ is less than $H(X)$. So data are helpful – they do not increase uncertainty, on average.

▷ Exercise 8.3.$^{[2,\text{ p.143}]}$ Prove the chain rule for entropy, equation (8.7). $[H(X,Y) = H(X) + H(Y\,|\,X)]$.

Exercise 8.4.$^{[2,\text{ p.143}]}$ Prove that the mutual information $I(X;Y) \equiv H(X) - H(X\,|\,Y)$ satisfies $I(X;Y) = I(Y;X)$ and $I(X;Y) \geq 0$.

[Hint: see exercise 2.26 (p.37) and note that

$$I(X;Y) = D_{\mathrm{KL}}(P(x,y)||P(x)P(y)).] \tag{8.11}$$

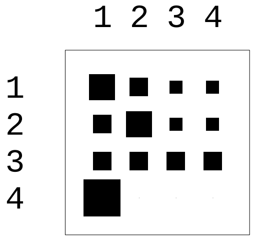
Exercise 8.5.$^{[4]}$ The 'entropy distance' between two random variables can be defined to be the difference between their joint entropy and their mutual information:

$$D_H(X,Y) \equiv H(X,Y) - I(X;Y). \tag{8.12}$$

Prove that the entropy distance satisfies the axioms for a distance – $D_H(X,Y) \geq 0$, $D_H(X,X)\!=\!0$, $D_H(X,Y)\!=\!D_H(Y,X)$, and $D_H(X,Z) \leq D_H(X,Y) + D_H(Y,Z)$. [Incidentally, we are unlikely to see $D_H(X,Y)$ again but it is a good function on which to practise inequality-proving.]

Exercise 8.6.$^{[2]}$ A joint ensemble XY has the following joint distribution.

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



What is the joint entropy $H(X,Y)$? What are the marginal entropies $H(X)$ and $H(Y)$? For each value of y, what is the conditional entropy $H(X\,|\,y)$? What is the conditional entropy $H(X\,|\,Y)$? What is the conditional entropy of Y given X? What is the mutual information between X and Y?

Exercise 8.7.[*2*, p.143] Consider the ensemble XYZ in which $\mathcal{A}_X = \mathcal{A}_Y = \mathcal{A}_Z = \{0, 1\}$, x and y are independent with $\mathcal{P}_X = \{p, 1 - p\}$ and $\mathcal{P}_Y = \{q, 1 - q\}$ and

$$z = (x + y) \bmod 2. \tag{8.13}$$

(a) If $q = 1/2$, what is $\mathcal{P}_Z$? What is $I(Z; X)$?

(b) For general p and q, what is $\mathcal{P}_Z$? What is $I(Z; X)$? Notice that this ensemble is related to the binary symmetric channel, with $x =$ input, $y =$ noise, and $z =$ output.

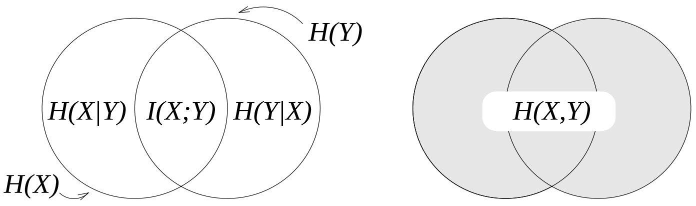


Figure 8.2. A misleading representation of entropies (contrast with figure 8.1).

Three term entropies

Exercise 8.8.[*3*, p.143] Many texts draw figure 8.1 in the form of a Venn diagram (figure 8.2). Discuss why this diagram is a misleading representation of entropies. Hint: consider the three-variable ensemble XYZ in which $x \in \{0, 1\}$ and $y \in \{0, 1\}$ are independent binary variables and $z \in \{0, 1\}$ is defined to be $z = x + y \bmod 2$.

▶ 8.3 Further exercises

The data-processing theorem

The data processing theorem states that data processing can only destroy information.

Exercise 8.9.[*3*, p.144] Prove this theorem by considering an ensemble WDR in which w is the state of the world, d is data gathered, and r is the processed data, so that these three variables form a *Markov chain*

$$w \rightarrow d \rightarrow r, \tag{8.14}$$

that is, the probability $P(w, d, r)$ can be written as

$$P(w, d, r) = P(w)P(d \mid w)P(r \mid d). \tag{8.15}$$

Show that the average information that R conveys about W, $I(W; R)$, is less than or equal to the average information that D conveys about W, $I(W; D)$.

This theorem is as much a caution about our definition of 'information' as it is a caution about data processing!

Inference and information measures

Exercise 8.10.$^{[2]}$ The three cards.

(a) One card is white on both faces; one is black on both faces; and one is white on one side and black on the other. The three cards are shuffled and their orientations randomized. One card is drawn and placed on the table. The upper face is black. What is the colour of its lower face? (Solve the inference problem.)

(b) Does seeing the top face convey *information* about the colour of the bottom face? Discuss the *information contents* and *entropies* in this situation. Let the value of the upper face's colour be u and the value of the lower face's colour be l. Imagine that we draw a random card and learn both u and l. What is the entropy of u, $H(U)$? What is the entropy of l, $H(L)$? What is the mutual information between U and L, $I(U;L)$?

Entropies of Markov processes

▷ Exercise 8.11.$^{[3]}$ In the guessing game, we imagined predicting the next letter in a document starting from the beginning and working towards the end. Consider the task of predicting the *reversed* text, that is, predicting the letter that precedes those already known. Most people find this a harder task. Assuming that we model the language using an N-gram model (which says the probability of the next character depends only on the $N-1$ preceding characters), is there any difference between the average information contents of the reversed language and the forward language?

▶ 8.4 Solutions

Solution to exercise 8.2 (p.140). See exercise 8.6 (p.140) for an example where $H(X\,|\,y)$ exceeds $H(X)$ (set $y=3$).

We can prove the inequality $H(X\,|\,Y) \leq H(X)$ by turning the expression into a relative entropy (using Bayes' theorem) and invoking Gibbs' inequality (exercise 2.26 (p.37)):

$$\begin{aligned}
H(X\,|\,Y) &\equiv \sum_{y\in\mathcal{A}_Y} P(y)\left[\sum_{x\in\mathcal{A}_X} P(x\,|\,y)\log\frac{1}{P(x\,|\,y)}\right] \\
&= \sum_{xy\in\mathcal{A}_X\mathcal{A}_Y} P(x,y)\log\frac{1}{P(x\,|\,y)} && (8.16)\\
&= \sum_{xy} P(x)P(y\,|\,x)\log\frac{P(y)}{P(y\,|\,x)P(x)} && (8.17)\\
&= \sum_{x} P(x)\log\frac{1}{P(x)} + \sum_{x} P(x)\sum_{y} P(y\,|\,x)\log\frac{P(y)}{P(y\,|\,x)} && (8.18)
\end{aligned}$$

The last expression is a sum of relative entropies between the distributions $P(y\,|\,x)$ and $P(y)$. So

$$H(X\,|\,Y) \leq H(X) + 0, \tag{8.19}$$

with equality only if $P(y\,|\,x) = P(y)$ for all x and y (that is, only if X and Y are independent).

Solution to exercise 8.3 (p.140). The chain rule for entropy follows from the decomposition of a joint probability:

$$\begin{aligned} H(X,Y) &= \sum_{xy} P(x,y) \log \frac{1}{P(x,y)} && (8.20)\\ &= \sum_{xy} P(x)P(y \,|\, x) \left[\log \frac{1}{P(x)} + \log \frac{1}{P(y \,|\, x)} \right] && (8.21)\\ &= \sum_{x} P(x) \log \frac{1}{P(x)} + \sum_x P(x) \sum_y P(y \,|\, x) \log \frac{1}{P(y \,|\, x)} && (8.22)\\ &= H(X) + H(Y \,|\, X). && (8.23) \end{aligned}$$

Solution to exercise 8.4 (p.140). Symmetry of mutual information:

$$\begin{aligned} I(X;Y) &= H(X) - H(X \,|\, Y) && (8.24)\\ &= \sum_x P(x) \log \frac{1}{P(x)} - \sum_{xy} P(x,y) \log \frac{1}{P(x \,|\, y)} && (8.25)\\ &= \sum_{xy} P(x,y) \log \frac{P(x \,|\, y)}{P(x)} && (8.26)\\ &= \sum_{xy} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}. && (8.27) \end{aligned}$$

This expression is symmetric in x and y so

$$I(X;Y) = H(X) - H(X \,|\, Y) = H(Y) - H(Y \,|\, X). \qquad (8.28)$$

We can prove that mutual information is positive two ways. One is to continue from

$$I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \qquad (8.29)$$

which is a relative entropy and use Gibbs' inequality (proved on p.44), which asserts that this relative entropy is ≥ 0, with equality only if $P(x,y) = P(x)P(y)$, that is, if X and Y are independent.

The other is to use Jensen's inequality on

$$-\sum_{x,y} P(x,y) \log \frac{P(x)P(y)}{P(x,y)} \geq -\log \sum_{x,y} \frac{P(x,y)}{P(x,y)} P(x)P(y) = \log 1 = 0. \qquad (8.30)$$

Solution to exercise 8.7 (p.141). $z = x + y \bmod 2$.

(a) If $q = 1/2$, $\mathcal{P}_Z = \{1/2, 1/2\}$ and $I(Z;X) = H(Z) - H(Z \,|\, X) = 1 - 1 = 0$.

(b) For general q and p, $\mathcal{P}_Z = \{pq + (1-p)(1-q), p(1-q) + q(1-p)\}$. The mutual information is $I(Z;X) = H(Z) - H(Z \,|\, X) = H_2(pq + (1-p)(1-q)) - H_2(q)$.

Three term entropies

Solution to exercise 8.8 (p.141). The depiction of entropies in terms of Venn diagrams is misleading for at least two reasons.

First, one is used to thinking of Venn diagrams as depicting sets; but what are the 'sets' $H(X)$ and $H(Y)$ depicted in figure 8.2, and what are the objects that are members of those sets? I think this diagram encourages the novice student to make inappropriate analogies. For example, some students imagine

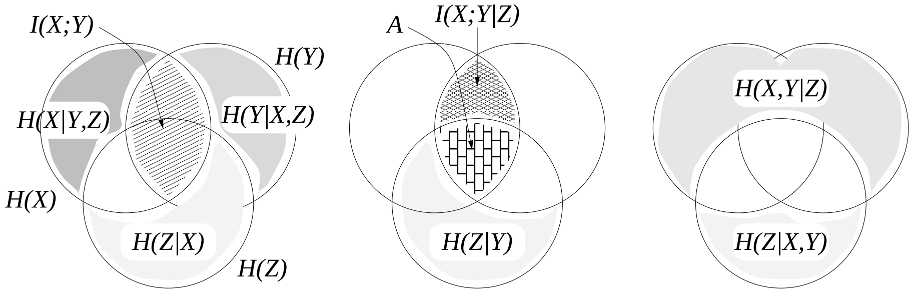


Figure 8.3. A misleading representation of entropies, continued.

that the random outcome (x, y) might correspond to a point in the diagram, and thus confuse entropies with probabilities.

Secondly, the depiction in terms of Venn diagrams encourages one to believe that all the areas correspond to positive quantities. In the special case of two random variables it is indeed true that $H(X \mid Y)$, $I(X;Y)$ and $H(Y \mid X)$ are positive quantities. But as soon as we progress to three-variable ensembles, we obtain a diagram with positive-looking areas that may actually correspond to negative quantities. Figure 8.3 correctly shows relationships such as

$$H(X) + H(Z \mid X) + H(Y \mid X, Z) = H(X, Y, Z). \tag{8.31}$$

But it gives the misleading impression that the conditional mutual information $I(X;Y \mid Z)$ is *less than* the mutual information $I(X;Y)$. In fact the area labelled A can correspond to a *negative* quantity. Consider the joint ensemble (X, Y, Z) in which $x \in \{0, 1\}$ and $y \in \{0, 1\}$ are independent binary variables and $z \in \{0, 1\}$ is defined to be $z = x + y \bmod 2$. Then clearly $H(X) = H(Y) = 1$ bit. Also $H(Z) = 1$ bit. And $H(Y \mid X) = H(Y) = 1$ since the two variables are independent. So the mutual information between X and Y is zero. $I(X;Y) = 0$. However, if z is observed, X and Y become correlated — knowing x, given z, tells you what y is: $y = z - x \bmod 2$. So $I(X;Y \mid Z) = 1$ bit. Thus the area labelled A must correspond to -1 bits for the figure to give the correct answers.

The above example is not at all a capricious or exceptional illustration. The binary symmetric channel with input X, noise Y, and output Z is a situation in which $I(X;Y) = 0$ (input and noise are uncorrelated) but $I(X;Y \mid Z) > 0$ (once you see the output, the unknown input and the unknown noise are intimately related!).

The Venn diagram representation is therefore valid only if one is aware that positive areas may represent negative quantities. With this proviso kept in mind, the interpretation of entropies in terms of sets can be helpful (Yeung, 1991).

Solution to exercise 8.9 (p.141). For any joint ensemble XYZ, the following chain rule for mutual information holds.

$$I(X;Y,Z) = I(X;Y) + I(X;Z \mid Y). \tag{8.32}$$

Now, in the case $w \rightarrow d \rightarrow r$, w and r are independent given d, so $I(W;R \mid D) = 0$. Using the chain rule twice, we have:

$$I(W;D,R) = I(W;D) \tag{8.33}$$

and

$$I(W;D,R) = I(W;R) + I(W;D \mid R), \tag{8.34}$$

so

$$I(W;R) - I(W;D) \leq 0. \tag{8.35}$$

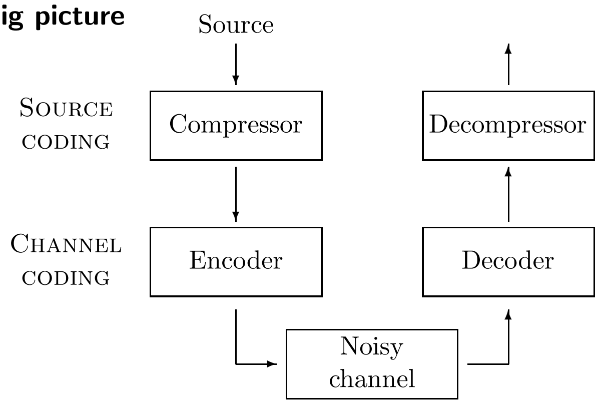
About Chapter 9

Before reading Chapter 9, you should have read Chapter 1 and worked on exercise 2.26 (p.37), and exercises 8.2–8.7 (pp.140–141).

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:

$P(x\,|\,y)$		x				$H(X\,|\,y)$/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
						$H(X\,|\,Y) = 11/8$

Note that whereas $H(X\,|\,y\!=\!4) = 0$ is less than $H(X)$, $H(X\,|\,y\!=\!3)$ is greater than $H(X)$. So in some cases, learning y can *increase* our uncertainty about x. Note also that although $P(x\,|\,y\!=\!2)$ is a different distribution from $P(x)$, the conditional entropy $H(X\,|\,y\!=\!2)$ 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 $H(X\,|\,Y) < H(X)$.

One may also evaluate $H(Y|X) = 13/8$ bits. The mutual information is $I(X;Y) = H(X) - H(X\,|\,Y) = 3/8$ 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 $P(y\,|\,x)$, one for each $x \in \mathcal{A}_X$.

These *transition probabilities* may be written in a matrix

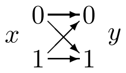
$$Q_{j|i} = P(y\!=\!b_j\,|\,x\!=\!a_i). \tag{9.1}$$

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:

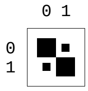
$$\mathbf{p}_Y = \mathbf{Q}\mathbf{p}_X. \tag{9.2}$$

Some useful model channels are:

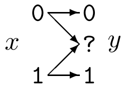
Binary symmetric channel. $\mathcal{A}_X = \{\mathtt{0},\mathtt{1}\}$. $\mathcal{A}_Y = \{\mathtt{0},\mathtt{1}\}$.



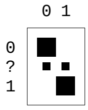
$$\begin{array}{rclrcl} P(y=\mathtt{0}\,|\,x=\mathtt{0}) &=& 1-f; & P(y=\mathtt{0}\,|\,x=\mathtt{1}) &=& f; \\ P(y=\mathtt{1}\,|\,x=\mathtt{0}) &=& f; & P(y=\mathtt{1}\,|\,x=\mathtt{1}) &=& 1-f. \end{array}$$



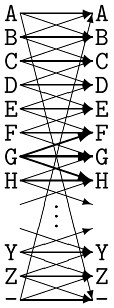
Binary erasure channel. $\mathcal{A}_X = \{\mathtt{0},\mathtt{1}\}$. $\mathcal{A}_Y = \{\mathtt{0},\mathtt{?},\mathtt{1}\}$.



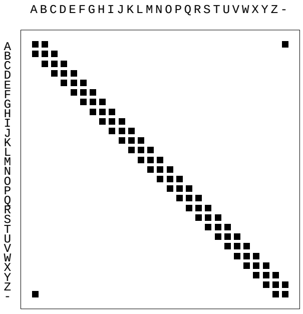
$$\begin{array}{rclrcl} P(y=\mathtt{0}\,|\,x=\mathtt{0}) &=& 1-f; & P(y=\mathtt{0}\,|\,x=\mathtt{1}) &=& 0; \\ P(y=\mathtt{?}\,|\,x=\mathtt{0}) &=& f; & P(y=\mathtt{?}\,|\,x=\mathtt{1}) &=& f; \\ P(y=\mathtt{1}\,|\,x=\mathtt{0}) &=& 0; & P(y=\mathtt{1}\,|\,x=\mathtt{1}) &=& 1-f. \end{array}$$



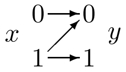
Noisy typewriter. $\mathcal{A}_X = \mathcal{A}_Y =$ the 27 letters $\{\mathtt{A}, \mathtt{B}, \ldots, \mathtt{Z}, \mathtt{-}\}$. The letters are arranged in a circle, and when the typist attempts to type B, what comes out is either A, B or C, with probability 1/3 each; when the input is C, the output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



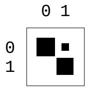
$$\begin{array}{rcl} & \vdots & \\ P(y=\mathtt{F}\,|\,x=\mathtt{G}) &=& 1/3; \\ P(y=\mathtt{G}\,|\,x=\mathtt{G}) &=& 1/3; \\ P(y=\mathtt{H}\,|\,x=\mathtt{G}) &=& 1/3; \\ & \vdots & \end{array}$$



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



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



▶ 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 $y=1$.

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

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

Exercise 9.2.$^{[\mathit{1},\ \text{p.157}]}$ 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.$^{[\mathit{1},\ \text{p.157}]}$ 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.

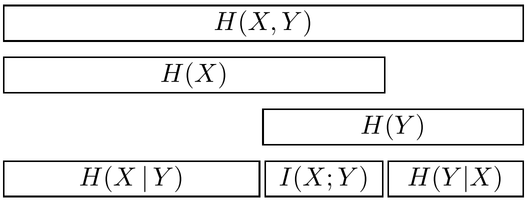


Figure 9.1. The relationship between joint information, marginal entropy, conditional entropy and mutual entropy. This figure is important, so I'm showing it twice.

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 $H_2(p) \equiv H(p, 1-p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{(1-p)}$.

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

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

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 $I(X;Y)$.]

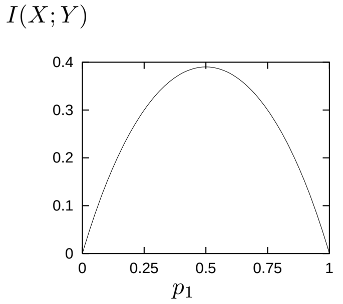


Figure 9.2. The mutual information $I(X;Y)$ for a binary symmetric channel with $f = 0.15$ as a function of the input distribution.

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

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

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 $I(X;Y)$,

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

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:

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

Then the mutual information is:

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

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 that 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.

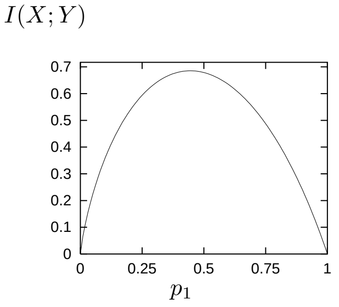


Figure 9.3. The mutual information $I(X;Y)$ for a Z channel with $f = 0.15$ as a function of the input distribution.

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 $P(s_{\rm in})$, is:

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

The maximal probability of block error is

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

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 $P(s \,|\, \mathbf{y})$.

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

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

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 $p_{\rm b}$ 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 $\mathbf{s}_{\rm out}$ is not equal to the corresponding bit of $\mathbf{s}_{\rm in}$ (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 $< \epsilon$.

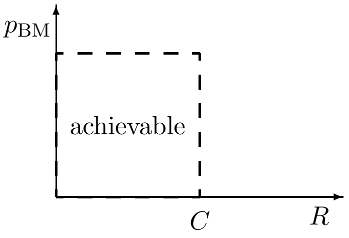


Figure 9.4. Portion of the $R, p_{\rm BM}$ plane asserted to be achievable by the first part of Shannon's noisy channel coding theorem.

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).

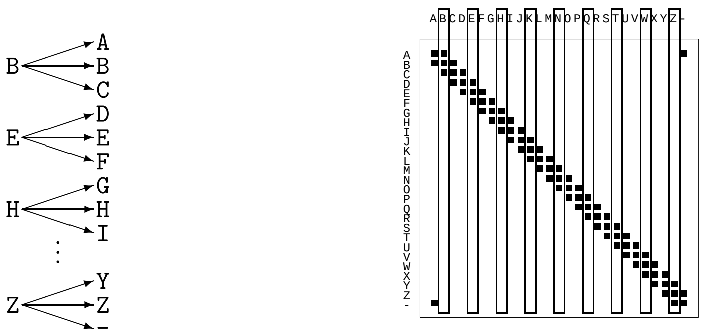


Figure 9.5. A non-confusable subset of inputs for the noisy typewriter.

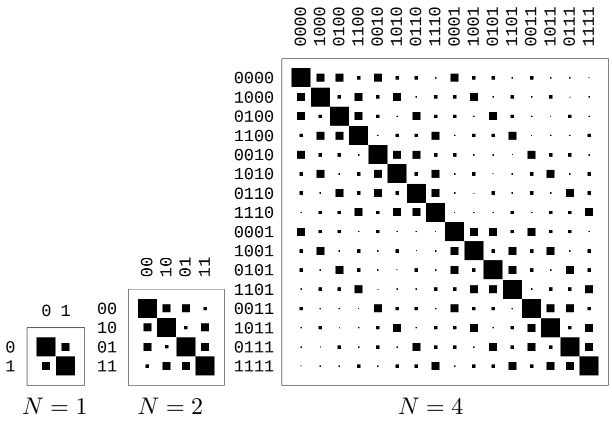


Figure 9.6. Extended channels obtained from a binary symmetric channel with transition probability 0.15.

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 block length 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 a given channel, we consider the *extended channel* corresponding to N uses of the given 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 $N = 4$.

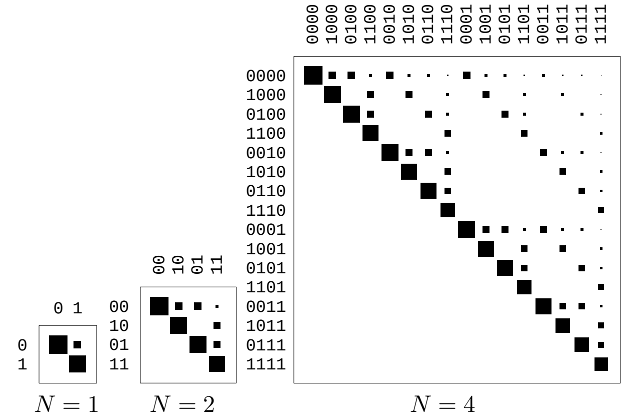


Figure 9.7. Extended channels obtained from a Z channel with transition probability 0.15. Each column corresponds to an input, and each row is a different output.

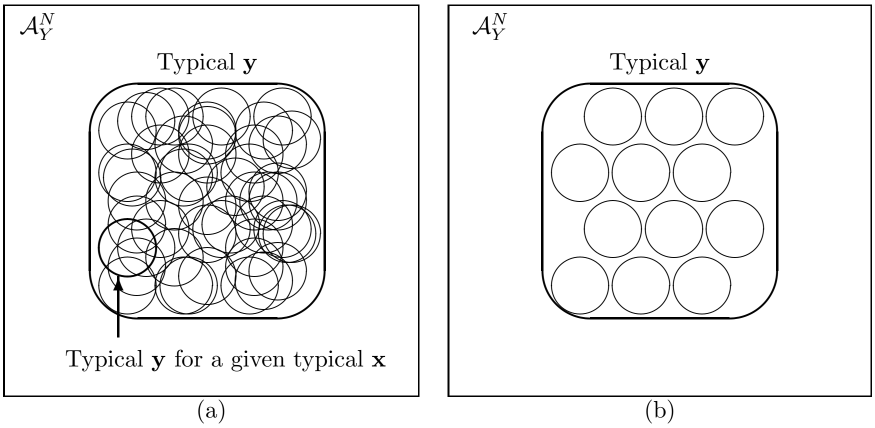


Figure 9.8. (a) Some typical outputs in $\mathcal{A}_Y^N$ corresponding to typical inputs $\mathbf{x}$. (b) A subset of the typical sets shown in (a) that do not overlap each other. This picture can be compared with the solution to the noisy typewriter in figure 9.5.

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 block lengths 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

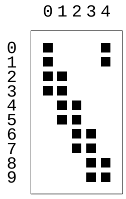
$$p_1^* = \frac{1/(1-f)}{1 + 2^{(H_2(f)/(1-f))}}. \tag{9.19}$$

(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.

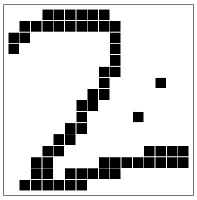
(b) Assume that a single bit is to be transmitted. What is the optimal decoder, and what is its probability of error? Express your answer in terms of the signal to noise ratio α^2/σ^2 and the error function (the cumulative probability function of the Gaussian distribution),

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

[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.

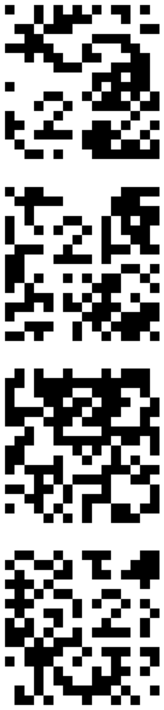


Figure 9.9. Some more 2s.

Random coding

Exercise 9.20.$^{[2, \text{ p.160}]}$ 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