About Chapter 10

Before reading Chapter 10, you should have read Chapters 4 and 9. Exercise 9.14 (p.153) is especially recommended.

Cast of characters

Q	the noisy channel
C	the capacity of the channel
X^N	an ensemble used to create a random code
$\mathcal{C}$	a random code
N	the length of the codewords
$\mathbf{x}^{(s)}$	a codeword, the sth in the code
s	the number of a chosen codeword (mnemonic: the *source* selects s)
$S = 2^K$	the total number of codewords in the code
$K = \log_2 S$	the number of bits conveyed by the choice of one codeword from S, assuming it is chosen with uniform probability
$\mathbf{s}$	a binary representation of the number s
$R = K/N$	the rate of the code, in bits per channel use (sometimes called R' instead)
$\hat{s}$	the decoder's guess of s

10

The Noisy-Channel Coding Theorem

10.1 The theorem

The theorem has three parts, two positive and one negative. The main positive result is the first.

1. For every discrete memoryless channel, the channel capacity
$$C = \max_{\mathcal{P}_X} I(X;Y) \tag{10.1}$$
has the following property. For any $\epsilon > 0$ and $R < C$, for large enough N, there exists a code of length N and rate $\geq R$ and a decoding algorithm, such that the maximal probability of block error is $< \epsilon$.

2. If a probability of bit error $p_{\rm b}$ is acceptable, rates up to $R(p_{\rm b})$ are achievable, where
$$R(p_{\rm b}) = \frac{C}{1 - H_2(p_{\rm b})}. \tag{10.2}$$

3. For any $p_{\rm b}$, rates greater than $R(p_{\rm b})$ are not achievable.

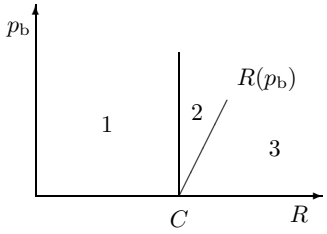


Figure 10.1. Portion of the $R, p_{\rm b}$ plane to be proved achievable (1, 2) and not achievable (3).

10.2 Jointly-typical sequences

We formalize the intuitive preview of the last chapter.

We will define codewords $\mathbf{x}^{(s)}$ as coming from an ensemble X^N, and consider the random selection of one codeword and a corresponding channel output $\mathbf{y}$, thus defining a joint ensemble $(XY)^N$. We will use a *typical-set decoder*, which decodes a received signal $\mathbf{y}$ as s if $\mathbf{x}^{(s)}$ and $\mathbf{y}$ are *jointly typical*, a term to be defined shortly.

The proof will then centre on determining the probabilities (a) that the true input codeword is *not* jointly typical with the output sequence; and (b) that a *false* input codeword is jointly typical with the output. We will show that, for large N, both probabilities go to zero as long as there are fewer than 2^{NC} codewords, and the ensemble X is the optimal input distribution.

Joint typicality. A pair of sequences $\mathbf{x}, \mathbf{y}$ of length N are defined to be jointly typical (to tolerance β) with respect to the distribution $P(x, y)$ if

$$\mathbf{x} \text{ is typical of } P(\mathbf{x}), \quad \text{i.e.,} \quad \left| \frac{1}{N} \log \frac{1}{P(\mathbf{x})} - H(X) \right| < \beta,$$
$$\mathbf{y} \text{ is typical of } P(\mathbf{y}), \quad \text{i.e.,} \quad \left| \frac{1}{N} \log \frac{1}{P(\mathbf{y})} - H(Y) \right| < \beta,$$
$$\text{and } \mathbf{x}, \mathbf{y} \text{ is typical of } P(\mathbf{x}, \mathbf{y}), \quad \text{i.e.,} \quad \left| \frac{1}{N} \log \frac{1}{P(\mathbf{x}, \mathbf{y})} - H(X, Y) \right| < \beta.$$

The jointly-typical set $J_{N\beta}$ is the set of all jointly-typical sequence pairs of length N.

Example. Here is a jointly-typical pair of length $N = 100$ for the ensemble $P(x, y)$ in which $P(x)$ has $(p_0, p_1) = (0.9, 0.1)$ and $P(y \mid x)$ corresponds to a binary symmetric channel with noise level 0.2.

```
x  1111111111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
y  0011111111100000000000000000000000000000000000000000000000000000000000000000000000000000011111111111111111111
```

Notice that $\mathbf{x}$ has 10 1s, and so is typical of the probability $P(\mathbf{x})$ (at any tolerance β); and $\mathbf{y}$ has 26 1s, so it is typical of $P(\mathbf{y})$ (because $P(y=1) = 0.26$); and $\mathbf{x}$ and $\mathbf{y}$ differ in 20 bits, which is the typical number of flips for this channel.

Joint typicality theorem. Let $\mathbf{x}, \mathbf{y}$ be drawn from the ensemble $(XY)^N$ defined by

$$P(\mathbf{x}, \mathbf{y}) = \prod_{n=1}^{N} P(x_n, y_n).$$

Then

1. the probability that $\mathbf{x}, \mathbf{y}$ are jointly typical (to tolerance β) tends to 1 as $N \to \infty$;
2. the number of jointly-typical sequences $|J_{N\beta}|$ is close to $2^{NH(X,Y)}$. To be precise,
$$|J_{N\beta}| \leq 2^{N(H(X,Y)+\beta)}; \tag{10.3}$$
3. if $\mathbf{x}' \sim X^N$ and $\mathbf{y}' \sim Y^N$, i.e., $\mathbf{x}'$ and $\mathbf{y}'$ are *independent* samples with the same marginal distribution as $P(\mathbf{x}, \mathbf{y})$, then the probability that $(\mathbf{x}', \mathbf{y}')$ lands in the jointly-typical set is about $2^{-NI(X;Y)}$. To be precise,
$$P((\mathbf{x}', \mathbf{y}') \in J_{N\beta}) \leq 2^{-N(I(X;Y)-3\beta)}. \tag{10.4}$$

Proof. The proof of parts 1 and 2 by the law of large numbers follows that of the source coding theorem in Chapter 4. For part 2, let the pair x, y play the role of x in the source coding theorem, replacing $P(x)$ there by the probability distribution $P(x, y)$.

For the third part,

$$\begin{aligned} P((\mathbf{x}', \mathbf{y}') \in J_{N\beta}) &= \sum_{(\mathbf{x},\mathbf{y}) \in J_{N\beta}} P(\mathbf{x})P(\mathbf{y}) && (10.5)\\ &\leq |J_{N\beta}|\, 2^{-N(H(X)-\beta)}\, 2^{-N(H(Y)-\beta)} && (10.6)\\ &\leq 2^{N(H(X,Y)+\beta)-N(H(X)+H(Y)-2\beta)} && (10.7)\\ &= 2^{-N(I(X;Y)-3\beta)}. \quad \square && (10.8) \end{aligned}$$

A cartoon of the jointly-typical set is shown in figure 10.2. Two independent typical vectors are jointly typical with probability

$$P((\mathbf{x}', \mathbf{y}') \in J_{N\beta}) \simeq 2^{-N(I(X;Y))} \tag{10.9}$$

because the *total* number of independent typical pairs is the area of the dashed rectangle, $2^{NH(X)}2^{NH(Y)}$, and the number of jointly-typical pairs is roughly $2^{NH(X,Y)}$, so the probability of hitting a jointly-typical pair is roughly

$$2^{NH(X,Y)}/2^{NH(X)+NH(Y)} = 2^{-NI(X;Y)}. \tag{10.10}$$

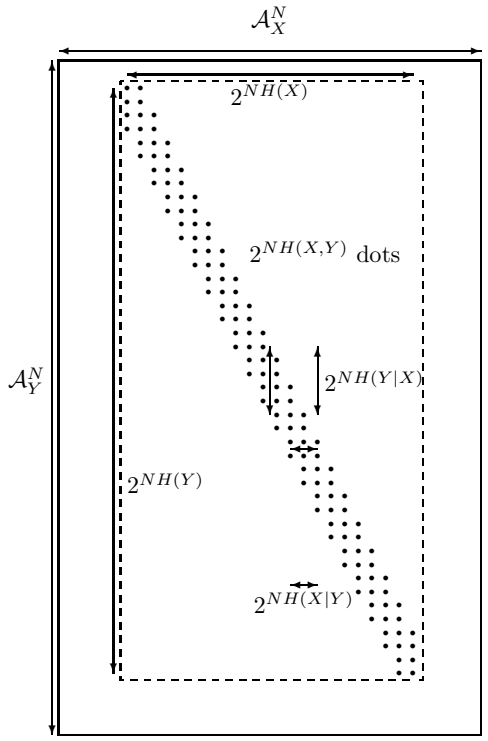


Figure 10.2. The jointly-typical set. The horizontal direction represents $\mathcal{A}_X^N$, the set of all input strings of length N. The vertical direction represents $\mathcal{A}_Y^N$, the set of all output strings of length N. The outer box contains all conceivable input–output pairs. Each dot represents a jointly-typical pair of sequences $(\mathbf{x}, \mathbf{y})$. The total number of jointly-typical sequences is about $2^{NH(X,Y)}$.

▶ 10.3 Proof of the noisy-channel coding theorem

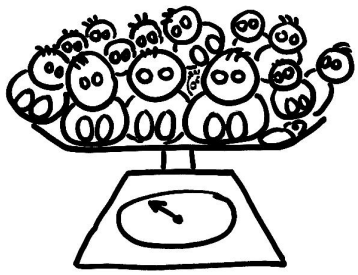


Figure 10.3. Shannon's method for proving one baby weighs less than 10 kg.

Analogy

Imagine that we wish to prove that there is a baby in a class of one hundred babies who weighs less than 10 kg. Individual babies are difficult to catch and weigh. Shannon's method of solving the task is to scoop up all the babies and weigh them all at once on a big weighing machine. If we find that their *average* weight is smaller than 10 kg, there must exist *at least one* baby who weighs less than 10 kg – indeed there must be many! Shannon's method isn't guaranteed to reveal the existence of an underweight child, since it relies on there being a tiny number of elephants in the class. But if we use his method and get a total weight smaller than 1000 kg then our task is solved.

From skinny children to fantastic codes

We wish to show that there exists a code and a decoder having small probability of error. Evaluating the probability of error of any particular coding and decoding system is not easy. Shannon's innovation was this: instead of constructing a good coding and decoding system and evaluating its error probability, Shannon calculated the average probability of block error of *all* codes, and proved that this average is small. There must then exist individual codes that have small probability of block error.

Random coding and typical-set decoding

Consider the following encoding–decoding system, whose rate is R'.

1. We fix $P(x)$ and generate the $S = 2^{NR'}$ codewords of a $(N, NR') =$

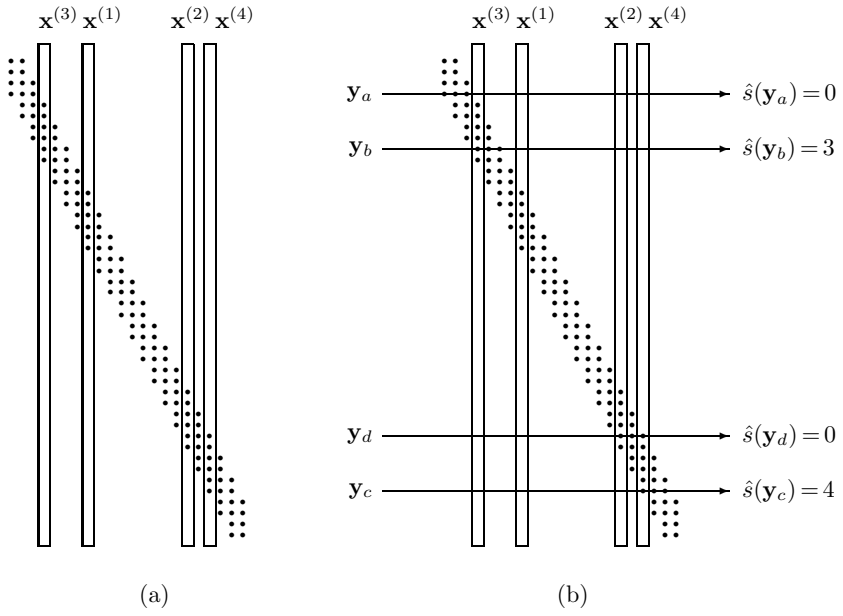


Figure 10.4. (a) A random code. (b) Example decodings by the typical set decoder. A sequence that is not jointly typical with any of the codewords, such as $\mathbf{y}_a$, is decoded as $\hat{s} = 0$. A sequence that is jointly typical with codeword $\mathbf{x}^{(3)}$ alone, $\mathbf{y}_b$, is decoded as $\hat{s} = 3$. Similarly, $\mathbf{y}_c$ is decoded as $\hat{s} = 4$. A sequence that is jointly typical with more than one codeword, such as $\mathbf{y}_d$, is decoded as $\hat{s} = 0$.

(N, K) code $\mathcal{C}$ at random according to

$$P(\mathbf{x}) = \prod_{n=1}^{N} P(x_n). \tag{10.11}$$

A random code is shown schematically in figure 10.4a.

2. The code is known to both sender and receiver.

3. A message s is chosen from $\{1, 2, \ldots, 2^{NR'}\}$, and $\mathbf{x}^{(s)}$ is transmitted. The received signal is $\mathbf{y}$, with

$$P(\mathbf{y} \,|\, \mathbf{x}^{(s)}) = \prod_{n=1}^{N} P(y_n \,|\, x_n^{(s)}). \tag{10.12}$$

4. The signal is decoded by *typical-set decoding*.

   Typical-set decoding. Decode $\mathbf{y}$ as $\hat{s}$ if $(\mathbf{x}^{(\hat{s})}, \mathbf{y})$ are jointly typical *and* there is no other s' such that $(\mathbf{x}^{(s')}, \mathbf{y})$ are jointly typical; otherwise declare a failure ($\hat{s} = 0$).

   This is not the optimal decoding algorithm, but it will be good enough, and easier to analyze. The typical-set decoder is illustrated in figure 10.4b.

5. A decoding error occurs if $\hat{s} \neq s$.

There are three probabilities of error that we can distinguish. First, there is the probability of block error for a particular code $\mathcal{C}$, that is,

$$p_{\mathrm{B}}(\mathcal{C}) \equiv P(\hat{s} \neq s \,|\, \mathcal{C}). \tag{10.13}$$

This is a difficult quantity to evaluate for any given code.

Second, there is the average over all codes of this block error probability,

$$\langle p_{\mathrm{B}} \rangle \equiv \sum_{\mathcal{C}} P(\hat{s} \neq s \,|\, \mathcal{C}) P(\mathcal{C}). \tag{10.14}$$

Fortunately, this quantity is much easier to evaluate than the first quantity $P(\hat{s} \neq s \,|\, \mathcal{C})$.

$\langle p_{\rm B} \rangle$ is just the probability that there is a decoding error at step 5 of the five-step process on the previous page.

Third, the maximal block error probability of a code $\mathcal{C}$,

$$p_{\rm BM}(\mathcal{C}) \equiv \max_s P(\hat{s} \neq s \,|\, s, \mathcal{C}), \tag{10.15}$$

is the quantity we are most interested in: we wish to show that there exists a code $\mathcal{C}$ with the required rate whose maximal block error probability is small.

We will get to this result by first finding the average block error probability, $\langle p_{\rm B} \rangle$. Once we have shown that this can be made smaller than a desired small number, we immediately deduce that there must exist *at least one* code $\mathcal{C}$ whose block error probability is also less than this small number. Finally, we show that this code, whose block error probability is satisfactorily small but whose maximal block error probability is unknown (and could conceivably be enormous), can be modified to make a code of slightly smaller rate whose maximal block error probability is also guaranteed to be small. We modify the code by throwing away the worst 50% of its codewords.

We therefore now embark on finding the average probability of block error.

Probability of error of typical-set decoder

There are two sources of error when we use typical-set decoding. Either (a) the output $\mathbf{y}$ is not jointly typical with the transmitted codeword $\mathbf{x}^{(s)}$, or (b) there is some other codeword in $\mathcal{C}$ that is jointly typical with $\mathbf{y}$.

By the symmetry of the code construction, the average probability of error averaged over all codes does not depend on the selected value of s; we can assume without loss of generality that $s = 1$.

(a) The probability that the input $\mathbf{x}^{(1)}$ and the output $\mathbf{y}$ are not jointly typical vanishes, by the joint typicality theorem's first part (p.163). We give a name, δ, to the upper bound on this probability, satisfying $\delta \to 0$ as $N \to \infty$; for any desired δ, we can find a blocklength $N(\delta)$ such that the $P((\mathbf{x}^{(1)}, \mathbf{y}) \notin J_{N\beta}) \leq \delta$.

(b) The probability that $\mathbf{x}^{(s')}$ and $\mathbf{y}$ are jointly typical, for a *given* $s' \neq 1$ is $\leq 2^{-N(I(X;Y)-3\beta)}$, by part 3. And there are $(2^{NR'} - 1)$ rival values of s' to worry about.

Thus the average probability of error $\langle p_{\rm B} \rangle$ satisfies:

$$\begin{aligned} \langle p_{\rm B} \rangle &\leq \delta + \sum_{s'=2}^{2^{NR'}} 2^{-N(I(X;Y)-3\beta)} && (10.16) \\ &\leq \delta + 2^{-N(I(X;Y)-R'-3\beta)}. && (10.17) \end{aligned}$$

The inequality (10.16) that bounds a total probability of error $P_{\rm TOT}$ by the sum of the probabilities $P_{s'}$ of all sorts of events s' each of which is sufficient to cause error,

$$P_{\rm TOT} \leq P_1 + P_2 + \cdots,$$

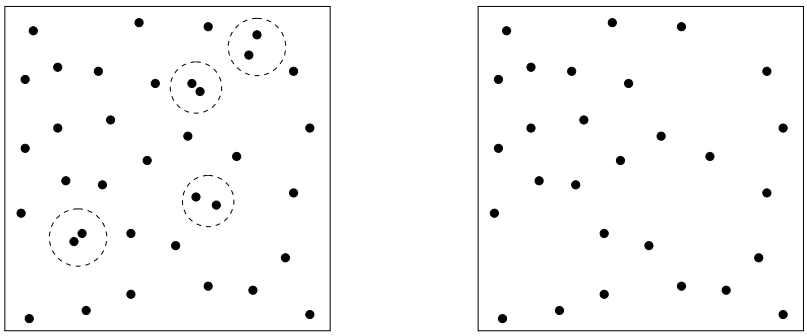
is called a *union bound*. It is only an equality if the different events that cause error never occur at the same time as each other.

The average probability of error (10.17) can be made $< 2\delta$ by increasing N if

$$R' < I(X;Y) - 3\beta. \tag{10.18}$$

We are almost there. We make three modifications:

1. We choose $P(x)$ in the proof to be the optimal input distribution of the channel. Then the condition $R' < I(X;Y) - 3\beta$ becomes $R' < C - 3\beta$.



(a) A random code ... ⇒ (b) expurgated

Figure 10.5. How expurgation works. (a) In a typical random code, a small fraction of the codewords are involved in collisions – pairs of codewords are sufficiently close to each other that the probability of error when either codeword is transmitted is not tiny. We obtain a new code from a random code by deleting all these confusable codewords. (b) The resulting code has slightly fewer codewords, so has a slightly lower rate, and its maximal probability of error is greatly reduced.

2. Since the average probability of error over all codes is $< 2\delta$, there must exist *a* code with mean probability of block error $p_{\rm B}(\mathcal{C}) < 2\delta$.

3. To show that not only the average but also the maximal probability of error, $p_{\rm BM}$, can be made small, we modify this code by throwing away the worst half of the codewords – the ones most likely to produce errors. Those that remain must all have *conditional* probability of error less than 4δ. We use these remaining codewords to define a new code. This new code has $2^{NR'-1}$ codewords, i.e., we have reduced the rate from R' to $R' - 1/N$ (a negligible reduction, if N is large), and achieved $p_{\rm BM} < 4\delta$. This trick is called *expurgation* (figure 10.5). The resulting code may not be the best code of its rate and length, but it is still good enough to prove the noisy-channel coding theorem, which is what we are trying to do here.

In conclusion, we can 'construct' a code of rate $R' - 1/N$, where $R' < C - 3\beta$, with maximal probability of error $< 4\delta$. We obtain the theorem as stated by setting $R' = (R + C)/2$, $\delta = \epsilon/4$, $\beta < (C - R')/3$, and N sufficiently large for the remaining conditions to hold. The theorem's first part is thus proved. □

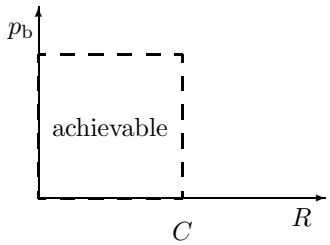


Figure 10.6. Portion of the $R, p_{\rm b}$ plane proved achievable in the first part of the theorem. [We've proved that the maximal probability of block error $p_{\rm BM}$ can be made arbitrarily small, so the same goes for the bit error probability $p_{\rm b}$, which must be smaller than $p_{\rm BM}$.]

▶ 10.4 Communication (with errors) above capacity

We have proved, for any discrete memoryless channel, the achievability of a portion of the $R, p_{\rm b}$ plane shown in figure 10.6. We have shown that we can turn any noisy channel into an essentially noiseless binary channel with rate up to C bits per cycle. We now extend the right-hand boundary of the region of achievability at non-zero error probabilities. [This is called *rate-distortion theory*.]

We do this with a new trick. Since we know we can make the noisy channel into a perfect channel with a smaller rate, it is sufficient to consider communication with errors over a *noiseless* channel. How fast can we communicate over a noiseless channel, if we are allowed to make errors?

Consider a noiseless binary channel, and assume that we force communication at a rate greater than its capacity of 1 bit. For example, if we require the sender to attempt to communicate at $R = 2$ bits per cycle then he must effectively throw away half of the information. What is the best way to do this if the aim is to achieve the smallest possible probability of bit error? One simple strategy is to communicate a fraction $1/R$ of the source bits, and ignore the rest. The receiver guesses the missing fraction $1 - 1/R$ at random, and

the average probability of bit error is

$$p_{\rm b} = \frac{1}{2}(1 - 1/R). \tag{10.19}$$

The curve corresponding to this strategy is shown by the dashed line in figure 10.7.

We can do better than this (in terms of minimizing $p_{\rm b}$) by spreading out the risk of corruption evenly among all the bits. In fact, we can achieve $p_{\rm b} = H_2^{-1}(1 - 1/R)$, which is shown by the solid curve in figure 10.7. So, how can this optimum be achieved?

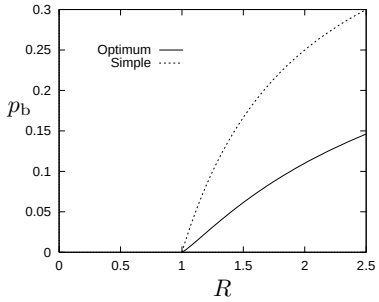


Figure 10.7. A simple bound on achievable points $(R, p_{\rm b})$, and Shannon's bound.

We reuse a tool that we just developed, namely the (N, K) code for a noisy channel, and we turn it on its head, using the *decoder* to define a lossy compressor. Specifically, we take an excellent (N, K) code for the binary symmetric channel. Assume that such a code has a rate $R' = K/N$, and that it is capable of correcting errors introduced by a binary symmetric channel whose transition probability is q. Asymptotically, rate-R' codes exist that have $R' \simeq 1 - H_2(q)$. Recall that, if we attach one of these capacity-achieving codes of length N to a binary symmetric channel then (a) the probability distribution over the outputs is close to uniform, since the entropy of the output is equal to the entropy of the source (NR') plus the entropy of the noise ($NH_2(q)$), and (b) the optimal decoder of the code, in this situation, typically maps a received vector of length N to a transmitted vector differing in qN bits from the received vector.

We take the signal that we wish to send, and chop it into blocks of length N (yes, N, not K). We pass each block through the *decoder*, and obtain a shorter signal of length K bits, which we communicate over the noiseless channel. To decode the transmission, we pass the K bit message to the *encoder* of the original code. The reconstituted message will now differ from the original message in some of its bits – typically qN of them. So the probability of bit error will be $p_{\rm b} = q$. The rate of this lossy compressor is $R = N/K = 1/R' = 1/(1 - H_2(p_{\rm b}))$.

Now, attaching this lossy compressor to our capacity-C error-free communicator, we have proved the achievability of communication up to the curve $(p_{\rm b}, R)$ defined by:

$$R = \frac{C}{1 - H_2(p_{\rm b})}. \qquad \square \tag{10.20}$$

For further reading about rate-distortion theory, see Gallager (1968), p. 451, or McEliece (2002), p. 75.

▶ 10.5 The non-achievable region (part 3 of the theorem)

The source, encoder, noisy channel and decoder define a Markov chain: $s \rightarrow \mathbf{x} \rightarrow \mathbf{y} \rightarrow \hat{s}$

$$P(s, \mathbf{x}, \mathbf{y}, \hat{s}) = P(s)P(\mathbf{x} \,|\, s)P(\mathbf{y} \,|\, \mathbf{x})P(\hat{s} \,|\, \mathbf{y}). \tag{10.21}$$

The data processing inequality (exercise 8.9, p.141) must apply to this chain: $I(s; \hat{s}) \leq I(\mathbf{x}; \mathbf{y})$. Furthermore, by the definition of channel capacity, $I(\mathbf{x}; \mathbf{y}) \leq NC$, so $I(s; \hat{s}) \leq NC$.

Assume that a system achieves a rate R and a bit error probability $p_{\rm b}$; then the mutual information $I(s; \hat{s})$ is $\geq NR(1 - H_2(p_{\rm b}))$. But $I(s; \hat{s}) > NC$ is not achievable, so $R > \frac{C}{1 - H_2(p_{\rm b})}$ is not achievable. $\square$

Exercise 10.1.[3] Fill in the details in the preceding argument. If the bit errors between $\hat{s}$ and s are independent then we have $I(s; \hat{s}) = NR(1 - H_2(p_{\rm b}))$.

What if we have complex correlations among those bit errors? Why does the inequality $I(s;\hat{s}) \geq NR(1-H_2(p_{\rm b}))$ hold?

▶ 10.6 Computing capacity

We have proved that the capacity of a channel is the maximum rate at which reliable communication can be achieved. How can we compute the capacity of a given discrete memoryless channel? We need to find its optimal input distribution. In general we can find the optimal input distribution by a computer search, making use of the derivative of the mutual information with respect to the input probabilities.

Sections 10.6–10.8 contain advanced material. The first-time reader is encouraged to skip to section 10.9 (p.172).

▷ Exercise 10.2.[2] Find the derivative of $I(X;Y)$ with respect to the input probability p_i, $\partial I(X;Y)/\partial p_i$, for a channel with conditional probabilities $Q_{j|i}$.

Exercise 10.3.[2] Show that $I(X;Y)$ is a concave $\frown$ function of the input probability vector $\mathbf{p}$.

Since $I(X;Y)$ is concave $\frown$ in the input distribution $\mathbf{p}$, any probability distribution $\mathbf{p}$ at which $I(X;Y)$ is stationary must be a global maximum of $I(X;Y)$. So it is tempting to put the derivative of $I(X;Y)$ into a routine that finds a local maximum of $I(X;Y)$, that is, an input distribution $P(x)$ such that

$$\frac{\partial I(X;Y)}{\partial p_i} = \lambda \ \text{ for all } i, \tag{10.22}$$

where λ is a Lagrange multiplier associated with the constraint $\sum_i p_i = 1$. However, this approach may fail to find the right answer, because $I(X;Y)$ might be maximized by a distribution that has $p_i = 0$ for some inputs. A simple example is given by the ternary confusion channel.

Ternary confusion channel. $\mathcal{A}_X = \{0, ?, 1\}$. $\mathcal{A}_Y = \{0, 1\}$.

$$\begin{array}{lll} P(y=0\,|\,x=0) = 1; & P(y=0\,|\,x=?) = 1/2; & P(y=0\,|\,x=1) = 0; \\ P(y=1\,|\,x=0) = 0; & P(y=1\,|\,x=?) = 1/2; & P(y=1\,|\,x=1) = 1. \end{array}$$

Whenever the input ? is used, the output is random; the other inputs are reliable inputs. The maximum information rate of 1 bit is achieved by making no use of the input ?.

▷ Exercise 10.4.[2, p.173] Sketch the mutual information for this channel as a function of the input distribution $\mathbf{p}$. Pick a convenient two-dimensional representation of $\mathbf{p}$.

The optimization routine must therefore take account of the possibility that, as we go up hill on $I(X;Y)$, we may run into the inequality constraints $p_i \geq 0$.

▷ Exercise 10.5.[2, p.174] Describe the condition, similar to equation (10.22), that is satisfied at a point where $I(X;Y)$ is maximized, and describe a computer program for finding the capacity of a channel.

Results that may help in finding the optimal input distribution

1. All outputs must be used.
2. $I(X;Y)$ is a convex $\smile$ function of the channel parameters.
3. There may be several optimal input distributions, but they all look the same at the output.

Reminder: The term 'convex $\smile$' means 'convex', and the term 'concave $\frown$' means 'concave'; the little smile and frown symbols are included simply to remind you what convex and concave mean.

▷ Exercise 10.6.[2] Prove that no output y is unused by an optimal input distribution, unless it is unreachable, that is, has $Q(y\,|\,x) = 0$ for all x.

Exercise 10.7.[2] Prove that $I(X;Y)$ is a convex $\smile$ function of $Q(y\,|\,x)$.

Exercise 10.8.[2] Prove that all optimal input distributions of a channel have the same output probability distribution $P(y) = \sum_x P(x)Q(y\,|\,x)$.

These results, along with the fact that $I(X;Y)$ is a concave $\frown$ function of the input probability vector $\mathbf{p}$, prove the validity of the symmetry argument that we have used when finding the capacity of symmetric channels. If a channel is invariant under a group of symmetry operations – for example, interchanging the input symbols and interchanging the output symbols – then, given any optimal input distribution that is not symmetric, i.e., is not invariant under these operations, we can create another input distribution by averaging together this optimal input distribution and all its permuted forms that we can make by applying the symmetry operations to the original optimal input distribution. The permuted distributions must have the same $I(X;Y)$ as the original, by symmetry, so the new input distribution created by averaging must have $I(X;Y)$ bigger than or equal to that of the original distribution, because of the concavity of I.

Symmetric channels

In order to use symmetry arguments, it will help to have a definition of a symmetric channel. I like Gallager's (1968) definition.

A discrete memoryless channel is a symmetric channel if the set of outputs can be partitioned into subsets in such a way that for each subset the matrix of transition probabilities has the property that each row (if more than 1) is a permutation of each other row and each column is a permutation of each other column.

Example 10.9. This channel

$$\begin{array}{lllllll} P(y=0\,|\,x=0) & = & 0.7\,; & P(y=0\,|\,x=1) & = & 0.1\,; \\ P(y=?\,|\,x=0) & = & 0.2\,; & P(y=?\,|\,x=1) & = & 0.2\,; \\ P(y=1\,|\,x=0) & = & 0.1\,; & P(y=1\,|\,x=1) & = & 0.7. \end{array} \tag{10.23}$$

is a symmetric channel because its outputs can be partitioned into $(0, 1)$ and ?, so that the matrix can be rewritten:

$$\begin{array}{|lllllll|} \hline P(y=0\,|\,x=0) & = & 0.7\,; & P(y=0\,|\,x=1) & = & 0.1\,; \\ P(y=1\,|\,x=0) & = & 0.1\,; & P(y=1\,|\,x=1) & = & 0.7\,; \\ \hline P(y=?\,|\,x=0) & = & 0.2\,; & P(y=?\,|\,x=1) & = & 0.2. \\ \hline \end{array} \tag{10.24}$$

Symmetry is a useful property because, as we will see in a later chapter, communication at capacity can be achieved over symmetric channels by *linear* codes.

Exercise 10.10.[2] Prove that for a symmetric channel with any number of inputs, the uniform distribution over the inputs is an optimal input distribution.

▷ Exercise 10.11.[2, p.174] Are there channels that are not symmetric whose optimal input distributions are uniform? Find one, or prove there are none.

▶ 10.7 Other coding theorems

The noisy-channel coding theorem that we proved in this chapter is quite general, applying to any discrete memoryless channel; but it is not very specific. The theorem only says that reliable communication with error probability ϵ and rate R can be achieved by using codes with *sufficiently large* blocklength N. The theorem does not say how large N needs to be to achieve given values of R and ϵ.

Presumably, the smaller ϵ is and the closer R is to C, the larger N has to be.

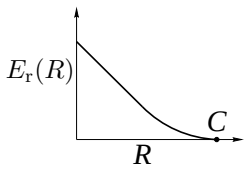


Figure 10.8. A typical random-coding exponent.

Noisy-channel coding theorem – version with explicit N-dependence

> For a discrete memoryless channel, a blocklength N and a rate R, there exist block codes of length N whose average probability of error satisfies:
>
> $$p_{\rm B} \leq \exp\left[-NE_{\rm r}(R)\right] \tag{10.25}$$
>
> where $E_{\rm r}(R)$ is the *random-coding exponent* of the channel, a convex $\smile$, decreasing, positive function of R for $0 \leq R < C$. The random-coding exponent is also known as the reliability function.
>
> [By an expurgation argument it can also be shown that there exist block codes for which the *maximal* probability of error $p_{\rm BM}$ is also exponentially small in N.]

The definition of $E_{\rm r}(R)$ is given in Gallager (1968), p. 139. $E_{\rm r}(R)$ approaches zero as $R \to C$; the typical behaviour of this function is illustrated in figure 10.8. The computation of the random-coding exponent for interesting channels is a challenging task on which much effort has been expended. Even for simple channels like the binary symmetric channel, there is no simple expression for $E_{\rm r}(R)$.

Lower bounds on the error probability as a function of blocklength

The theorem stated above asserts that there are codes with $p_{\rm B}$ smaller than $\exp\left[-NE_{\rm r}(R)\right]$. But how small can the error probability be? Could it be much smaller?

> For any code with blocklength N on a discrete memoryless channel, the probability of error assuming all source messages are used with equal probability satisfies
>
> $$p_{\rm B} \gtrsim \exp[-NE_{\rm sp}(R)], \tag{10.26}$$

where the function $E_{\rm sp}(R)$, the *sphere-packing exponent* of the channel, is a convex $\smile$, decreasing, positive function of R for $0 \leq R < C$.

For a precise statement of this result and further references, see Gallager (1968), p. 157.

▶ 10.8 Noisy-channel coding theorems and coding practice

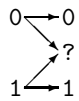
Imagine a customer who wants to buy an error-correcting code and decoder for a noisy channel. The results described above allow us to offer the following service: if he tells us the properties of his channel, the desired rate R and the desired error probability $p_{\rm B}$, we can, after working out the relevant functions C, $E_{\rm r}(R)$, and $E_{\rm sp}(R)$, advise him that there exists a solution to his problem using a particular blocklength N; indeed that almost any randomly chosen code with that blocklength should do the job. Unfortunately we have not found out how to implement these encoders and decoders in practice; the cost of implementing the encoder and decoder for a random code with large N would be exponentially large in N.

Furthermore, for practical purposes, the customer is unlikely to know exactly what channel he is dealing with. So Berlekamp (1980) suggests that the sensible way to approach error-correction is to design encoding-decoding systems and plot their performance on a *variety* of idealized channels as a function of the channel's noise level. These charts (one of which is illustrated on page 568) can then be shown to the customer, who can choose among the systems on offer without having to specify what he really thinks his channel is like. With this attitude to the practical problem, the importance of the functions $E_{\rm r}(R)$ and $E_{\rm sp}(R)$ is diminished.

▶ 10.9 Further exercises

Exercise 10.12.[2] A binary erasure channel with input x and output y has transition probability matrix:

$$\mathbf{Q} = \begin{bmatrix} 1-q & 0 \\ q & q \\ 0 & 1-q \end{bmatrix}$$



Find the *mutual information* $I(X;Y)$ between the input and output for general input distribution $\{p_0, p_1\}$, and show that the *capacity* of this channel is $C = 1 - q$ bits.

A Z channel has transition probability matrix:

$$\mathbf{Q} = \begin{bmatrix} 1 & q \\ 0 & 1-q \end{bmatrix}$$

Show that, using a $(2, 1)$ code, **two** uses of a Z channel can be made to emulate **one** use of an erasure channel, and state the erasure probability of that erasure channel. Hence show that the capacity of the Z channel, $C_{\rm Z}$, satisfies $C_{\rm Z} \geq \frac{1}{2}(1 - q)$ bits.

Explain why the result $C_{\rm Z} \geq \frac{1}{2}(1 - q)$ is an inequality rather than an equality.

Exercise 10.13.[3, p.174] A transatlantic cable contains $N = 20$ indistinguishable electrical wires. You have the job of figuring out which wire is which, that is, to create a consistent labelling of the wires at each end. Your only tools are the ability to connect wires to each other in groups of two or more, and to test for connectedness with a continuity tester. What is the smallest number of transatlantic trips you need to make, and how do you do it?

How would you solve the problem for larger N such as $N = 1000$?

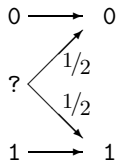
As an illustration, if N were 3 then the task can be solved in two steps by labelling one wire at one end a, connecting the other two together, crossing the Atlantic, measuring which two wires are connected, labelling them b and c and the unconnected one a, then connecting b to a and returning across the Atlantic, whereupon on disconnecting b from c, the identities of b and c can be deduced.

This problem can be solved by persistent search, but the reason it is posed in this chapter is that it can also be solved by a greedy approach based on maximizing the acquired *information*. Let the unknown permutation of wires be x. Having chosen a set of connections of wires $\mathcal{C}$ at one end, you can then make measurements at the other end, and these measurements y convey *information* about x. How much? And for what set of connections is the information that y conveys about x maximized?

▶ 10.10 Solutions

Solution to exercise 10.4 (p.169). If the input distribution is $\mathbf{p} = (p_0, p_?, p_1)$, the mutual information is

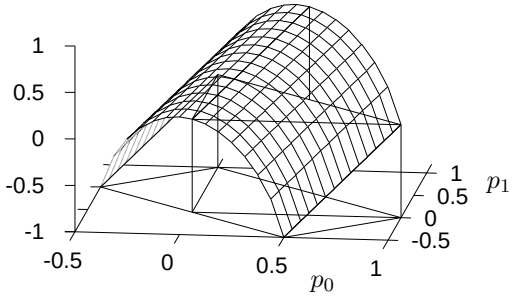
$$I(X;Y) = H(Y) - H(Y|X) = H_2(p_0 + p_?/2) - p_?. \tag{10.27}$$

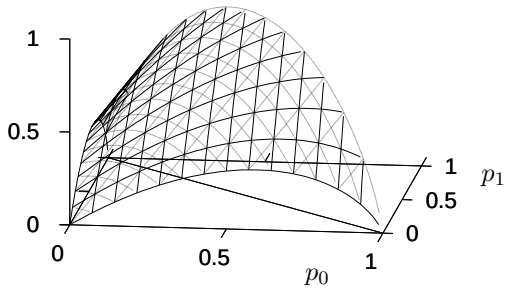


We can build a good sketch of this function in two ways: by careful inspection of the function, or by looking at special cases.

For the plots, the two-dimensional representation of $\mathbf{p}$ I will use has p_0 and p_1 as the independent variables, so that $\mathbf{p} = (p_0, p_?, p_1) = (p_0, (1-p_0-p_1), p_1)$.

By inspection. If we use the quantities $p_* \equiv p_0 + p_?/2$ and $p_?$ as our two degrees of freedom, the mutual information becomes very simple: $I(X;Y) = H_2(p_*) - p_?$. Converting back to $p_0 = p_* - p_?/2$ and $p_1 = 1 - p_* - p_?/2$, we obtain the sketch shown at the left below. This function is like a tunnel rising up the direction of increasing p_0 and p_1. To obtain the required plot of $I(X;Y)$ we have to strip away the parts of this tunnel that live outside the feasible simplex of probabilities; we do this by redrawing the surface, showing only the parts where $p_0 > 0$ and $p_1 > 0$. A full plot of the function is shown at the right.





Special cases. In the special case $p_? = 0$, the channel is a noiseless binary channel, and $I(X;Y) = H_2(p_0)$.

In the special case $p_0 = p_1$, the term $H_2(p_0 + p_?/2)$ is equal to 1, so $I(X;Y) = 1 - p_?$.

In the special case $p_0 = 0$, the channel is a Z channel with error probability 0.5. We know how to sketch that, from the previous chapter (figure 9.3).

These special cases allow us to construct the skeleton shown in figure 10.9.

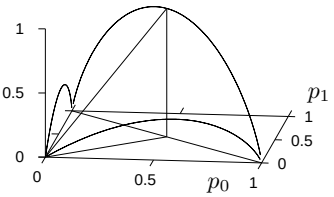


Figure 10.9. Skeleton of the mutual information for the ternary confusion channel.

Solution to exercise 10.5 (p.169). Necessary and sufficient conditions for $\mathbf{p}$ to maximize $I(X;Y)$ are

$$\left.\begin{array}{rcll} \frac{\partial I(X;Y)}{\partial p_i} & = & \lambda & \text{and} \quad p_i > 0 \\ \frac{\partial I(X;Y)}{\partial p_i} & \leq & \lambda & \text{and} \quad p_i = 0 \end{array}\right\} \quad \text{for all } i, \tag{10.28}$$

where λ is a constant related to the capacity by $C = \lambda + \log_2 e$.

This result can be used in a computer program that evaluates the derivatives, and increments and decrements the probabilities p_i in proportion to the differences between those derivatives.

This result is also useful for lazy human capacity-finders who are good guessers. Having guessed the optimal input distribution, one can simply confirm that equation (10.28) holds.

Solution to exercise 10.11 (p.171). We certainly expect nonsymmetric channels with uniform optimal input distributions to exist, since when inventing a channel we have $I(J-1)$ degrees of freedom whereas the optimal input distribution is just $(I-1)$-dimensional; so in the $I(J-1)$-dimensional space of perturbations around a symmetric channel, we expect there to be a subspace of perturbations of dimension $I(J-1) - (I-1) = I(J-2) + 1$ that leave the optimal input distribution unchanged.

Here is an explicit example, a bit like a Z channel.

$$\mathbf{Q} = \begin{bmatrix} 0.9585 & 0.0415 & 0.35 & 0.0 \\ 0.0415 & 0.9585 & 0.0 & 0.35 \\ 0 & 0 & 0.65 & 0 \\ 0 & 0 & 0 & 0.65 \end{bmatrix} \tag{10.29}$$

Solution to exercise 10.13 (p.173). The labelling problem can be solved for any $N > 2$ with just two trips, one each way across the Atlantic.

The key step in the information-theoretic approach to this problem is to write down the information content of one *partition*, the combinatorial object that is the connecting together of subsets of wires. If N wires are grouped together into g_1 subsets of size 1, g_2 subsets of size 2, ..., then the number of such partitions is

$$\Omega = \frac{N!}{\prod_r (r!)^{g_r} \, g_r!}, \tag{10.30}$$

and the information content of one such partition is the log of this quantity. In a greedy strategy we choose the first partition to maximize this information content.

One game we can play is to maximize this information content with respect to the quantities g_r, treated as real numbers, subject to the constraint $\sum_r g_r r = N$. Introducing a Lagrange multiplier λ for the constraint, the derivative is

$$\frac{\partial}{\partial g_r} \left(\log \Omega + \lambda \sum_r g_r r \right) = -\log r! - \log g_r + \lambda r, \tag{10.31}$$

which, when set to zero, leads to the rather nice expression

$$g_r = \frac{e^{\lambda r}}{r!}; \tag{10.32}$$

the optimal g_r is proportional to a Poisson distribution! We can solve for the Lagrange multiplier by plugging g_r into the constraint $\sum_r g_r r = N$, which gives the implicit equation

$$N = \mu\, e^{\mu}, \tag{10.33}$$

where $\mu \equiv e^{\lambda}$ is a convenient reparameterization of the Lagrange multiplier. Figure 10.10a shows a graph of $\mu(N)$; figure 10.10b shows the deduced non-integer assignments g_r when $\mu = 2.2$, and nearby integers $g_r = \{1, 2, 2, 1, 1\}$ that motivate setting the first partition to (a)(bc)(de)(fgh)(ijk)(lmno)(pqrst).

This partition produces a random partition at the other end, which has an information content of $\log \Omega = 40.4$ bits, which is a lot more than half the total information content we need to acquire to infer the transatlantic permutation, $\log 20! \simeq 61$ bits. [In contrast, if all the wires are joined together in pairs, the information content generated is only about 29 bits.] How to choose the second partition is left to the reader. A Shannonesque approach is appropriate, picking a random partition at the other end, using the same $\{g_r\}$; you need to ensure the two partitions are as unlike each other as possible.

If $N \neq 2$, 5 or 9, then the labelling problem has solutions that are particularly simple to implement, called Knowlton–Graham partitions: partition $\{1, \ldots, N\}$ into disjoint sets in two ways A and B, subject to the condition that at most one element appears both in an A set of cardinality j and in a B set of cardinality k, for each j and k (Graham, 1966; Graham and Knowlton, 1968).

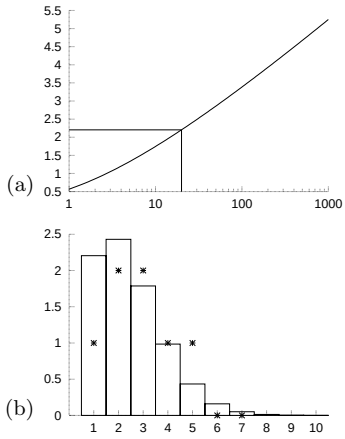


Figure 10.10. Approximate solution of the cable-labelling problem using Lagrange multipliers. (a) The parameter μ as a function of N; the value $\mu(20) = 2.2$ is highlighted. (b) Non-integer values of the function $g_r = \mu^r/r!$ are shown by lines and integer values of g_r motivated by those non-integer values are shown by crosses.