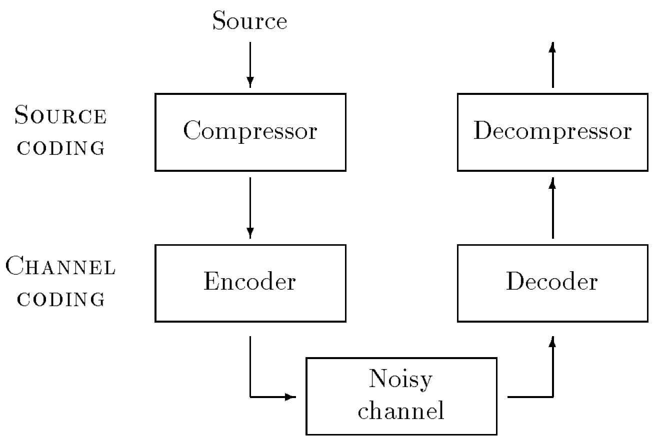
1 The big picture

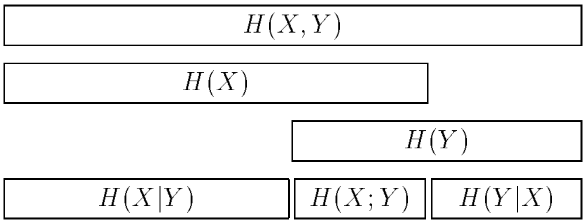


In lectures 2, 3 and 4, we have discussed source coding with block codes, symbol codes and arithmetic codes. We will now spend two lectures on the subject of channel coding—the fundamental possibilities and limitations of error-free communication through a noisy channel.

2 Review of probability theory and conditional, joint and mutual information

Please see notes of lecture 1.

A useful picture breaks down the total information content $H(X,Y)$ of a joint ensemble thus:



Example of a joint ensemble

A joint ensemble XY has joint distribution $P(x,y)$:

$P(x,y)$		x = 1	2	3	4	$P(y)$
	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 = 1	2	3	4	$H(X|y)$
	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 $H(X)$. On average though, y does convey information about x, as $H(X|Y) < H(X)$.

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

3 Communication over a noisy channel

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 $x \in \mathcal{A}_X$.

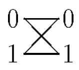
The transition probabilities may be written in a matrix

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

Our aim here is to show that for any channel there is a finite rate, the capacity $C(Q)$, up to which information can be sent with arbitrarily small probability of error.

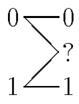
Some useful model channels are:

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



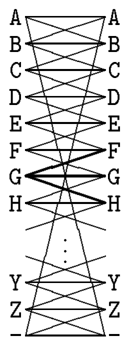
$P(y{=}0|x{=}0){=}1-q$; $P(y{=}0|x{=}1){=}q$;
$P(y{=}1|x{=}0){=}q$; $P(y{=}1|x{=}1){=}1-q$.

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



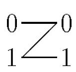
$P(y{=}0|x{=}0){=}1-q$; $P(y{=}0|x{=}1){=}0$;
$P(y{=}?|x{=}0){=}q$; $P(y{=}?|x{=}1){=}q$;
$P(y{=}1|x{=}0){=}0$; $P(y{=}1|x{=}1){=}1-q$.

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



$P(y{=}\texttt{F}|x{=}\texttt{G}){=}1/3$;
$P(y{=}\texttt{G}|x{=}\texttt{G}){=}1/3$;
$P(y{=}\texttt{H}|x{=}\texttt{G}){=}1/3$.

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



$P(y{=}0|x{=}0){=}1$; $P(y{=}0|x{=}1){=}q$;
$P(y{=}1|x{=}0){=}0$; $P(y{=}1|x{=}1){=}1-q$.

Inferring a single transmission

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

$$P(x,y) = P(y|x)P(x).$$

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

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} = \frac{P(y|x)P(x)}{\sum_{x'} P(y|x')P(x')}$$

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

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

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

Alternatively, assume we observe $y{=}0$.

$$\begin{aligned} P(x{=}1|y{=}0) &= \frac{P(y{=}0|x{=}1)P(x{=}1)}{\sum_{x'} P(y|x')P(x')} \\ &= \frac{0.15 \times 0.1}{0.15 \times 0.1 + 0.85 \times 0.9} \\ &= \frac{0.015}{0.78} = 0.019. \end{aligned}$$

Example 2: Consider a Z channel with probability of error $q = 0.15$. Let the input ensemble be $\mathcal{P}_X : \{p_0{=}0.9, p_1{=}0.1\}$. Assume we observe $y{=}1$.

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

Given the output $y = 1$ we become certain of the input. Alternatively, assume we observe $y = 0$.

$$\begin{aligned} P(x{=}1|y{=}0) &= \frac{0.15 \times 0.1}{0.15 \times 0.1 + 1.0 \times 0.9} \\ &= \frac{0.015}{0.915} = 0.016. \end{aligned}$$

Information conveyed by the 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:

$$H(X;Y) \equiv H(X) - H(X|Y) = H(Y) - H(Y|X)$$

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

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

Example 1: Consider the BSC again, with $q{=}0.15$ 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:

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

This may be contrasted with the entropy of the source $H(X) = H_2(0.15) = 0.47$ bits.

Example 2: And now the Z channel. $P(y{=}1){=}0.085$.

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

The entropy of the source is $H(X) = 0.47$ bits.

Maximizing the mutual information

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:

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

The distribution $\mathcal{P}_X$ that achieves the maximum is called the optimal input distribution, $\mathcal{P}^*_X$.

We will eventually show that the capacity does indeed measure the maximum amount of error-free information that is transmittable over the channel per unit time.

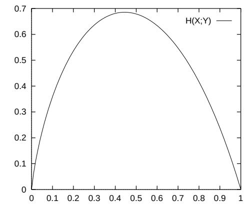
Example 1: Consider the BSC with $q{=}0.15$. Above we considered $\mathcal{P}_X : \{p_0{=}0.9, p_1{=}0.1\}$, and found $H(X;Y) = 0.15$ bits. How much better can we do? By symmetry, or by explicit maximization, the optimal input distribution is $\{0.5, 0.5\}$ and $C(Q_{\text{BSC}}) = H_2(0.5) - H_2(0.15) = 1.0 - 0.61 = 0.39$ bits.

Example 2: The noisy typewriter. The optimal input distribution is a uniform distribution over x, and gives $C = \log_2 9$ bits. For this channel, we can easily create an error-free communication strategy: we only use the letters `B`, `E`, `H`, ...`Z`, *i.e.*, every third letter. This is a *non-confusable subset* of the input alphabet. 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 $= C$.

Example 3: Consider the Z channel with q=0.15. Identifying the optimal input distribution is not so straightforward. We evaluate $H(X;Y)$ explicitly for $\mathcal{P}_X = \{p_0, p_1\}$.

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

This is a non-trivial function of p_1:



It is maximized for $q = 0.15$ by $p_1^* = 0.445$. We find $C(Q_Z) = 0.685$.

Exercise a: Why is p_1^* less than 0.5? In the case of general q, show that the optimal input distribution is

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

What happens to p_1^* as the noise level q tends to 1?
Exercise b: Show that the capacity of the Binary Erasure Channel with $q = 0.15$ is $C_{\rm BEC} = 0.85$. What is its capacity for general q? Comment.
Exercise c: Sketch graphs of the capacity of the Z channel, the BSC and the BEC as a function of q.

4 The Channel Coding Theorem

An (N, K) **Block Code** for a channel Q is a list of $M = 2^K$ codewords $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots \mathbf{x}^{(2^K)}\}$, each of length N: $\mathbf{x}^{(m)} \in \mathcal{A}_X^N$. The signal to be encoded is assumed to come from an alphabet of size 2^K; signal m is encoded as $\mathbf{x}^{(m)}$.

The **rate** of the code is $R = K/N$.

A linear (N, K) **block code** is a block code in which all the codewords $\{\mathbf{x}^{(m)}\}$ lie in a K-dimensional subspace of $\mathcal{A}_X^N$. The encoding operation can be represented by a binary matrix $\mathbf{G}$ such that if the signal to be encoded, in binary notation, is $\mathbf{s}$ (a vector of length K bits), then the encoded signal is $\mathbf{Gs}$ modulo 2.

For example the (7,4) Hamming code takes $K = 4$ signal bits, $\mathbf{s}$, and transmits them followed by three parity check bits. The $N = 7$ transmitted symbols are given by $\mathbf{Gs} \bmod 2$, where:

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}.$$

A decoder for an (N, K) Block Code is a mapping from $\mathcal{A}_Y^N$ to $m \in \{0, 1, 2, \ldots 2^K\}$.

The extra symbol $m = 0$ 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 $m_{\rm in}$, is:

$$p_{\rm B} = \sum_{m_{\rm in}} P(m_{\rm in}) P(m_{\rm out} \neq m_{\rm in} | m_{\rm in})$$

The maximal probability of block error is

$$p_{\rm BM} = \max_{m_{\rm in}} P(m_{\rm out} \neq m_{\rm in} | m_{\rm in})$$

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 m that has maximum posterior probability $P(m|\mathbf{y})$. This decoder is often called the maximum likelihood decoder, since a uniform distribution on m is usually assumed.

The probability of bit error p_b is defined assuming that m is represented by a binary vector $\mathbf{s}$; it is the average probability that a bit of $\mathbf{s}_{\rm out}$ is not equal to the corresponding bit of $\mathbf{s}_{\rm in}$.

The Theorem. Associated with each discrete memoryless channel, there is a nonnegative number C (called channel capacity) with 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 $< \epsilon$.

The other bit of the theorem. Error free communication above capacity is impossible. It is not possible to achieve a rate R with probability of bit error smaller than

$$p_b(R) = H_2^{-1}(1 - C/R).$$

This is the equation of the curve separating achievable from non-achievable points (bit error probability p_b, rate R). For example, for the binary symmetric channel with cross-over probability q, the curve can be written:

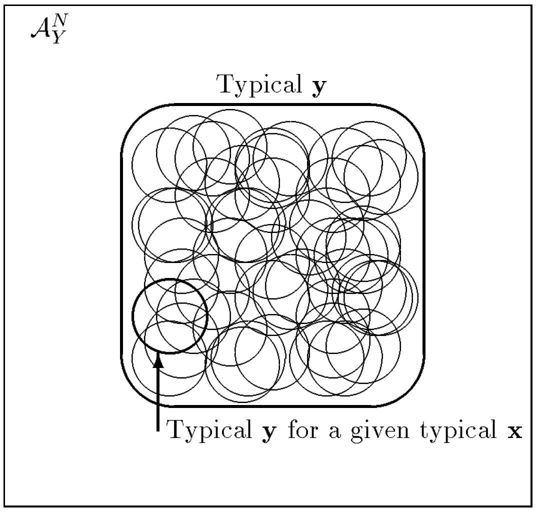
$$R(p_b) = \frac{1 - H_2(q)}{1 - H_2(p_b)}.$$

Intuitive preview of proof

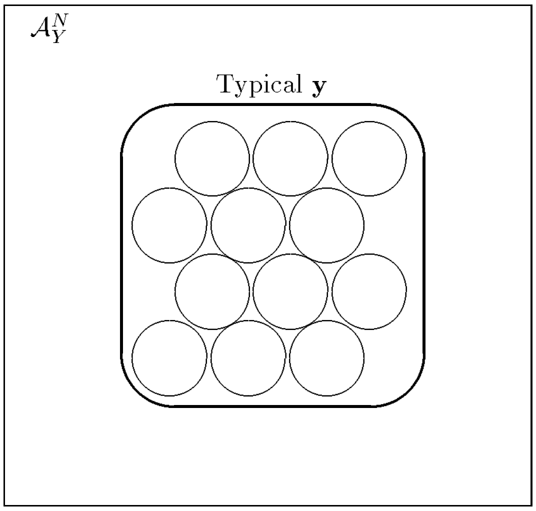
We make use of large block lengths N. We can consider the extended channel with $|\mathcal{A}_X|^N$ possible inputs $\mathbf{x}$ and $|\mathcal{A}_Y|^N$ possible outputs. The intuitive idea is that *this channel looks a lot like the noisy typewriter.* Any particular input $\mathbf{x}$ produces, with high probability, an output in a small subspace of the output alphabet. 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 lecture 2, and consider the number of probable output sequences $\mathbf{y}$. The total number of typical output sequences $\mathbf{y}$, when $\mathbf{x}$ comes from the ensemble X^N, 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 the figure below.



We now imagine restricting ourselves to a subset of the typical inputs $\mathbf{x}$ such that the corresponding typical output sets, as in the noisy typewriter, do not overlap. 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 the 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 $\leq 2^{NH(Y)-NH(Y|X)} = 2^{NH(X;Y)}$.



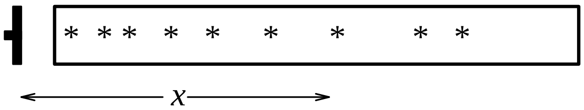
The maximum value of this bound is achieved if X is the ensemble that maximizes $H(X;Y)$, in which case the number of non-confusable inputs is $\leq 2^{NC}$. Thus we anticipate that assymptotically up to C bits per cycle, and no more, can be communicated with vanishing error probability.

5 Inference problems

Please have a think about these questions:

Inferring a decay constant

Unstable particles are emitted from a source and decay at a distance x that has an exponential probability distribution with characteristic length λ. Decay events can only be observed if they occur in a window extending from $x = 1\text{cm}$ to $x = 20\text{cm}$. N decays are observed at locations $\{x_1 \ldots x_N\}$. What is λ?



Genetic test evidence

Two people have left traces of their own blood at the scene of a crime. Their blood groups can be reliably identified from these traces and are found to be of type 'O' (a common type in the local population, having frequency 60%) and of type 'AB' (a rare type, with frequency 1%). A suspect is tested and found to have type 'O' blood. Do these data (type 'O' and 'AB' blood were found at scene) give evidence in favour of the proposition that this suspect was one of the two people present at the crime?

6 Possible topics for final lecture or two

- More outrageous results in information theory: communication by multiple users over a single channel; compression of correlated information without collaboration.
- Continuous channels and sources.
- A variational free energy decoding algorithm.
- Neural networks as information storage devices.
- Theory of computation. Turing machines. Uncomputability and the halting problem.
- Complexity theory. NP-complete problems.
- Bayesian inference and maximum entropy.

Feedback to `mackay@mrao.cam.ac.uk`, please.