Information Theory

http://www.inf.ed.ac.uk/teaching/courses/it/

Week 6

Communication channels and Information

Iain Murray, 2012

School of Informatics, University of Edinburgh

Noisy channel communication

Input File

↓ compress

$\mathbf{x} = 00110001$

↓ add redundancy

00110001
00110001
00110001

→ Noisy Channel →

00011001
01110011
10110100

transmitted message
often called $\mathbf{t}$ or $\mathbf{x}$

decode ↓

$\hat{\mathbf{x}} = 00110001$

decompress ↓

If all errors corrected: **Original file**

Some notes on the noisy channel setup:

Noisy communication was outlined in lecture 1, then abandoned to cover compression, representing messages for a noiseless channel.

Why compress, remove all redundancy, just to add it again?

Firstly remember that repetition codes require a *lot* of repetitions to get a negligible probability of error. We are going to have to add better forms of redundancy to get reliable communication at good rates. Our files won't necessarily have the right sort of redundancy.

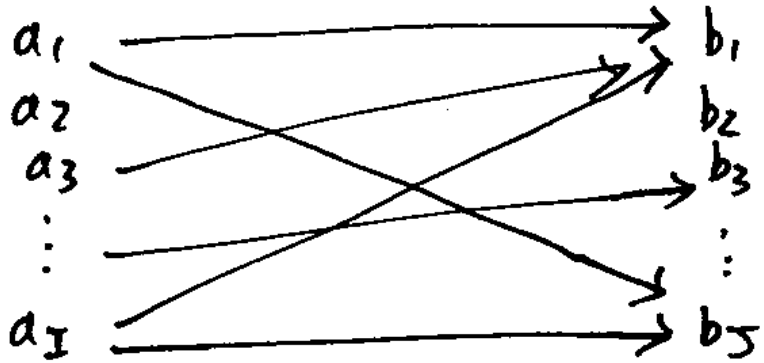
It is often useful to have modular designs. We can design an encoding/decoding scheme for a noisy channel separately from modelling data. Then use a compression system to get our file appropriately distributed over the required alphabet.

It is possible to design a combined system that takes redundant files and encodes them for a noisy channel. MN codes do this:
`http://www.inference.phy.cam.ac.uk/mackay/mncN.pdf`
These lectures won't discuss this option.

Discrete Memoryless Channel, Q

Discrete: Inputs x and Outputs y have discrete (sometimes binary) alphabets:

$x \in \mathcal{A}_X$



$y \in \mathcal{A}_Y$

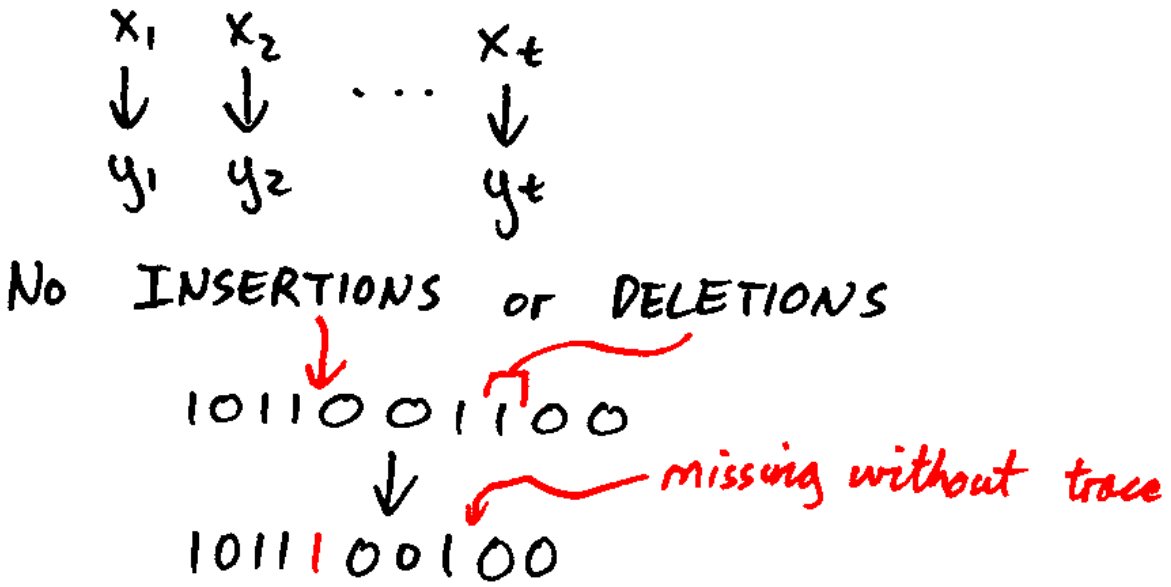
$$Q_{j|i} = P(y = b_j \,|\, x = a_i)$$

Memoryless: outputs always drawn using fixed Q matrix

We also assume channel is **synchronized**

Synchronized channels

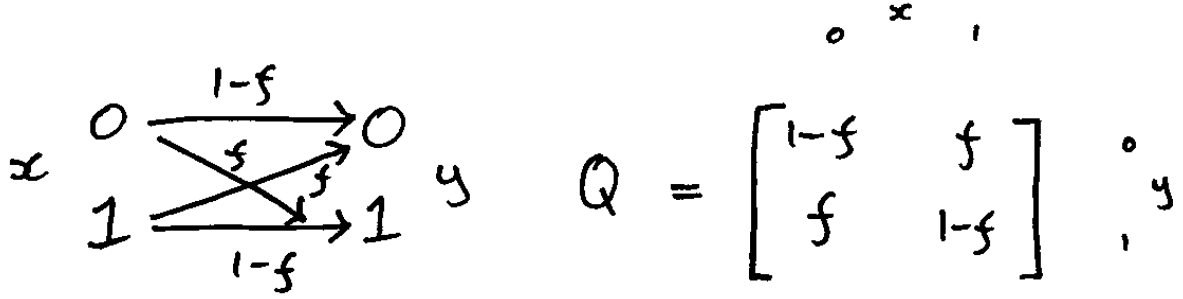
We know that a sequence of inputs was sent and which outputs go with them.



Dealing with insertions and deletions is a tricky topic, an active area of research that we will avoid

Binary Symmetric Channel (BSC)

A natural model channel for binary data:



Alternative view:

noise drawn from $p(n) = \begin{cases} 1-f & n=0 \\ f & n=1 \end{cases}$

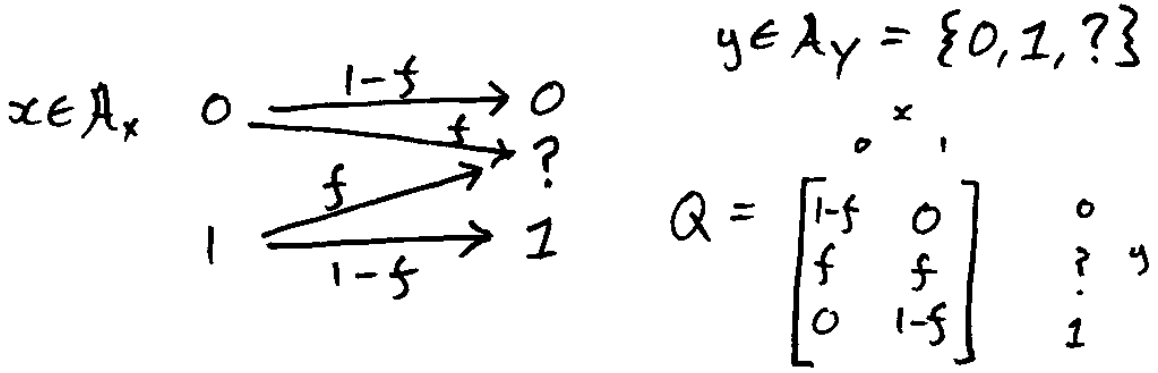
$$y \;=\; (x+n) \bmod 2 \;=\; x \;\texttt{XOR}\; n$$

```
% Matlab/Octave
y = mod(x+n, 2);
y = bitxor(x, n);
```

```
/* C (or Python) */
y = (x+n) % 2;
y = x ^ n;
```

Binary Erasure Channel (BEC)

An example of a non-binary alphabet:



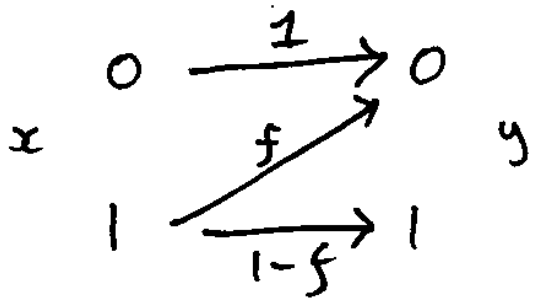
With this channel corruptions are obvious

Feedback: could ask for retransmission
Care required: negotiation could be corrupted too
Feedback sometimes not an option: hard disk storage

The BEC is not the *deletion channel*. Here symbols are replaced with a placeholder, in the deletion channel they are removed entirely and it is no longer clear at what time symbols were transmitted.

Z channel

Cannot always treat symbols symmetrically



$$Q = \begin{matrix} \\ x \end{matrix} \begin{matrix} \begin{bmatrix} 1 & f \\ 0 & 1-f \end{bmatrix} \end{matrix} \begin{matrix} \\ y \end{matrix}$$

"Ink gets rubbed off, but never added"

Channel Probabilities

Channel definition:

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

Assume there's nothing we can do about Q.
We can choose what to throw at the channel.

Input distribution: $\mathbf{p}_X = \begin{pmatrix} p(x = a_1) \\ \vdots \\ p(x = a_I) \end{pmatrix}$

Joint distribution: $P(x, y) = P(x)\, P(y \,|\, x)$

Output distribution: $P(y) = \sum_x P(x, y)$

vector notation: $\mathbf{p}_Y = Q\, \mathbf{p}_X$

(the usual relationships for any two variables x and y)

A little more detail on channel probabilities:

More detail on why the output distribution can be found by a matrix multiplication:

$$\begin{aligned} p_{Y,j} = P(y{=}b_j) &= \sum_i P(y{=}b_j, x{=}a_i) \\ &= \sum_i P(y{=}b_j \,|\, x{=}a_i)\, P(x{=}a_i) \\ &= \sum_i Q_{j|i}\, p_{X,i} \\ \mathbf{p}_Y &= Q\, \mathbf{p}_X \end{aligned}$$

Care: some texts (but not MacKay) use the transpose of our Q as the transition matrix, and so use left-multiplication instead.

Channels and Information

Three distributions: $P(x)$, $P(y)$, $P(x, y)$
Three observers: sender, receiver, omniscient outsider

Average surprise of receiver: $H(Y) = \sum_y P(y) \log {}^1\!/_{P(y)}$
Partial information about sent file and added noise

Average information of file: $H(X) = \sum_x P(x) \log {}^1\!/_{P(x)}$
Sender observes all of this, but no information about noise

Omniscient outsider experiences total joint entropy of file and noise: $H(X, Y) = \sum_{x,y} P(x, y) \log {}^1\!/_{P(x,y)}$

Joint Entropy

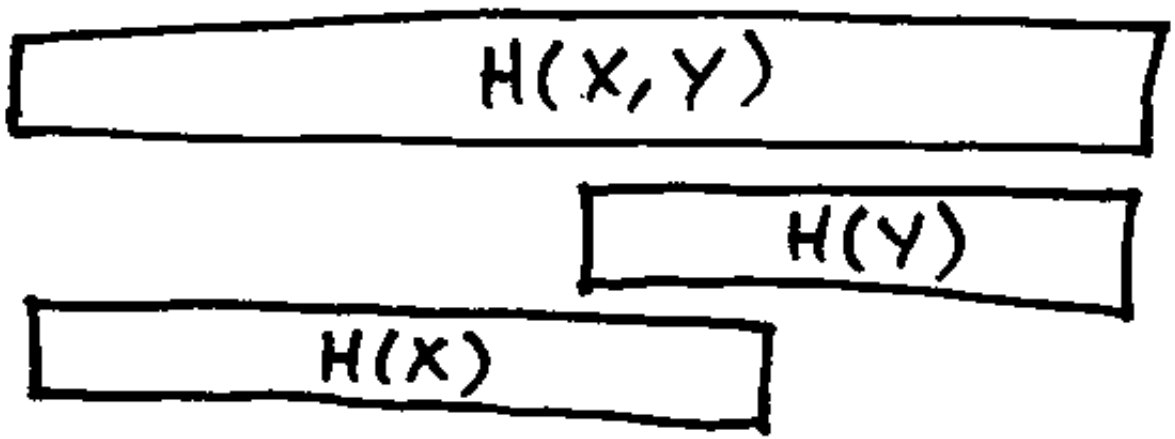
Omniscient outsider gets more information on average than an observer at one end of the channel: $H(X,Y) \geq H(X)$

Outsider can't have more information than both ends combined:

$$H(X,Y) \leq H(X) + H(Y)$$

with equality only if X and Y are independent

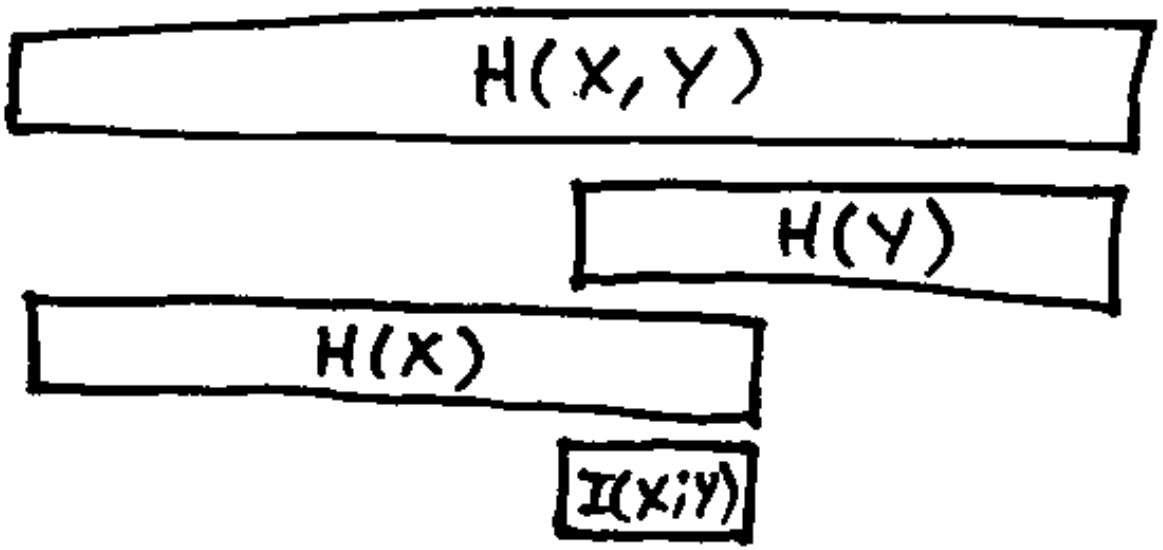
(independence useless for communication!)



Mutual Information (1)

How much too big is $H(X) + H(Y) \neq H(X, Y)$?

Overlap: $I(X;Y) = H(X) + H(Y) - H(X, Y)$
is called the **mutual information**



It's the average information content "shared" by the dependent X and Y ensembles. (more insight to come)

Inference in the channel

The receiver doesn't know x, but on receiving y can update the prior $P(x)$ to a posterior:

$$P(x \mid y) = \frac{P(x, y)}{P(y)} = \frac{P(y \mid x)\, P(x)}{P(y)}$$

e.g. for BSC with $P(x=1) = 0.5$, $P(x \mid y) = \begin{cases} 1-f & x = 0 \\ f & x = 1 \end{cases}$

other channels may have less obvious posteriors

Another distribution we can compute the entropy of!

Conditional Entropy (1)

We can condition every part of an expression on the setting of an arbitrary variable:

$$H(X \mid y) = \sum_x P(x \mid y) \log {}^1\!/_{P(x \mid y)}$$

Average information available from seeing x, given that we already know y.

On average this is written:

$$H(X \mid Y) = \sum_y P(y)\, H(X \mid y) = \sum_{x,y} P(x, y) \log {}^1\!/_{P(x \mid y)}$$

Conditional Entropy (2)

Similarly

$$H(Y \mid X) = \sum_{x,y} P(x,y) \log {}^{1}/_{P(y \mid x)}$$

is the average uncertainty about the output that the sender has, given that she knows what she sent over the channel.

Intuitively this should be less than the average surprise that the receiver will experience, $H(Y)$.

Conditional Entropy (3)

The *chain rule* for entropy:

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

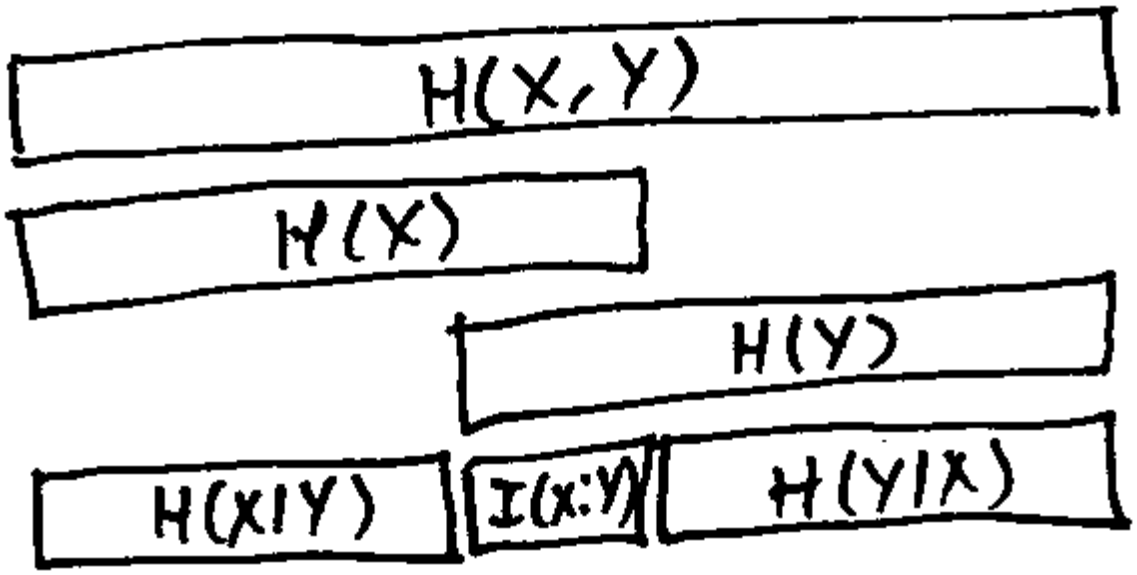
"The average coding cost of a pair is the same regardless of whether you treat them as a joint event, or code one and then the other."

Proof:

$$H(X,Y) = \sum_x \sum_y p(x)\, p(y \mid x) \left[\log \frac{1}{p(x)} + \log \frac{1}{p(y \mid x)} \right]$$

$$= \sum_x p(x) \log \frac{1}{p(x)} \cancelto{1}{\sum_y p(y \mid x)} + \sum_x \sum_y p(x,y) \log \frac{1}{p(y \mid x)}$$

Mutual Information (2)



The receiver thinks: $I(X;Y) = H(X) - H(X \mid Y)$

The mutual information is, on average, the information content of the input minus the part that is still uncertain after seeing the output. That is, the average information that we can get about the input over the channel.

$I(X;Y) = H(Y) - H(Y \mid X)$ is often easier to calculate

The Capacity

Where are we going?

$I(X;Y)$ depends on the channel and input distribution $\mathbf{p}_X$

The Capacity: $C(Q) = \max_{\mathbf{p}_X} I(X;Y)$

C gives the maximum average amount of information we can get in one use of the channel.

We will see that reliable communication is possible at C bits per channel use.

Lots of new definitions

When dealing with extended ensembles, independent identical copies of an ensemble, entropies were easy: $H(X^K) = K\,H(X)$.

Dealing with channels forces us to extend our notions of information to collections of dependent variables. For every joint, conditional and marginal probability we have a different entropy and we'll want to understand their relationships.

Unfortunately this meant seeing a lot of definitions at once.
They are summarized on pp138–139 of MacKay. And also in the following tables.

The probabilities associated with a channel

Very little of this is special to channels, it's mostly results for any pair of dependent random variables.

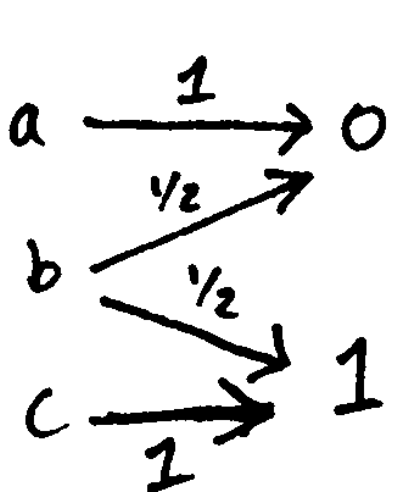
Distribution	Where from?	Interpretation / Name
$P(x)$	We choose	Input distribution
$P(y \mid x)$	Q, channel definition	Channel noise model Sender's beliefs about output
$P(x, y)$	$p(y \mid x)\, p(x)$	Omniscient outside observer's joint distribution
$P(y)$	$\sum_x p(x, y) = Q\, \mathbf{p}_X$	(Marginal) output distribution
$P(x \mid y)$	$p(y \mid x)\, p(x)/p(y)$	Receiver's beliefs about input. "Inference"

Corresponding information measures

$H(X)$	$\sum_x p(x) \log {}^1\!/_{p(x)}$	Ave. info. content of source
		Sender's ave. surprise on seeing x
$H(Y)$	$\sum_y p(y) \log {}^1\!/_{p(y)}$	Ave. info. content of output
		Partial info. about x *and* noise
		Ave. surprise of receiver
$H(X,Y)$	$\sum_{x,y} p(x,y) \log {}^1\!/_{p(x,y)}$	Ave. info. content of (x,y)
		or "source and noise".
		Ave. surprise of outsider
$H(X \mid y)$	$\sum_x p(x \mid y) \log {}^1\!/_{p(x \mid y)}$	Uncertainty after seeing output
$H(X \mid Y)$	$\sum_{x,y} p(x,y) \log {}^1\!/_{p(x \mid y)}$	Average, $\mathbb{E}_{p(y)}[H(X \mid y)]$
$H(Y \mid X)$	$\sum_{x,y} p(x,y) \log {}^1\!/_{p(y \mid x)}$	Sender's ave. uncertainty about y
$I(X;Y)$	$H(X)+H(Y)-H(X,Y)$	'Overlap' in ave. info. contents
	$H(X) - H(X \mid Y)$	Ave. uncertainty reduction by y
		Ave info. about x over channel.
	$H(Y) - H(Y \mid X)$	Often easier to calculate

And review the diagram relating all these quantities!

Ternary confusion channel



$$Q = \begin{bmatrix} 1 & 1/2 & 0 \\ 0 & 1/2 & 1 \end{bmatrix}$$

Assume $\mathbf{p}_X = [1/3, 1/3, 1/3]$. What is $I(X;Y)$?

$$H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) = 1 - 1/3 = 2/3$$

Optimal input distribution: $\mathbf{p}_X = [1/2, 0, 1/2]$

For which $I(X;Y) = 1$, the *capacity* of the channel.