Information Theory and Networks
Lecture 23: Channel Information Capacity

Matthew Roughan
<matthew.roughan@adelaide.edu.au>
http://www.maths.adelaide.edu.au/matthew.roughan/Lecture_notes/InformationTheory/

School of Mathematical Sciences,
University of Adelaide

October 8, 2013

Part I

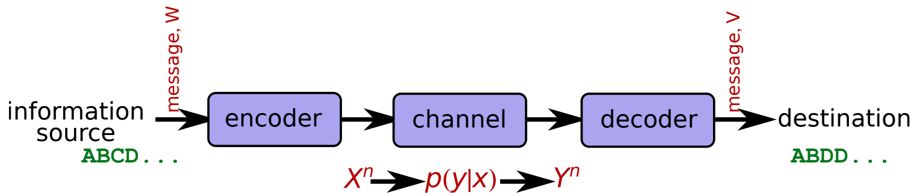
Channel Information Capacity

> To make no mistakes is not in the power of man; but from their errors and mistakes the wise and good learn wisdom for the future.
> *Plutarch*

Section 1

Information Capacity

Digital Communications Channels

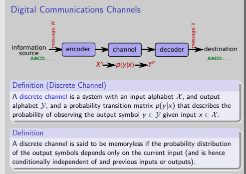


Definition (Discrete Channel)

A discrete channel is a system with an input alphabet $\mathcal{X}$, and output alphabet $\mathcal{Y}$, and a probability transition matrix $p(y|x)$ that describes the probability of observing the output symbol $y \in \mathcal{Y}$ given input $x \in \mathcal{X}$.

Definition

A discrete channel is said to be memoryless if the probability distribution of the output symbols depends only on the current input (and is hence conditionally independent of and previous inputs or outputs).



$p(y|x)$ is a little like the transition matrix in a Markov chain, but

1. the input and output states don't have to be the same set
2. we only go through one step, so there is no "chain"

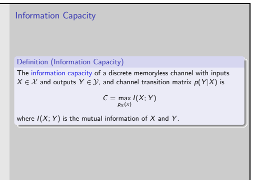
Information Capacity

Definition (Information Capacity)

The information capacity of a discrete memoryless channel with inputs $X \in \mathcal{X}$ and outputs $Y \in \mathcal{Y}$, and channel transition matrix $p(Y|X)$ is

$$C = \max_{p_X(x)} I(X;Y)$$

where $I(X;Y)$ is the mutual information of X and Y.



We will soon learn that information capacity and operational capacity are the same, so we will just call them **channel capacity**.

Reminder: mutual information is defined to be

$$\begin{aligned} I(X;Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p_X(x)p_Y(y)} \\ &= D\big(p(x,y) \big\| q(x,y)\big) \\ &= E\left[\log \frac{p(X|Y)}{p(X)}\right], \end{aligned}$$

where $q(x,y) = p_X(x)p_Y(y)$, where $p_X(x)$ and $p_Y(y)$ are the marginal distributions of X and Y respectively. Remember also that

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

Example 1: Binary Symmetric Channel

$$P(Y|X) = \begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix}$$

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_x p_X(x) H(Y|X=x) \\ &= H(Y) - \sum_x p_X(x) H(\alpha) \\ &= H(Y) - H(\alpha) \\ &\le 1 - H(\alpha) \end{aligned}$$

because Y is a binary random variable. Hence

$$C \le 1 - H(\alpha) \text{ bits}$$

We'll do a bit more on this in a moment, but for the moment note the extreme cases:

- When $\alpha = 0$, the channel is noiseless, and $C = 1$ (i.e., we can send 1 bit per symbol)
- When $\alpha = 1/2$, then $H(\alpha) = 1$ and the bound implies $C = 0$, which should be obvious as when $\alpha = 1/2$ we learn nothing about the input from each output symbol.

Example 2: Binary Erasure Channel

$$P(Y|X) = \begin{pmatrix} 1-\alpha & \alpha & 0 \\ 0 & \alpha & 1-\alpha \end{pmatrix}$$

- The probability we end up in state ? is α regardless of $P_X(x)$, so $H(Y|X=0) = H(Y|X=1) = H(\alpha)$.
- The entropy $H(Y)$ will be

  $$H(Y) = -p_Y(0)\log_2 p_Y(0) - p_Y(1)\log_2 p_Y(1) - \alpha \log_2 \alpha$$

  as for entropy of Bernoulli, this is maximised when $p_Y(0) = p_Y(1)$, which requires they both are $= (1-\alpha)/2$

  $$H(Y) = -(1-\alpha)[\log_2(1/2) + \log_2(1-\alpha)] - \alpha \log_2 \alpha = (1-\alpha)H(1/2) + H(\alpha)$$

- So the capacity will be

  $$\begin{aligned} C &= \max_{p_X(x)} H(Y) - H(Y|X) \\ &= (1-\alpha)H(1/2) + H(\alpha) - H(\alpha) \\ &= 1 - \alpha \end{aligned}$$

- This capacity makes sense, as we are loosing capacity in proportion to the probability symbols are erased
- But it isn't immediately obvious that we could achieve this rate without loss of information
- One approach is to use feedback (see earlier to see it can achieve this rate)
- It turns out we can achieve this even without feedback

Example 3: Non-Overlapping Output

$$P(Y|X) = \begin{pmatrix} 2/3 & 1/3 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$

- The channel appears to be noisy, but isn't really
  - The input symbol is determined by the output
  - So $H(X|Y) = 0$
- So we get

$$C = \max_{p_X} I(X;Y) = \max_{p_X} H(X) - H(X|Y) = \max_{p_X} H(X)$$

which we get for the uniform distribution over X, so

$$C = H(1/2) = 1 \text{ bit}$$

Example 5: Z Channel

$$P(Y|X) = \begin{pmatrix} 1 & 0 \\ \alpha & 1-\alpha \end{pmatrix}$$

Conditional entropy

$$H(Y|X) = p_X(1) H(\alpha)$$

Entropy

$$H(Y) = H\big(p_X(1)(1-\alpha)\big)$$

Capacity

$$\begin{aligned} C &= \max_{p_X(x)} H(Y) - H(Y|X) \\ &= \max_{p_X(x)} H\big(p_X(1)(1-\alpha)\big) - p_X(1) H(\alpha) \end{aligned}$$

From the definition of $H(Y|X)$

$$\begin{aligned} H(Y|X) &= p_X(0)\,[p(0|0)\log_2 p(0|0) + p(1|0)\log_2 p(1|0)] \\ &\quad + p_X(1)\,[p(0|1)\log_2 p(0|0) + p(1|1)\log_2 p(1|0)] \\ &= p_X(0)\,[1\log_2 1 + 0]\; p_X(1)\,[\alpha \log_2 \alpha + (1-\alpha)\log_2(1-\alpha)] \\ &= p_X(1) H(\alpha) \end{aligned}$$

and entropy WRT the output symbols is

$$\begin{aligned} H(Y) &= p_Y(0)\log_2 p_Y(0) + p_Y(1)\log_2 p_Y(1) \\ &= (p_X(0) + \alpha p_X(1))\log_2(p_X(0) + \alpha p_X(1)) \\ &\quad + (1-\alpha)p_X(1)\log_2(1-\alpha)p_X(1) \\ &= (1-(1-\alpha)p_X(1))\log_2(1-(1-\alpha)p_X(1)) \\ &\quad + (1-\alpha)p_X(1)\log_2(1-\alpha)p_X(1) \\ &= H\big(p_X(1)(1-\alpha)\big) \end{aligned}$$

Example 5: Z Channel

Take

$$
\begin{aligned}
c(p) &= H\big(p(1-\alpha)\big) - pH(\alpha) \\
\frac{dc}{dp} &= (1-\alpha)H'\big(p(1-\alpha)\big) - H(\alpha) \\
&= (1-\alpha)\log\frac{p(1-\alpha)}{1-p(1-\alpha)} - H(\alpha)
\end{aligned}
$$

We maximise $c(p)$ when $dc/dp = 0$, so

$$
\begin{aligned}
(1-\alpha)\log\frac{p(1-\alpha)}{1-p(1-\alpha)} &= H(\alpha) \\
\frac{p(1-\alpha)}{1-p(1-\alpha)} &= 2^{H(\alpha)/(1-\alpha)} \\
p &= \frac{1}{(1-\alpha)\left(1+2^{H(\alpha)/(1-\alpha)}\right)}
\end{aligned}
$$

$$
\begin{aligned}
H'(\alpha) &= \frac{d}{d\alpha}\left[\alpha\log\alpha + (1-\alpha)\log(1-\alpha)\right] \\
&= \log\alpha + 1 - \log(1-\alpha) - 1 \\
&= \log\alpha - \log(1-\alpha) \\
&= \log\frac{\alpha}{1-\alpha}
\end{aligned}
$$

Example 5: Z Channel (small α approximation)

Capacity

$$
C = \max_{p_X(x)} H\big(p_X(1)(1-\alpha)\big) - p_X(1)H(\alpha)
$$

which is maximised when

$$
p_X(1) = \frac{1}{(1-\alpha)\left(1+2^{H(\alpha)/(1-\alpha)}\right)}
$$

For small α (small error probability) C can be approximated by

$$
C \simeq 1 - 0.5H(\alpha)
$$

Compare this to the binary symmetric channel with

$$
C \leq 1 - H(\alpha)
$$

Section 2

Symmetry

Section 2

Symmetry

Symmetric Channels

Definition (Symmetric Channel)

We say a channel is symmetric if the rows and columns of the channel transition matrix are permutations of each other.
It is said to be weakly symmetric if every row is a permutation of the others, and all the column sums $\sum_x p(y|x)$ are equal.

Symmetric Channels

Definition (Symmetric Channel)
We say a channel is symmetric if the rows and columns of the channel transition matrix are permutations of each other.
It is said to be weakly symmetric if every row is a permutation of the others, and all the column sums $\sum_x p(y|x)$ are equal.

Example 1: Binary Symmetric Channel

$$P(Y|X) = \begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix}$$

This example is symmetric

- we can get either row or column by a permutation of $(\alpha, 1-\alpha)$

Example 1: Binary Symmetric Channel
$P(Y|X) = \begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix}$
This example is symmetric
- we can get either row or column by a permutation of $(\alpha, 1-\alpha)$

Example 2: Binary Erasure Channel

$$P(Y|X) = \begin{pmatrix} 1-\alpha & \alpha & 0 \\ 0 & \alpha & 1-\alpha \end{pmatrix}$$

This example not symmetric

- we just have to look at column sums, which are not equal

Example 2: Binary Erasure Channel
$P(Y|X) = \begin{pmatrix} 1-\alpha & \alpha & 0 \\ 0 & \alpha & 1-\alpha \end{pmatrix}$
This example not symmetric
- we just have to look at column sums, which are not equal

Example 3: Non-Overlapping Output

$$P(Y|X) = \begin{pmatrix} 2/3 & 1/3 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$

This example not symmetric

- we just have to look at column sums, which are not equal

Example 3: Non-Overlapping Output
$P(Y|X) = \begin{pmatrix} 2/3 & 1/3 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$
This example not symmetric
- we just have to look at column sums, which are not equal

Example 4: Noisy Typewriter

$$P(Y|X) = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & \cdots & 0 & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & \cdots & 0 & 0 & 0 \\ & & \ddots & \ddots & \ddots & \ddots & & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & 0 & 0 & 0 & 0 & \cdots & 0 & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

This example is symmetric

- we can get either row or column by a permutation of $(1/3, 1/3, 1/3, 0, \ldots, 0)$

Example 4: Noisy Typewriter
This example is symmetric
- we can get either row or column by a permutation of (1/3,1/3,1/3,0,...,0)

Example 5: Z Channel

$$P(Y|X) = \begin{pmatrix} 1 & 0 \\ \alpha & 1-\alpha \end{pmatrix}$$

This example not symmetric

- we just have to look at column sums, which are not equal

Example 5: Z Channel
$P(Y|X) = \begin{pmatrix} 1 & 0 \\ \alpha & 1-\alpha \end{pmatrix}$
This example not symmetric
- we just have to look at column sums, which are not equal

Example 6

$$P(Y|X) = \begin{pmatrix} 1/3 & 1/6 & 1/2 \\ 1/3 & 1/2 & 1/6 \end{pmatrix}$$

This example is only weakly symmetric

- the rows are all a permutation of $(1/3, 1/6, 1/2)$
- the columns are not all permutations of each other
- but the columns all sum to $2/3$

Example 6
$P(Y|X) = \begin{pmatrix} 1/3 & 1/6 & 1/2 \\ 1/3 & 1/2 & 1/6 \end{pmatrix}$
This example is only weakly symmetric
- the rows are all a permutation of $(1/3, 1/6, 1/2)$
- the columns are not all permutations of each other
- but the columns all sum to $2/3$

[CT91, p.190]

Theorem

For a weakly symmetric channel

$$C = \log |\mathcal{Y}| - H(\mathbf{r})$$

where $\mathbf{r}$ *is any row of the channel transition matrix.*
This capacity is achieved on a uniform distribution on the input alphabet.

Proof.

First note that the entropy of a permuted PMF is (by our Axioms) unchanged, so $H(\mathbf{r})$ will be the same for any row $\mathbf{r}$ of a weakly symmetric channel.
Now remember that

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(\mathbf{r}) \\ &\leq \log |\mathcal{Y}| - H(\mathbf{r}) \end{aligned}$$

with equality only if the output is uniform.

[CT91, p.190]

Proof.

Now note that if the PMF for X is uniform, then

$$p_X(x) = \frac{1}{|\mathcal{X}|}$$

and from the Law of Total Probability

$$\begin{aligned} p_Y(y) &= \sum_x p(y|x) p_X(x) \\ &= \frac{1}{|\mathcal{X}|} \sum_x p(y|x) \\ &= \frac{1}{|\mathcal{X}|} c \\ &= \frac{1}{|\mathcal{Y}|} \end{aligned}$$

where $c = \sum_x p(y|x)$ is guaranteed by weak symmetry. □

So we see that uniformity of the input implies uniformity for the output of a weakly symmetric system.

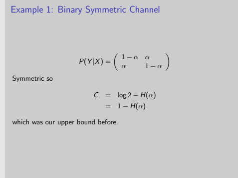
Example 1: Binary Symmetric Channel

$$P(Y|X) = \begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix}$$

Symmetric so

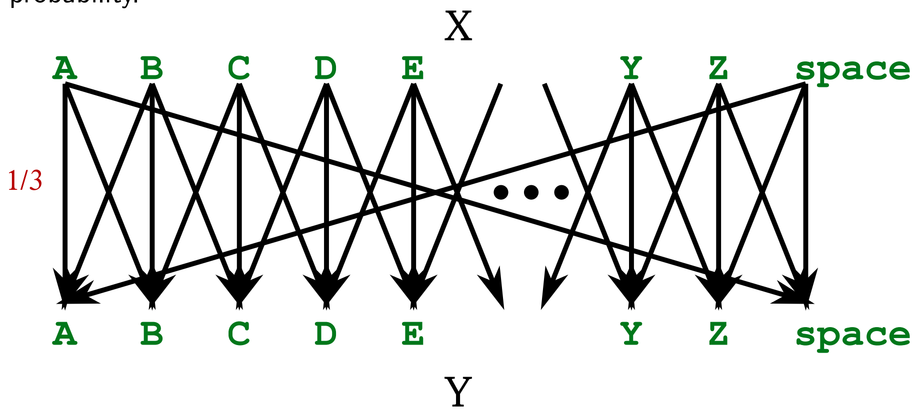
$$\begin{aligned} C &= \log 2 - H(\alpha) \\ &= 1 - H(\alpha) \end{aligned}$$

which was our upper bound before.



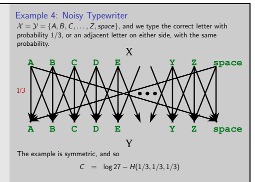
Example 4: Noisy Typewriter

$\mathcal{X} = \mathcal{Y} = \{A, B, C, \ldots, Z, space\}$, and we type the correct letter with probability $1/3$, or an adjacent letter on either side, with the same probability.



The example is symmetric, and so

$$C = \log 27 - H(1/3, 1/3, 1/3)$$



[CT91, 8.1.3, pp.185-186] or [Mac11, p.148].

$$P(Y|X) = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & \cdots & 0 & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & \cdots & 0 & 0 & 0 \\ & & \ddots & \ddots & \ddots & \ddots & & & \\ & & & \ddots & \ddots & \ddots & \ddots & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & 0 & 0 & 0 & 0 & \cdots & 0 & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

Example 6

$$P(Y|X) = \begin{pmatrix} 1/3 & 1/6 & 1/2 \\ 1/3 & 1/2 & 1/6 \end{pmatrix}$$

This example is weakly symmetric so

$$C = \log 3 - H(1/3, 1/2, 1/6)$$

[CT91, p.190]

Section 3

Other Properties of Channel Capacity

Bounds

$$0 \leq C \leq \min \Big[\log |\mathcal{X}|, \log |\mathcal{Y}| \Big]$$

- The lower bound arise because $I(X;Y) \geq 0$
- The upper bound arises because

$$C = \max I(X;Y) \leq \max H(X) = \log |\mathcal{X}|$$

and

$$C = \max I(X;Y) \leq \max H(Y) = \log |\mathcal{Y}|$$

Does a C always exist?

Remember that

- $I(X;Y)$ is a continuous function of $p_X(x)$
- $I(X;Y)$ is a concave function of $p_X(x)$ (for fixed $p(y|x)$)
- As noted above $I(X;Y)$ is bounded above

Given these condition, a local maximum is always a global maximum, and given it is finite we don't have to talk about the supremum.

Can we find C?

Obviously, finding it could be hard analytically, but it is numerically tractable:

- $-C$ is convex
- standard restrictions on probabilities are linear

$$p_i \geq 0 \text{ and } \sum p_i = 1$$

This allows standard convex optimisation approaches:

- Karush-Kuhn-Tucker conditions;
- Gradient projection algorithm.

Can we find C?
Obviously, finding it could be hard analytically, but it is numerically tractable:
- $-C$ is convex
- standard restrictions on probabilities are linear
$p_i \geq 0$ and $\sum p_i = 1$
This allows standard convex optimisation approaches:
- Karush-Kuhn-Tucker conditions;
- Gradient projection algorithm.

Further reading I

Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, John Wiley and Sons, 1991.

David J. MacKay, *Information theory, inference, and learning algorithms*, Cambridge University Press, 2011.