# Formal Modeling in Cognitive Science

Lecture 29: Noisy Channel Model and Applications; Kullback-Leibler Divergence; Cross-entropy

Frank Keller

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

March 14, 2006









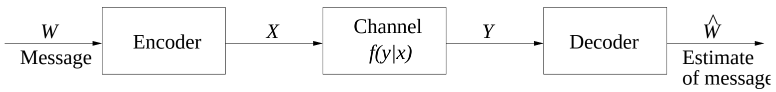


## Noisy Channel Model

So far, we have looked at encoding a message efficiently, put what about *transmitting* the message?

The transmission of a message can be modeled using a *noisy channel:*

- a message $W$ is encoded, resulting in a string $X$;
- $X$ is transmitted through a channel with the probability distribution $f(y|x)$;
- the resulting string $Y$ is decoded, yielding an estimate of the message $\hat{W}$.








## Channel Capacity

We are interested in the mathematical properties of the channel used to transmit the message, and in particular in its capacity.

**Definition: Discrete Channel**

A discrete channel consists of an input alphabet $X$, an output alphabet $Y$ and a probability distribution $f(y|x)$ that expresses the probability of observing symbol $y$ given that symbol $x$ is sent.

**Definition: Channel Capacity**

The channel capacity of a discrete channel is:

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

The capacity of a channel is the maximum of the mutual information of $X$ and $Y$ over all input distributions $f(x)$.

## Channel Capacity

### Example: Noiseless Binary Channel

Assume a binary channel whose input is reproduced exactly at the output. Each transmitted bit is received without error:


The channel capacity of this channel is:

$$C = \max_{f(x)} I(X;Y) = 1 \text{ bit}$$

This maximum is achieved with $f(0) = \frac{1}{2}$ and $f(1) = \frac{1}{2}$.

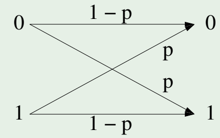




## Channel Capacity

### Example: Binary Symmetric Channel

Assume a binary channel whose input is flipped (0 transmitted a 1 or 1 transmitted as 0) with probability $p$:




The mutual information of this channel is bounded by:

$$\begin{aligned} I(X;Y) &= H(Y) - H(X|Y) = H(Y) - \sum_x f(x)H(Y|X=x) \\ &= H(Y) - \sum_x f(x)H(p) = H(Y) - H(p) \leq 1 - H(p) \end{aligned}$$

The channel capacity is therefore:

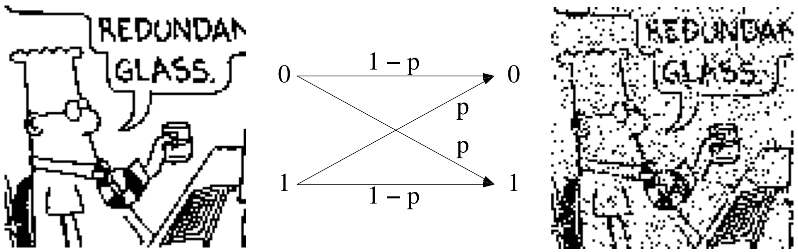
$$C = \max_{f(x)} I(X;Y) = 1 - H(p) \text{ bits}$$





## Channel Capacity

A binary data sequence of length 10,000 transmitted over a binary symmetric channel with $p = 0.1$:








## Properties Channel Capacity

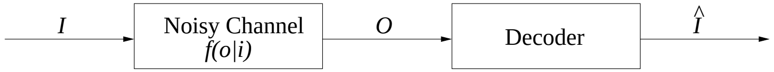
### Theorem: Properties of Channel Capacity

1. $C \geq 0$ since $I(X;Y) \geq 0$;
2. $C \leq \log |X|$, since $C = \max I(X;Y) \leq \max H(X) \leq \log |X|$;
3. $C \leq \log |Y|$ for the same reason.

# Applications of the Noisy Channel Model

The noisy channel can be applied to decoding processes involving linguistic information. A typical formulation of such a problem is:

- we start with a linguistic input $I$;
- $I$ is transmitted through a noisy channel with the probability distribution $f(o|i)$;
- the resulting output $O$ is decoded, yielding an estimate of the input $\hat{I}$.








# Applications of the Noisy Channel Model

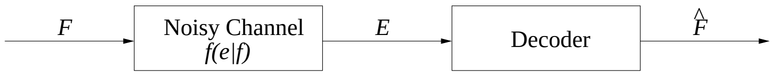
| Application | Input | Output | $f(i)$ | $f(o\|i)$ |
|---|---|---|---|---|
| Machine translation | target language word sequences | source language word sequences | target language model | translation model |
| Optical character recognition | actual text | text with mistakes | language model | model of OCR errors |
| Part of speech tagging | POS sequences | word sequences | probability of POS sequences | $f(w\|t)$ |
| Speech recognition | word sequences | speech signal | language model | acoustic model |





# Applications of the Noisy Channel Model

Let's look at machine translation in more detail. Assume that the French text ($F$) passed through a noisy channel and came out as English ($E$). We decode it to estimate the original French ($\hat{F}$):




We compute $\hat{F}$ using Bayes' theorem:

$$\hat{F} = \arg\max_f f(f|e) = \arg\max_f \frac{f(f)f(e|f)}{f(e)} = \arg\max_f f(f)f(e|f)$$

Here $f(e|f)$ is the translation model, $f(f)$ is the French language model, and $f(e)$ is the English language model (constant).





# Example Output: Spanish–English

we all know very well that the current treaties are insufficient and that , in the future , it will be necessary to develop a better structure and different for the european union , a structure more constitutional also make it clear what the competences of the member states and which belong to the union . messages of concern in the first place just before the economic and social problems for the present situation , and in spite of sustained growth , as a result of years of effort on the part of our citizens . the current situation , unsustainable above all for many self-employed drivers and in the area of agriculture , we must improve without doubt . in itself , it is good to reach an agreement on procedures , but we have to ensure that this system is not likely to be used as a weapon policy . now they are also clear rights to be respected . i agree with the signal warning against the return , which some are tempted to the intergovernmental methods . there are many of us that we want a federation of nation states .

## Example Output: Finnish–English

the rapporteurs have drawn attention to the quality of the debate and also the need to go further : of course , i can only agree with them . we know very well that the current treaties are not enough and that in future , it is necessary to develop a better structure for the union and , therefore perustuslaillisempi structure , which also expressed more clearly what the member states and the union is concerned . first of all , kohtaamiemme economic and social difficulties , there is concern , even if growth is sustainable and the result of the efforts of all , on the part of our citizens . the current situation , which is unacceptable , in particular , for many carriers and responsible for agriculture , is in any case , to be improved . agreement on procedures in itself is a good thing , but there is a need to ensure that the system cannot be used as a political lyomaaseena . they also have a clear picture of the rights of now , in which they have to work . i agree with him when he warned of the consenting to return to intergovernmental methods . many of us want of a federal state of the national member states .





## Kullback-Leibler Divergence

**Definition: Kullback-Leibler Divergence**

For two probability distributions $f(x)$ and $g(x)$ for a random variable $X$, the Kullback-Leibler divergence or relative entropy is given as:

$$D(f||g) = \sum_{x \in X} f(x) \log \frac{f(x)}{g(x)}$$

The KL divergence compares the entropy of two distributions over the same random variable.

Intuitively, the KL divergence number of additional bits required when encoding a random variable with a distribution $f(x)$ using the alternative distribution $g(x)$.





## Kullback-Leibler Divergence

**Theorem: Properties of the Kullback-Leibler Divergence**

1. $D(f||g) \geq 0$;
2. $D(f||g) = 0$ iff $f(x) = g(x)$ for all $x \in X$;
3. $D(f||g) \neq D(g||f)$;
4. $I(X;Y) = D(f(x,y)||f(x)f(y))$.

So the mutual information is the KL divergence between $f(x,y)$ and $f(x)f(y)$. It measures how far a distribution is from independence.





## Kullback-Leibler Divergence

**Example**

For a random variable $X = \{0, 1\}$ assume two distributions $f(x)$ and $g(x)$ with $f(0) = 1 - r$, $f(1) = r$ and $g(0) = 1 - s$, $g(1) = s$:

$$\begin{aligned} D(f||g) &= (1-r)\log\frac{1-r}{1-s} + r\log\frac{r}{s} \\ D(g||f) &= (1-s)\log\frac{1-s}{1-r} + s\log\frac{s}{r} \end{aligned}$$

If $r = s$ then $D(f||g) = D(g||f) = 0$. If $r = \frac{1}{2}$ and $r = \frac{1}{4}$:

$$\begin{aligned} D(f||g) &= \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{3}{4}} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{4}} = 0.2075 \\ D(g||f) &= \frac{3}{4}\log\frac{\frac{3}{4}}{\frac{1}{2}} + \frac{1}{4}\log\frac{\frac{1}{4}}{\frac{1}{2}} = 0.1887 \end{aligned}$$

## Cross-entropy

### Definition: Cross-entropy

For a random variable $X$ with the probability distribution $f(x)$ the cross-entropy for the probability distribution $g(x)$ is given as:

$$H(X,g) = -\sum_{x \in X} f(x) \log g(x)$$

The cross-entropy can also be expressed in terms of entropy and KL divergence:

$$H(X,g) = H(X) + D(f||g)$$

Intuitively, the cross-entropy is the total number of bits required when encoding a random variable with a distribution $f(x)$ using the alternative distribution $g(x)$.





## Cross-entropy

### Example

In the last lecture, we constructed a code for the following distribution using Huffman coding:

| $x$ | a | e | i | o | u |
|---|---|---|---|---|---|
| $f(x)$ | 0.12 | 0.42 | 0.09 | 0.30 | 0.07 |
| $-\log f(x)$ | 3.06 | 1.25 | 3.47 | 1.74 | 3.84 |

The entropy of this distribution is $H(X) = 1.995$. Now compute the distribution $g(x) = 2^{-l(x)}$ associated with the Huffman code:

| $x$ | a | e | i | o | u |
|---|---|---|---|---|---|
| $C(x)$ | 001 | 1 | 0001 | 01 | 0000 |
| $l(x)$ | 3 | 1 | 4 | 2 | 4 |
| $g(x)$ | $\frac{1}{8}$ | $\frac{1}{2}$ | $\frac{1}{16}$ | $\frac{1}{4}$ | $\frac{1}{16}$ |





## Cross-entropy

### Example

Then the cross-entropy for $g(x)$ is:

$$\begin{aligned} H(X,g) &= -\textstyle\sum_{x \in X} f(x) \log g(x) \\ &= -(0.12 \log \tfrac{1}{8} + 0.43 \log \tfrac{1}{2} + 0.09 \log \tfrac{1}{16} + 0.30 \log \tfrac{1}{4} \\ &\quad + 0.07 \log \tfrac{1}{16}) \\ &= 2.030 \end{aligned}$$

The KL divergence is:

$$D(f||g) = H(X,g) - H(X) = 0.035$$

This means we are losing on average 0.035 bits by using the Huffman code rather then the theoretically optimal code given by the Shannon information.





## Summary

- The noisy channel can model the errors and loss when transmitting a message with input $X$ and output $Y$;
- the capacity of the channel is given by the maximum of the mutual information of $X$ and $Y$;
- a binary symmetric channel is one where each bit is flipped with probability $p$;
- the noisy channel model can be applied to linguistic problems, e.g., machine translation;
- the Kullback-Leibler divergence is the distance between two distributions (the cost of encoding $f(x)$ through $g(x)$).