So far, we have looked at encoding a message efficiently, but 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 $f(x)$. 
**Example: Noisy Binary Channel**

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

\[
\begin{array}{ccc}
0 & 1-p & 0 \\
1-p & 0 & 1 \\
p & p & 1
\end{array}
\]

The mutual information of this channel is bounded by:

\[
I(X; Y) = H(Y) - H(X | Y) = H(Y) - \sum_x f(x)H(Y | X = x) \\
= H(Y) - \sum_x f(x)H(p) = H(Y) - H(p) \leq 1 - H(p)
\]

The channel capacity is therefore:

\[
C = \max I(X; Y) = 1 - H(p) \text{ bits}
\]
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 $\hat{I}$.

$$\begin{align*}
I & \xrightarrow{\text{Noisy Channel}} O & \xrightarrow{\text{Decoder}} \hat{I}
\end{align*}$$

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 ($\hat{F}$):

$$\begin{align*}
F & \xrightarrow{\text{Noisy Channel}} E & \xrightarrow{\text{Decoder}} \hat{F}
\end{align*}$$

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

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

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.
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, perustuslailisiemi structure, which also expressed more clearly what the member states and the union is concerned. First of all, kohtamiemme 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.
**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 $g(x)$.

**Example**

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

<table>
<thead>
<tr>
<th>x</th>
<th>a</th>
<th>e</th>
<th>i</th>
<th>o</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(x)$</td>
<td>0.12</td>
<td>0.42</td>
<td>0.09</td>
<td>0.30</td>
<td>0.07</td>
</tr>
<tr>
<td>$- \log f(x)$</td>
<td>3.06</td>
<td>1.25</td>
<td>3.47</td>
<td>1.74</td>
<td>3.84</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>x</th>
<th>a</th>
<th>e</th>
<th>i</th>
<th>o</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C(x)$</td>
<td>001</td>
<td>1</td>
<td>0001</td>
<td>01</td>
<td>0000</td>
</tr>
<tr>
<td>$l(x)$</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>$g(x)$</td>
<td>$\frac{1}{8}$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{16}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{16}$</td>
</tr>
</tbody>
</table>

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 than 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 $g(x)$).