Chapter 4

The Infomax principle: Maximization of Mutual Information

The approach presented here could be paraphrased under the motto "The brain has to process information, thus evolution will have taken care that it is as optimal in the sense of information theory as possible", roots back on the initiative of Linsker (1986, 1988, 1989). The approach is the application of information theory (Shannon 1948, see also Henze Homuth 1970) on exemplaric architectures of neural networks.

4.1 Introduction: Information entropie and Mutual Information

Many technical, but also biological systems can be described as information-transmissing or information-processing systems. By means of Shannon's information entropy (or mutual information, MI) one can "measure" the content of information, such that a quantitative description – including a description by extremal principles – becomes possible.

Such an *information-theoretical approach* can be applied to all systems that allow for a formal decomposition into Sender – Channel – Receiver, and for which the information-processing channel can be fully described by a transition probability matrix as follows.

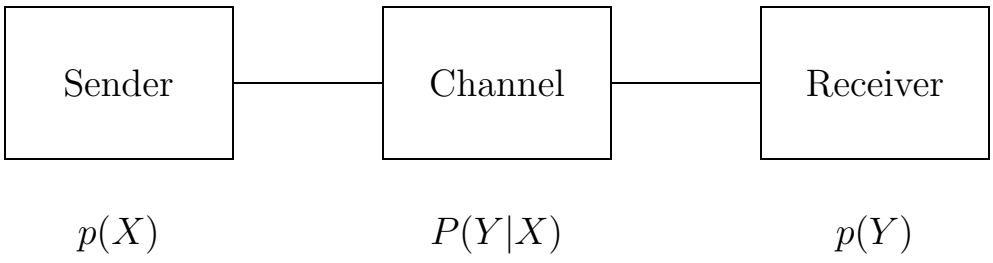

Figure 4.1: The formal decomposition Sender-Channel-Receiver

4.1.1 Description of an information-processing system

We consider a given sender with a finite character set or pattern set ("alphabet") x_i, and the corresponding probability density $p(x_i)$. Correspondingly, the receiver has an alphabet y_i, and the transmission channel is assumed to be characterized by the conditional probability $P(y_j|x_i)$, that – given the letter x_i is sent – the letter y_j is received. Hence $P(y_j|x_i)$ is the probability matrix describing the transfer through the channel. As we assume that for each letter x_i sent by the sender, for sure some letter is received by the sender, we have the normalization condition $\sum_j P(y_j|x_i) = 1$. Then the (joint) probability that in the system sender-receiver the pair of letters (x_i, y_j) appears, is just given by

$$p(x_i, y_j) = P(y_j|x_i) \cdot p(x_i). \tag{4.1}$$

Collecting together, the total probability for receipt of letter y_i is given by

$$p(y_j) = \sum_i p(x_i, y_j) = \sum_i P(y_j|x_i) \cdot p(x_i). \tag{4.2}$$

4.1.2 Entropy and Information

If a physical systems assumes each of 2^n states with equal probability, i.e., with probability $1/2^n$, its entropy is given by

$$S = -k_B \sum_{i=1}^{2^n} \rho_i \ln \rho_i = -k_B \sum_{i=1}^{2^n} \frac{1}{2^n} \ln \frac{1}{2^n} = n \cdot (k_B \ln 2). \tag{4.3}$$

Here $\rho_i = 1/2^n$ are the coefficients of the (quantum mechanical) density matrix and k_B is the Boltzmann constant. The entropie therefore is proportional to the number of yes/no questions one would need to specify the system's state. To localize, say, the position of a player on a chessboard, one needs 6, as $2^6 = 64$.

To quantifiy information in general, one defones for an arbitrary distribution $p(x_i)$ the Shannon entropy

$$I^{Sh}(X) = -\sum_i p(x_i) \mathrm{ld} p(x_i),$$

which hovever, apart from a factor ($I(X) = \ln 2 \cdot I^{Sh}(X)$) coincides with the expression

$$I(X) := -\sum_i p(x_i) \ln p(x_i) \tag{4.4}$$

which we will use in the remainder. In analogy, one can define

$$I(Y) := -\sum_j p(y_j) \ln p(y_j) \tag{4.5}$$

$$I(X, Y) := -\sum_{i,j} p(x_i, y_j) \ln p(x_i, y_j) \tag{4.6}$$

4.1.3 The Transinformation (Mutual Information)

The Transinformation is defined as follows:

$$T(X, Y) := I(X) + I(Y) - I(X, Y). \tag{4.7}$$

To fully acknowledge the importance of this expression, we consider two limiting cases:

Ideal Channel: If input alphabet and output alphabet coincide ($\forall_i x_i = y_i$), an ideal (working without errors) channel has the matrix $P(y_j|x_i) = \delta_{ij}$. Therefore $p(x_i, y_j) = \delta_{ij}p(x_i)$, and herefrom:

$$I(X,Y) = -\sum_{i,j} \delta_{ij}p(x_i) \ln \delta_{ij}p(x_i) = -\sum_j p(x_j) \ln p(x_j) = I(X) \tag{4.8}$$

(note that $\lim_{x\to 0} x \cdot \ln x = 0$) and

$$p(y_j) = \sum_i p(x_i, y_j) = \sum_i \delta_{ij}p(x_i) = p(x_j), \quad \text{also} \quad I(Y) = I(X). \tag{4.9}$$

The mutual information thus is $T(X,Y) = I(X)$, and equals the information of the input signal.

Totally disturbed channel: Here we assume that the output signal by no means is influenced by the input signal. Such a setup is described by $P(y_j|x_i) = p(y_j)$, being equivalent to $p(x_i, y_j) = p(x_i) \cdot p(y_j)$. Then we have:

$$\begin{aligned} I(X,Y) &= -\sum_{i,j} p(x_i)p(y_j) \ln(p(x_i)p(y_j)) \\ &= -\sum_i p(x_i) \underbrace{(\sum_j p(y_j))}_{=1} \ln p(x_i) - \sum_j p(y_j) \underbrace{(\sum_i p(x_i))}_{=1} \ln p(y_j) \\ &= I(X) + I(Y). \end{aligned} \tag{4.10}$$

Hence the mutual information of a totally disturbed channel is zero.

We can conclude that the transinformation (mutual information) is a measure quantifying the information transmitted by the channel. If one now maximizes (over all possible input distributions), for a given channel $P(y_j|x_i)$ the mutual information, one obtains the *channel capacity* C :

$$C := \max_{\{p(x_i)\}} T(X,Y). \tag{4.11}$$

For a digital system, $C/\ln 2$ equals the numver of binary data lines, or with of the data bus, in bit. Consequently, the channel capacity is the maximal amount of information per bus clock cycle (or transmitted symbol) that can be transferred, if one chooses an optimal input distribution, or, equivalently, chooses an optimal coding language.

4.1.4 Maximization of mutual information

Maximization of mutual information oder maximization of the channel capacity? This is the main issue of choice to be clarified; and it depends on the considered system, which quantity is crucial:

- If the number of "data lines" is crucial, as in the case of technical message transmission, or ling nerve fibres, and on the other hand, a detailed preprocessing or re-coding is feasible, one is going to optimize the channel capacity.
- Otherwise, if the inputs are already specified, as in the retina or other sensory fields, the subsequent processing layer has to organize in a way that the mutual information is maximized.

Both extremal principles always will have to be seen in context with some cost functions for the realizion of implementation, and eventually will turn out to be biologically "unimplementable". if the derived models exhibit parameters or operations that are not compatible with a biological implementation. In such a situation, one has to develop models that come close to the optimum, but are based on experimentally verified interactions.

For a feedforward structure, that processes data from a given input space, the central principle for an informations-theoretical optimum reads:

> For a given alphabet $\{x_i\}, \{y_i\}$ and a given input distribution $p(x_i)$, determine the matrix $P(y_j|x_i)$ in such a way that the mutual information becomes maximized.

Reformulation of the expression for the mutual information:

$$\begin{aligned}
T(X,Y) &= I(X) + I(Y) - I(X,Y) \\
&= -\sum_i \underbrace{(\sum_j P(y_j|x_i))}_{=1} p(x_i) \ln p(x_i) - \sum_{i,j} P(y_j|x_i)p(x_i) \ln[\sum_k P(y_j|x_k)p(x_k)] \\
&\quad + \sum_{i,j} P(y_j|x_i)p(x_i) \ln[P(y_j|x_i)p(x_i)] \\
&= \sum_{i,j} P(y_j|x_i)p(x_i) \left(-\ln p(x_i) - \ln[\sum_k P(y_j|x_k)p(x_k)] + \ln P(y_j|x_i) + \ln p(x_i) \right) \\
&= \sum_i p(x_i) \sum_j P(y_j|x_i) \ln \frac{P(y_j|x_i)}{[\sum_k P(y_j|x_k)p(x_k)]}. \qquad (4.12)
\end{aligned}$$

For continuous patterns (with $\int_{-\infty}^{+\infty} p(X)\mathrm{d}X = 1$) one has:

$$T(X,Y) = \int_{-\infty}^{+\infty} \mathrm{d}X \; p(X) \int_{-\infty}^{+\infty} \mathrm{d}Y \; P(Y|X) \ln \frac{P(Y|X)}{\int_{-\infty}^{+\infty} \mathrm{d}\tilde{X} \; P(Y|\tilde{X})p(\tilde{X})}. \qquad (4.13)$$