Information theory and Visual Plasticity

Nathan Intrator
Computer Science Department
Tel-Aviv University and
Institute for Brain and Neural Systems
Brown University

November, 1999

1 Introduction

The relevance of information theory to neural networks has become more apparent in recent years. This theory has become important in analysing and understanding the nature of the neuronal code that is relayed between cortical layers and the nature of the learning goals that guide neuronal learning and synaptic modification. The rapid advance in recent years of single and multiple electrode recording as well as other non-invasive techniques provides a window on neuronal activity and synaptic modification. However, the puzzle is still unsolved; we do not know what are neuronal learning goals, how they are being incorporated and most importantly, what is the nature of the neuronal code and how it is being formed and interpreted by successive layers.

When we try to understand the nature of synaptic learning rules, we should be concerned with possible goals that may underly synaptic changes. It is conceivable that knowing what could be a useful goal under different input environments could serve for distinguishing between synaptic plasticity theories. Only after this distinction between goals is indicated, can one continue further and distinguish between learning rules aimed at achieving the same objective, on the basis of their detailed mathematical properties, or computational complexity etc. In this paper, we concentrate on the analysis of different neuronal goals and attempt to provide an update on the current related issues and future directions.

2 Brief review on information theory

Information theory was developed about 50 years ago for the study of communication channels (Shannon, 1948). Shannon considered information as a loss of *uncertainty* and gave a definition as a function of the probability distribution of the code-words. If for example, the probability distribution $P(X)$ is concentrated on a single value, then the information we can transmit when choosing values from this distribution is zero since we always transmit the same value. Thus, the amount of information is a function of the variability of the distribution and actually the exact shape of the distribution. This quantity which we denote by $H(X)$ should satisfy an additivity constraint which states that when two random variables are independent, the information contained

in both of them should be the sum of the information contained in each of them, namely

$$P(X_1, X_2) = P_1(X_1)P_2(X_2) \Rightarrow H(X_1, X_2) = H(X_1) + H(X_2). \quad (1)$$

Shannon has shown that the only function that is consistent with this condition and with few other simple constraints is the Boltzmann entropy of statistical mechanics.[1] The entropy[2] in the continuous and discrete cases respectively is given by:

$$H(X) = -\int P(x) \log P(x) dx, \quad H(X) = -\sum_{i=1}^{K} p(x_i) \log p(x_i), \quad (2)$$

where $p(x_i)$ is the probability of observing the value x_i out of a possible K discrete values of the random variable X. An intuitive way to look at this function is by considering the average number of bits that is needed to produce an efficient code; It is desirable to use a small number of bits for sending those words that appear with high probability, and use larger number of bits for sending words that appear with lower probability. In the special case of n words arriving at the same probability, the number of bits that are required for each word is $\log_2 n$.

Shannon formulated this idea for the problem of information flow through a bottleneck, having to optimize the code so as to send the smallest number of bits on average. This led to questions such as how does the receiver, given the transmitted information only, maximize his knowledge about the data available at the sender side. For our purpose, we formulate the mutual information idea in terms of a neural network of a single layer. Let $\mathbf{d}^i \in R^n$ be an input vector to the network occurring with a probability distribution P_d, and let $\mathbf{c}^i \in R^k$ be the corresponding k-dimensional network activity with its probability distribution P_c. The *relative entropy* or the *Kullback Leibler distance* between the two probability distributions is defined as[3]

$$D(P_d \parallel P_c) = \sum_{\mathbf{d}_i} P_d(\mathbf{d}_i) \log \frac{P_d(\mathbf{d}_i)}{P_c(\mathbf{c}_i)} = E_{P_d}[\log(P_d) - \log(P_c)]. \quad (3)$$

Consider now the joint probability distribution of the input and output random variables $P(\mathbf{d}, \mathbf{c})$ such that $P_{\mathbf{d}}$ and $P_{\mathbf{c}}$ are the corresponding marginal distributions. The *mutual information* $I(\mathbf{d}, \mathbf{c})$ is the relative entropy between the joint distribution and the product distribution, namely,

$$\begin{aligned} I(\mathbf{d}, \mathbf{c}) &= D(P(\mathbf{d}, \mathbf{c}) \parallel P(\mathbf{d})P(\mathbf{c})) \\ &= \sum_{\mathbf{d}^i} \sum_{\mathbf{c}^j} P(\mathbf{d}^i, \mathbf{c}^j) \log \frac{P(\mathbf{d}^i, \mathbf{c}^j)}{P(\mathbf{d}^i)P(\mathbf{c}^j)} \\ &= \sum_{\mathbf{d}^i} \sum_{\mathbf{c}^j} P(\mathbf{d}^i, \mathbf{c}^j) \log \frac{P(\mathbf{d}^i|\mathbf{c}^j)}{P(\mathbf{d}^i)} \\ &= H(\mathbf{d}) - H(\mathbf{d}|\mathbf{c}). \end{aligned} \quad (4)$$

Additional properties[4] of mutual information can be found in (Cover and Thomas, 1991).

[1] E. T. Jaynes has demonstrated the connection between information theory, statistics and statistical mechanics in two papers from 1957 (Jaynes, 1957a; Jaynes, 1957b) which are also in a forthcoming book about his work (Jaynes, 1999).

[2] In information theory, it is customary to neglect the Boltzmann constant which sets up the units correctly, and to use the logarithm of base 2 so that the information is measured in bits.

[3] Note that this is not symmetric and does not satisfy the triangle inequality.

[4] Note that by symmetry $I(\mathbf{d}, \mathbf{c}) = H(\mathbf{c}) - H(\mathbf{c}|\mathbf{d})$, and $I(\mathbf{d}, \mathbf{c}) = H(\mathbf{c}) + H(\mathbf{d}) - H(\mathbf{d}, \mathbf{c})$.

By maximizing the mutual information, we effectively minimize $H(\mathbf{d}|\mathbf{c})$ namely we reduce the uncertainty about the input $\mathbf{d}$ by knowing the output $\mathbf{c}$. Thus, given a constrained situation where the output $\mathbf{c}$ carries less data than the input $\mathbf{d}$, information theory tells us what the optimal output should be for a given input so as to have, on average, maximal knowledge about the input. Synaptic modification rules can be derived from solving the mutual information maximization problem under various assumptions about the probability distribution of input words. The solution to such learning rules is based on gradient ascent or a more sophisticated optimization algorithm, e.g., conjugate gradient.

2.1 Distributions that maximize entropy under various constraints

When we observe a certain distribution, it is natural to ask if this distribution represents a redundant coding or does it maximize entropy under certain constraints. In some cases we may be interested to recover the constraints under which the distribution maximizes the entropy. In this section we mention some of the most common constraints and the distributions which are naturally connected with these constraints. Entropy maximization, or as it is sometimes called the MAXENT principle, is a powerful statistical inference tool; Given a certain set of constraints on a random variable, it suggests the *only possible* underlying distribution for the process. If the observed distribution is different, this implies that there are additional or different constraints governing the process. An excellent review with connection to statistical mechanics can be found in (Jaynes, 1957a) or in his forthcoming book (Jaynes, 1999). Applications of this inference tool are many, see for example (Skilling, 1989).

Bounded distributions The uniform distribution maximizes the entropy of a random variable with bounded values. Note that when discritizing a r.v., its distribution becomes automatically bounded, but it is the non-discritized distribution which governs the process, thus, we would expect a maximal entropy distribution of 8 bit gray level pictures to have a Gaussian and not a uniform distribution.

Positive valued random variables Distributions that take only positive values or more generally, are bounded from below, are a special important case as they include for example to distributions of spike counts over a certain measurement window. It turns out that under mean value constraint the Poisson distribution maximizes the entropy. Under a variance constraint (of positive valued distribution) the Gibbs distribution maximizes the entropy. This distribution occurs often when a non-negative functional (also called an Energy or a Hamiltonian) can be associated with a configuration state of a physical system. A famous example is the annealing process (Brillouin, 1956) and a numerical algorithm called simulated annealing (Kirkpatrick and Jr., 1983; Geman and Geman, 1984).

Fixed variance constraint Under a fixed mean and variance and no bounds on the values the random variable, the Gaussian distribution maximizes the entropy. This makes a strong connection between minimizing entropy and searching for distributions that are far from Gaussian. It also shows that a linear layered networks receiving Gaussian distributed inputs should extract the projections which maximize the variance, namely find the principal components of the data in order to maximize the entropy of the projections.

3 The statistics of natural images and optimal retina-cortex maps

A coding scheme which generates a probability distribution that is different from the one that maximizes entropy under the appropriate coding constraints is said to be *redundant*. It is thus natural to study the signal distribution at various information junctions in the brain and determine the rate of redundancy of code at those locations. The first place to start in the visual pathway is of course the images themselves. It was demonstrated recently (Ruderman, 1994) that the single log intensity of pixel distribution of a small collection of natural images is not Gaussian. indicates that the multi-dimensional distribution of pixel images is not Gaussian, since otherwise, every (linear) projection (including single pixel projections) should have been Gaussian. Ruderman suggests a transformation of the pixel intensity, based on its local variance, which makes the new distribution Gaussian. While the optimality of such a transformation is demonstrated from information theory considerations, the biological basis of such transformation has yet to be found.

Atick and Redlich (1992) hypothesis that the main goal of retinal transformations is to eliminate redundancy in input signals, particularly that due to pairwise correlations among pixels (second order correlation). Their discussion of the optimal response of Ganglion cells is motivated by information maximization.

Field (1987) suggests an interesting match between the spectrum of natural images, and the log polar mapping from retina to cortex. Based on a small number of images, he observes that the power spectrum goes down like $1/f^2$ where f is the frequency of the changes of grey level in the image. Assuming that the coding is done similarly in each of the frequency bands, this implies that differently frequency bands do not carry the same amount of information, thus, leading to suboptimal coding of the information. He suggests that log polar retinotopic mapping, in which, the bandwidth of each frequency band is a fraction of the central frequency, causes each frequency band to carry the same amount of information, and is thus, optimal from information theory view point. More recently Field suggested that the redundancy can be utilized to produce a more constant response from highly varying spectra that is frequency dependent (Field and Brady, 1997).

Linsker (1986) had presented a set of equations stemming from a simple Hebbian rule with the addition of some weight constraints for stability, as a framework for synaptic modification in early layers of visual cortex. The neuronal goal was to maximize the mutual information between the inputs and outputs of each layer in his network. He showed that on/off cells, simple and some types of complex cells could emerge from this goal.

The above examples hypothesise that the goal of neuronal learning and data relay maybe to reduce redundancy and transfer a non-Gaussian distribution into a Gaussian one or to utilize the redundancy to gain other desired properties such as sensitivity to varying spatial frequency. So far, we have not seen a reduction in the amount of information relay, but merely a recoding that makes the code more efficient. We now turn to methods that actually attempt to *reduce* the amount of relayed-information by extracting *important* information (based on some criteria to be discussed) and ignoring the rest. These methods actually emphasize the parts of the data that are not Gaussian in a manner that is described below.

4 Minimal description length

We have seen how information theory can suggest an optimal coding **c** for a given input **d** based on the probability distribution of the inputs. In this case, only the resulting code is being transmitted through the bottleneck transmission line to the receiver which then tries to reconstruct the inputs. This formulation does not take into account the complexity of the code that is being sent, and the complexity of constructing or decoding this code. A different information theoretic formulation, which is more appropriate for supervised learning, does take the above considerations into account. This formulation is based on the *minimal description length* principle (Rissanen, 1984) which states that the way to choose a better model for the data is by minimizing concurrently the cost of describing the model and the cost of describing the misfit between the model and the data. In terms of the information bottleneck that we have described before, we can view the current situation as if there is a teacher and student network, and the teacher is trying to send to the student the network to solve a certain problem. Under a supervised setup, the assumption is that both the student and teacher can see the input data (zero cost) but only the teacher knows what the output should be. For the student to reconstruct the output, he would need to have a good model (network) namely small misfit between network output and desired output and for a quick learning, the model should be simple. Both these properties can be measured by the entropy of sending the information about data misfit and about the model. In classical information theory it is often assumed that the cost of learning a model can be neglected as learning takes place only once while data is sent continuously, however when modeling learning, it is clear that the cost of learning plays an important role and should not be neglected. It remains to be see whether measuring the model (or learning) cost using the cost of sending the model parameters (entropy of the weight distribution) would turn out to be a useful constraint and a useful neuronal learning goal.

Note that measuring model cost by the entropy of its parameters, may be radically different than measuring model cost by the actual or effective number of parameters, as has been proposed before (Akaike, 1974; Moody, 1992).

5 Projection pursuit and cortical plasticity

While we have so far discussed maximization of entropy under various conditions it turns out that sometimes minimization of the entropy is sought. This occurs when a classification is sought, when we want the output to have little ambiguity, when searching for independent components (Comon, 1994) and most notably, when looking for structure in the data by searching for interesting projections. It is due to the central limit theorem which says that given a list of independent random variables with the same distribution, their mean is normally distributed. Thus, a random projection of a high dimensional data would yield a single dimensional Gaussian distribution unless there is a strong dependency between the projection vector and the data. Similarly, in the case of independent components, the linear summation of sources makes the resulting data more Gaussian than the original data and thus, to recover the original data, one has to search for largest deviations from Gaussian distribution which is done by minimizing some approximation to the entropy. A general framework that is useful for the above examples is exploratory projection pursuit (Friedman, 1987) and its supervised version, the projection pursuit regression (Friedman and Stuetzle, 1981).

Projection pursuit (PP) methods seek features which emphasize the non-Gaussian nature of distributions (Huber, 1985, for review). They seek structure that is exhibited by (semi) linear

projections of the data. The relevance to neural network theory is clear, since the activity of a neuron is largely believed to be a semi linear function of the projection of the inputs on the vector of synaptic weights. Diaconis and Freedman (1984) have shown that for most high-dimensional clouds (of points), most low-dimensional projections are approximately Gaussian. This finding suggests that important information in the data is conveyed in those directions whose single dimensional projected distribution is far from Gaussian. Polynomial moments are good candidates for measuring deviation from Gaussian distribution, for example, skewness and kurtosis which are functions of the first four moments of the distribution, are frequently used in this respect.

Intrator and Cooper (1992) have shown that a BCM neuron can find structure in the input distribution that exhibits deviation from Gaussian distribution in the form of multi-modality in the projected distributions. Since clusters can not be found directly in the data due to its sparsity, this type of deviation, which is measured by the first three moments of the distribution, is particularly useful for finding clusters in high dimensional data and is thus useful for classification or recognition tasks.

6 Summary

We have demonstrated the important role of information theory in conveying information throughout visual cortex. We have presented cases in which information theory considerations led people to seek methods for Gaussianizing the input distribution, and in other cases led people to seek learning goals for non-Gaussian distributions. We have presented the MDL principle as a goal for learning which takes into account the complexity of the decoding network, and have presented projection pursuit methods as a framework for seeking projections that are far from Gaussian (minimize entropy).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Atick, J. J. and Redlich, N. (1992). What does the retina know about natural scenes. *Neural Computation*, 4:196–210.

Brillouin, L. (1956). *Science and Information Theory.* Academic Press, New York.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36:287–314.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory.* Wiley.

Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Annals of Statistics*, 12:793–815.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America*, 4:2379–2394.

Field, D. J. and Brady, N. (1997). Wavelets, blur and the sources of variability in the amplitude spectra of natural scenes. *Vision Research*, 37:3367–3383.

Friedman, J. H. (1987). Exploratory projection pursuit. *Journal of the American Statistical Association*, 82:249–266.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Huber, P. J. (1985). Projection pursuit. (with discussion). *The Annals of Statistics*, 13:435–475.

Intrator, N. and Cooper, L. N. (1992). Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5:3–17.

Jaynes, E. T. (1957a). Information theory and statistical mechanics I. *Phys. Rev.*, 106:620–530.

Jaynes, E. T. (1957b). Information theory and statistical mechanics II. *Phys. Rev.*, 108:171–190.

Jaynes, E. T. (1999). *Probability theory.* Preprint: ftp://bayes.wustl.edu/Jaynes.book.

Kirkpatrick, S. and Jr., C. D. G. (1983). Optimization by simulated annealing. *Science*, 220:671–680.

Linsker, R. (1986). From basic network principles to neural architecture (series). *Proceedings of the National Academy of Science*, 83:7508–7512, 8390–8394, 8779–8783.

Moody, J. E. (1992). The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In Moody, J. E., Hanson, S. J., and Lippmann, R. P., editors, *Advances in Neural Information Processing Systems*, volume 4, pages 847–854. Morgan Kaufmann, San Mateo, CA.

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30:629–636.

Ruderman, D. L. (1994). The statistics of natural images. *Network*, 5:517–548.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423 and 623–656.

Skilling, J. (1989). *Maximum Entropy and Bayesian Methods.* Kluwer Academic, Dordrecht.

Contents