Mathematical modeling of cognitive learning and memory

Vipin Srivastava[1,2] **(vpssp@uohyd.ernet.in)** & **Suchitra Sampath**[1] **(cc09pc01@uohyd.ernet.in)**
[1]Centre for Neural and Cognitive Sciences & [2]School of Physics, University of Hyderabad, India.

Abstract

We demonstrate how basic cognitive functions of learning and memory can be modeled mathematically and how such models are first built from a bare minimum of essential information and then developed systematically in a step by step manner to include more and more realistic features.

Keywords: Cognitive learning and memory; Hopfield model; orthogonalization; attractor neural network; LTP; LTD; spin glass

Introduction

Learning and memory are amongst the basic cognitive attributes of our brain, yet we are only beginning to understand the physiological mechanisms underlying them. Whatever little success we have achieved in recent years in this direction can be attributed, to some extent, to mathematical and computational modelling of these and related phenomena. We offer a glimpse of how one approaches this problem through mathematical modelling.

Developing mathematical models

Learning and memory are related in that we first learn, and, if what we learn stays in the brain and can be recalled then we say that we are able to memorise. A major break in understanding "learning" was given by Donald Hebb in 1949 (Hebb, 1949), who pointed out that the synapse connecting the neurons are plastic in nature and that their strength can change in an irreversible manner. These changes, termed as long-term potentiation (LTP) and long-term depression (LTD), are manifestations of 'learning'. Leon Cooper (Cooper, 1973) cast the Hebbian hypothesis in the following mathematical form,

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^{(\mu)}\xi_j^{(\mu)}, \tag{1}$$

where, in a fully connected network of N neurons, J_{ij} represent the synaptic strengths between neurons i and j, and $\xi_i{}^{(\mu)}$ represents the activity of the i^{th} neuron, which is taken as 1 if the neuron fires and -1 otherwise; μ is index for a pattern/vector of ± 1s. The J_{ij} thus depends on the activities of the neurons that are connected by it as was hypothesized by Hebb (Hebb, 1949). It changes cumulatively as new patterns μ are presented successively to the network.

John Hopfield (Hopfield, 1982) used mathematical framework of a physics system called 'spin glass' (Edwards & Anderson, 1975) and incorporated this prescription for learning in a simple model of firing/not firing (i.e. ± 1) neurons to account for numerous memories that the brain can accommodate at the same time. In particular, it helped us understand 'content addressability' or 'associative recall' in which if the network encounters a pattern that is similar to but not the same as an imprinted pattern, then it can associate the new pattern with the imprinted one. This accounts for a common feature of cognitive memory in which we can identify a familiar (or memorized) object from its partial, or obscured, or noisy appearance. In fact we can imagine around each imprinted pattern a collection of patterns that bear similarity with the imprinted pattern in varying degrees. This region in the configuration space is called "basin of attraction" and the Hopfield network is called attractor neural network (ANN). A noisy version of an imprinted pattern falling within a certain range around it, if presented to the ANN, will by and by converge to the imprinted pattern following the retrieval/recall prescription,

$$h_i^{(\nu)} = \frac{1}{2}\sum_{\substack{j=1 \\ j\neq i}}^{N} J_{ij}\xi_j^{(\nu)}, \tag{2}$$

where $h_i^{(\nu)}$ is the local field (or post-synaptic potential) on neuron i due to activities on all the other $(N-1)$ neurons (in an arbitrary pattern ν) projecting onto i via J_{ij}'s. If ν^{th} pattern is not one of the the imprinted patterns then the condition,

$$h_i^{(\nu)}\xi_i^{(\nu)} > 0 \tag{3}$$

will not be met for all i's. In that case $\{1\ \mathrm{sgn}(h_i^{(\nu)})\}$ are fed on the right hand side of equation (2) as $\{\xi_i^{(\nu)}\}$ and condition (3) is checked with the new set $\{h_i^{(\nu)}\}$. After a few iterations the ν^{th} pattern converges to the imprinted pattern in whose basin of attraction the ν^{th} pattern happens to fall. In physics terms this means that the imprinted pattern, say μ, corresponds to a minimum of the following total energy function (or Hamiltonian),

$$H = \frac{1}{2}\sum_{\substack{j=1 \\ (j\neq i)}}^{N} J_{ij}\xi_i^{(\mu)}\xi_j^{(\mu)}, \tag{4}$$

This energy function is akin to that of spin-glass (Edwards & Anderson, 1975). What makes it useful as a model for memory is that it has an exponentially large number of minima, which correspond to different configurations of up and down spins or ± 1, being fed in through (1).

Random sets of $\{\xi_i^{(\mu)}\}$ minimize H as long as the number of imprinted patterns does not exceed a critical limit (Amit, Gutfreund, & Sompolinsky, 1985). As new patterns are imprinted according to (1) noise builds up in the system and beyond a stage $(p/N > 0.14)$ the noise submerges the signals and we end up in a situation where none of the imprinted patterns is retrieved. This catastrophic loss of memory is cognitively unrealistic.

In figure (1) we show a simulation of the Hopfield model. It shows the variation of the number of patterns that are retrieved with 100% accuracy as a function of the number of stored patterns, p normalised by N. Note that beyond p/N=0.1, the fraction of stored patterns that are retrieved accurately begins to reduce, and drops rather steeply for p/N >0.15. Close to p/N=0.3 hardly any of the stored patterns is retrieved. This marks the memory catastrophe.

Going beyond, with corrections

To remove the above hurdle we have improved the Hopfield model to eliminate the noise from the system, which is produced by "cross-talks" between the imprinted patterns. Our hypothesis is that when an information comes to be recorded, it is first "orthogonalized" with respect to all the information in the memory, and then the orthogonalized version is stored in the memory following the Hebbian hypothesis (1). Orthogonalization is a mathematical transformation that converts a set of vectors into a mutually perpendicular set. Orthogonalization amounts to identifying similarities and differences that the new pattern may have with all those in the memory and then storing these similarities and differences in the synapses. While the mathematical details can be found in (Srivastava & Edwards, 2000), we will highlight here a curiously interesting aspect of our hypothesis.

Suppose a set of vectors $\{\xi^{(\mu)}\}$ is to be stored in the Hopfield like neural network. In the orthogonalization hypothesis $\vec{\xi}^{(\mu)}$'s will be orthogonalized sequentially (for μ=1,2,3...p) following Gram-Schmidt's procedure (Srivastava & Edwards, 2000). This will give us a set $\{\vec{\eta}^{(\mu)}\}$, which will be inscribed/stored in the network following (1) using $\{\eta^{(\mu)}\}$ instead of $\{\xi^{(\mu)}\}$. However, we find that we can study the retrieval, or recall, of the original vectors $\vec{\xi}^{(\mu)}$'s from the network. Most significantly N $\vec{\xi}^{(\mu)}$'s in a network of N are retrieved efficiently, i.e. with 100% accuracy in a single iteration of prescription (2).

The red plot in figure (1) displays that the fraction of p stored patterns that can be retrieved perfectly stays at 1 for all values of p upto p=N when the orthogonalized versions of the given p patterns are stored. The orthogonalization scheme gives new insight into the basins of attraction of $\vec{\xi}^{(\mu)}$'s and their stability conditions (Sampath & Srivastava, manuscript in preparation).

In sum we have shown that mathematical modeling plays a crucial role in understanding the mechanisms of cognitive functions. Such models not only provide quantitative results which can be substantiated by experiments, but also have almost indefinite scope for improvement and generalization to include new parameters, and relax approximations and simplifying assumptions to make the model more and more biologically realistic. In the present model, for instance, we need to (a) dilute the connectivities between neurons (a neuron is typically connected to 15% or less of other neurons), (b) take into account the fact that J_{ij} need not be equal to J_{ji}, and (c) treat a multinary neuron rather than binary to account for the fact that a neuron may fire at different rates, etc. Also, an incoming new information need not be orthogonalized with respect to all the previously stored information; it should be orthogonalized with respect to a selected set of old and stored information - i.e., the memory ought to have a hierarchical structure. Moreover, we should generalize the model to go beyond sequential learning symbolized by Gram-Schmidt orthogonalization and include, e.g. episodal memories.

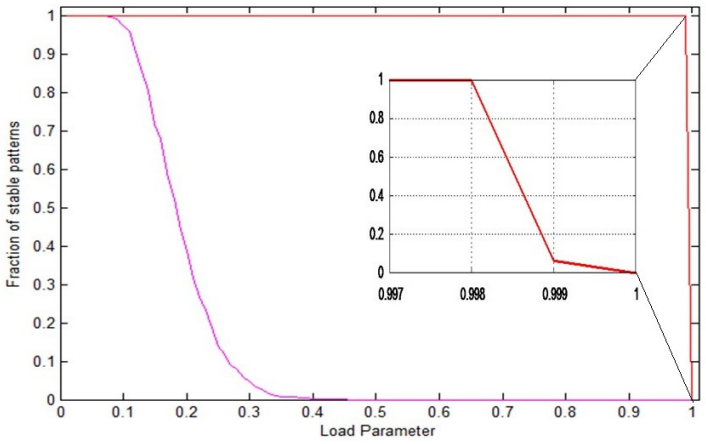

Figure 1: Fraction of perfect retrieval vs load parameter (p/N), in conventional Hopfield model (magenta) and after introducing orthogonalization for learning (red), (N=1000) .

Acknowledgements : This work is supported by the Royal Society (London, UK) and the cognitive science research initiative of the Department of Science and Technology, Government of India.

References

Amit, D., Gutfreund, H., & Sompolinsky, H. (1985). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, *55*, 1530–1533.

Cooper, L. (1973). A possible organization of animal memory and learning. In *Nobel symposium on collective properties of physical systems* (pp. 62–84). Aspensagaerden, Sweden: The Nobel Foundation.

Edwards, S., & Anderson, P. (1975). Theory of spin glasses. *Journal of Physics F: Metal Physics*, *5*, 965–974.

Hebb, D. (1949). *Organization of behaviour*. New York: Wiley.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*, 25542558.

Sampath, S., & Srivastava, V. (manuscript in preparation). New results from basins of attraction in attractor neural networks.

Srivastava, V., & Edwards, S. (2000). A model of how the brain discriminates and categorises. *Physica A: Statistical Mechanics and its Applications*, *276*, 352–358.