A Palimpsest Memory based on an Incremental Bayesian Learning Rule

A. Sandberg[a] A. Lansner[a] K. M. Petersson[b] Ö. Ekeberg[a]

[a]*Department of Numerical Analysis and Computer Science, Royal Institute of Technology, 100 44 Stockholm, Sweden*
{asa,ala,orjan}@nada.kth.se

[b]*PET Cognitive Neurophysiology, Dept of Clin Neurophysiology, Karolinska Sjukhuset, S-171 76 Stockholm, Sweden*
karlmp@neuro.ks.se

Abstract

Capacity limited memory systems need to gradually forget old information in order to avoid catastrophic forgetting where all stored information is lost. This can be achieved by allowing new information to overwrite old, as in a so-called palimpsest memory. This paper describes a new such learning rule employed in an attractor neural network. The network does not exhibit catastrophic forgetting, has a capacity dependent on the learning time constant and exhibits recency effects in retrieval.

Key words: Bayesian Confidence Propagation, Palimpsest memory

1 Introduction

Auto-associative attractor ANN:s have been proposed as models for biological associative memory [16, 8, 1, 4, 14]. Simulations have indicated that a network of cortical pyramidal and basket cells can operate as an attractor network and its connectivity structure captures many aspects of cortical functional architecture [6, 5, 3, 2].

The standard correlation based learning rules for attractor neural networks suffer from catastrophic forgetting, i.e. all memories are lost as the system gets overloaded. This runs contrary to what is believed to happen in working and intermediate memory such as the hippocampus, where new information needs to be continuously learned, stored and retrieved but not necessarily remembered beyond a certain time [12]. What is needed is learning where new patterns are stored on top of older ones which are gradually overwritten and become inaccessible, a so-called "palimpsest memory" [13].

A learning rule for attractor networks derived from Bayes rule has previously been developed, the BCPNN (Bayesian Confidence Propagation Neural Network [9, 10, 11, 7]. The learning rule is based on a probabilistic view of learning and retrieval, with input and output unit activities representing confidence of feature detection and posterior probabilities of outcomes, respectively. Weights are set based on probability estimates from the training data. Besides in ANN models, it has also been used for simulation of cortical dynamics and cell assemblies [3, 2].

This paper studies the properties of an attractor network based on an incremental version of the Bayesian learning rule, which exhibits palimpsest properties, a capacity dependent on learning rate and convergence speed dependent on memory load and recency.

2 Bayesian Learning

The traditional BCPNN learning rule estimates probabilities of individual features P_i and co-occurrences of them P_{ij} by counting the number of occurrences in the training data and dividing by the number of examples. Here we replace the fixed probability estimates with time varying estimates of rates of events (spikes) that are updated as new information arrives. We want an estimate that i) will converge towards $P_i(t)$ and $P_{ij}(t)$ in a stationary environment, ii) gives more weight to recent than remote information and iii) has a time scale that leads to smoothing and adaptation to trends in a non-stationary environment. The simplest such estimate is exponential smoothing of unit activities. Using this, the resulting network dynamics becomes

$$\eta \frac{do_i}{dt} = \Theta\left(\beta_i(t) + \sum_{j=1}^{N} w_{ij}(t) o_j(t)\right) - o_i(t) \tag{1}$$

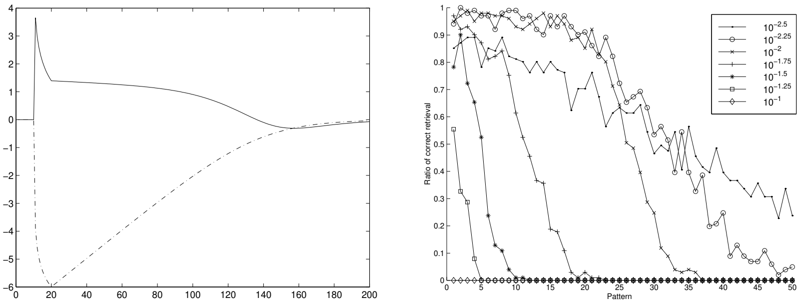
$$\tau \frac{d\Lambda_i}{dt} = [(1-\lambda_0)o_i(t) + \lambda_0] - \Lambda_i(t) \tag{2}$$

$$2\tau \frac{d\Lambda_{ij}}{dt} = [(1-\lambda_0^2)o_i(t)o_j(t) + \lambda_0^2] - \Lambda_{ij}(t) \tag{3}$$

$$\beta_i(t) = \log(\Lambda_i(t)) \tag{4}$$

$$w_{ij}(t) = \begin{cases} \log\left(\frac{\Lambda_{ij}(t)}{\Lambda_i(t)\Lambda_j(t)}\right) & i \neq j \\ 0 & i = j \end{cases} \tag{5}$$

where η is the passive time constant of a unit, λ_0 a low background activity, o_i is the output of unit i, $\Theta(x)$ is a clipped exponential function, Λ_i and Λ_{ij} rate



(a) Weight (b) Forgetting curves

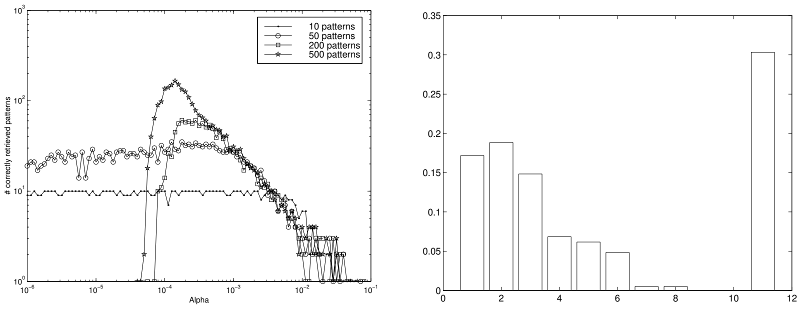
Fig. 1. (a) Weights between two units that are active together at $10 \leq t \leq 20$ (solid line) and when only one unit is active (dot-dash line), $\alpha = 0.05$. (b) Forgetting curve during continuous learning. A 100-unit network was trained with 50 sparse patterns and recall measured as the ratio of correct retrievals from noisy input patterns for different values of α.

estimates of $P_i(t)$ and $P_{ij}(t)$ respectively, and τ the learning time constant. For convenience, $\alpha = 1/\tau$ will be used in the following; during testing of the net α was set to 0.

3 Results

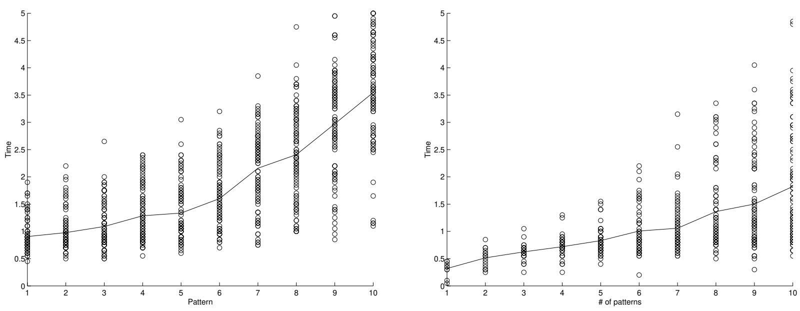
Figure 1a demonstrates the change over time of a weight between two units where one or both of them are active for a brief interval. A network with such decaying connections can exhibit palimpsest memory properties. As new data arrives, old data is forgotten and the ability to learn remains unchanged (figure 1b). Stored memories remained retrievable for a period of time set by the time constant until they rapidly decayed. For short time constants the capacity of the memory grew roughly linearly with the time constant, until it reached the limit set by the network size and sparsity, i.e. that of the standard counting BCPNN (figure 2a). For long time constants the behavior approached the counting BCPNN, which means that large training sets were averaged together and hence became impossible to retrieve, while smaller training sets were retained.

When activated with random stimuli, the probability of ending up in a given attractor state depended on its age; recent patterns were more likely to be retrieved than old ones (figure 2b). As more patterns were learned, the basins of attraction of old patterns became smaller, producing a change of convergence speed. The average time it took for the network to converge to a learned attractor state from a partial input increased with the age of the attractor and the number of attractors already learned (figure 3).



(a) Capacity (b) Random convergence

Fig. 2. (a) Number of correctly retrieved patterns as a function of α for different sizes of the training set. Pattern 1 is the most recent. (b) Distribution of states after convergence from a random initial state in a network trained with 10 patterns, pattern 11 corresponds to convergence to the low activity steady state.



(a) Recency (b) Memory load

Fig. 3. Convergence time for retrieval of stored patterns. (a) shows the effect of recency at full memory load. Pattern 1 is the most recently learned pattern. The line represent the mean of the convergence times. (b) shows the effect of increasing the memory load. The network is trained with 10 patterns.

4 Discussion

We have proposed and characterized a continuous, real-time extension to a previous Bayesian learning rule (BCPNN). It has some attractive features in comparison with previous palimpsest memories. For instance, it is derived from a statistical inference framework, it does not discard information, and it does not rely on unbiological scaling of weight changes. By changing the time constant the capacity can be regulated. Yet, at this stage there is little quantitative relation either to synaptic plasticity or to human memory phenomena. However, we expect this incremental learning rule to be useful in models of e.g. working memory and hippocampal-neocortical teach-back.

In our simulations with a fast learning and forgetting "working memory" we have found that the average convergence time increases with memory load. This relates to the classical finding by Sternberg of a reaction time dependence on list length [15]. The proposed learning rule is sensitive to temporal ordering of stimuli. However, it is likely that to reproduce memory effects in e.g. list learning, additional asymmetric components of the learning rule are required.

References

[1] Daniel J. Amit. *Modeling Brain Function: The World of Attractor Neural Networks.* Cambridge University Press, 1989.

[2] Erik Fransén and Anders Lansner. Low spiking rates in a population of mutually exciting pyramidal cells. *Network*, 6(2):271–288, 1995.

[3] Erik Fransén and Anders Lansner. A model of cortical associative memory based on a horizontal network of conneted columns. *Network*, 9:235–264, 1998.

[4] W. J. Freeman. The physiology of perception. *Scientific American*, 264:78–85, 1991.

[5] L. B. Haberly and J. M. Bower. Olfactory cortex: model circuit for study of associative memory? *Trends Neurosci*, 12(7):258–64, July 1989.

[6] M. E. Hasselmo., B. P. Anderson, and J. M. Bower. Cholinergic modulation of cortical associative memory function. *J. Neurophysiol.*, 67:1230–1246, 1992.

[7] Anders Holst. *The Use of a Bayesian Neural Network Model for Classification Tasks.* PhD thesis, Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden, September 1997. TRITA-NA-P9708.

[8] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci U S A*, 79(8):2554–8, April 1982.

[9] Anders Lansner and Örjan Ekeberg. An associative network solving the "4-Bit ADDER problem". In Maureen Caudill and Charles Butler, editors, *IEEE First International Conference on Neural Networks*, pages II–549, San Diego, CA, June 21–24 1987.

[10] Anders Lansner and Örjan Ekeberg. A one-layer feedback artificial neural network with a bayesian learning rule. *Int. J. Neural Systems*, 1(1):77–87, 1989.

[11] Anders Lansner and Anders Holst. A higher order Bayesian neural network with spiking units. *Int. J. Neural Systems*, 7(2):115–128, May 1996.

[12] James L. McClelland and Nigel H. Goddard. Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus*, 6:654–665, 1997.

[13] J.P Nadal, G. Toulouse, J.P. Changeux, and S. Dehaene. Networks of

formal neurons and memory palimpsests. *Europhysics Letters*, 1(10):535–542, 1986.

[14] P. Quinlan. *Connectionism and Psychology. A Psychological Perspective on Connectionist Research.* Harvester, Wheatsheaf, New York, 1991.

[15] S. Sternberg. High-speed scanning in human memory. *Science*, 153(736):652–4, Aug 5 1966.

[16] D.J. Willshaw, O.P. Buneman, and H.C. Longuet-Higgins. Non-holographic associative memory. *Nature*, 222:960, 1969.

Anders Sandberg
Anders Sandberg has a M.Sc. in Computer Science from Stockholm University. He is pursuing a PhD at the SANS group at the Department of Numerical Analysis and Computing Science of the Royal Institute of Technology in Stockholm. His main research interest involves the interplay between the medial temporal lobe and neocortex in memory consolidation, and how this relates to modulation of neural plasticity.

Anders Lansner
Anders Lansner received his M.Sc. in Engineering Chemistry and Biochemistry in 1974 and his Ph.D. in Computer Science from the Royal Institute of Technology in 1986. Since 1992 he has been associate professor in Computer Science and project manager of the SANS group (Studies of Artificial Neural Systems), which he founded in 1987. His research interests range from neural computation and control to detailed computational models of specific biological systems and functions.

Karl-Magnus Petersson
Karl Magnus Petersson received his B.Sc. in Mathematics and Physics in 1988 at Stockholm University and his M.D. at the Karolinska Institute in 1996. He is currently pursuing a Ph.D. at the Cognitive Neurophysiology group at the Department of Clinical Neuroscience of the Karolinska Institute. His research interests range from artificial neural computation to the functional organization of the human brain. Current work focuses on memory and learning in the human brain as well as language function.

Örjan Ekeberg
Örjan Ekeberg is an associate professor in Computer Science at the Royal Institute of Technology in Stockholm since 1994. His research interests ranges from artificial to biological neural networks with an emphasis on integrating the two lines of research. Current work focuses on the development of techniques for integrated neuro-mechanical simulations.