Statistics, Probability and Game Theory
IMS Lecture Notes – Monograph Series (1996) Volume 30

BAYESIAN UPDATINGS IN HOPFIELD-LIKE ASSOCIATE MEMORY MODELS

ISAAC MEILIJSON
Tel Aviv University

Abstract

This article summarizes and explains in statistical terminology two papers written jointly with Eytan Ruppin, presenting a Bayesian outlook on the performance of Hopfield-like attractor neural networks. Restricting attention to the evaluation of performance after two iterations rather than studying thermodynamical limits, we are able to extend the analysis to more general models than those usually considered: input patterns applied to small subsets of neurons, general connectivity architectures of the synaptic network and more efficient use of history. We show that the optimal signal that a Bayesian neuron should emit has a *slanted sigmoidal* shape as a function of its current field value (or posterior odds), and provide an intuitive account of activation functions with such non-monotone shapes.

1. Dedication. I studied in Berkeley at the Department of Statistics from 1967 to 1969. During this period I was a teaching and research assistant, mostly working under David Blackwell and Lester Dubins, to whom I equally owe the light that illuminates most of what I study. I had the privilege of being Blackwell's Ph.D. student, writing my thesis "A bargaining problem" under his guidance. I was also fortunate to be one of his teaching assistants in the course Stat 2, during the gestation period and writing of his marvelous textbook "Basic Statistics", the lucid, clear, elementary introduction to Bayesian thought. From Blackwell I learned stochastic modeling and Dynamic Programming, and got his Bayesian attitude into my bloodstream.

The following paper is warmly and gratefully dedicated to David Blackwell. It is a summary of an attempt with Eytan Ruppin to replace some ad-hoc dynamics transferred from the Statistical Mechanics language of spin models of Neural Networks, to a more Blackwellian approach: neurons have prior and posterior beliefs about their dichotomous hidden states, and these beliefs dictate their signals and decisions.

2. Introduction. In an *associative memory* model, there is a *storage phase* in which patterns are stored, and a *retrieval stage* in which distorted versions of the patterns are presented to the network, that is expected to recognize the correct input pattern by a sequence of "associations".

More formally, in the storage phase, a *pattern* is a vector ξ of length N (for *neurons*) with ± 1 entries. During the training period, the network is presented M (for *memories*) such patterns $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(M)}$, of which only the *Hebbian weights*

$$W_{ij} = \sum_{\mu=1}^{M} \xi_i^{(\mu)} \xi_j^{(\mu)}, \ 1 \leq i \neq j \leq N, \ W_{ii} \equiv 0 \tag{1}$$

are recorded (Hebb [3]). In other words, the *synapse* between neurons i and j records the excess W_{ij} of the patterns in which both neurons show the same bit ($\xi_i \xi_j = 1$) over those in which they show opposite sign ($\xi_i \xi_j = -1$), by simply adding or subtracting 1 from the current weight, upon the presentation of each new pattern. We will assume throughout the paper

Assumption 1.

The $M * N$ pattern bits $\xi_i^{(\mu)}$; $1 \leq i \leq N$, $1 \leq \mu \leq M$, are independent and identically distributed, with $P(\xi = +1) = P(\xi = -1) = \frac{1}{2}$.

Assumption 1 can and should be weakened. A more realistic model (see Tsodyks and Feigel'man [9]) has pattern bits 1 ("fire") and 0 ("don't fire") rather than ± 1, and uses sparse patterns - with a small fraction of 1's.

In the retrieval phase, the network receives a distorted version X of one of the M patterns, say ξ. To keep the model simple, we assume

Assumption 2.

(i) The pattern ξ is chosen at random, with equal probabilities $1/M$, from the collection of stored patterns.

(ii) For each $1 \leq i \leq N$, the distorted bit X_i is equal to ξ_i or $-\xi_i$ with respective probabilities $(1+\epsilon)/2$ and $(1-\epsilon)/2$, independently of everything else.

Assumption 2(ii) will be weakened in Section 3, where only some of the neurons (the *active* ones) receive the initial distorted pattern.

The problem is to recognize ξ from X, or at least to create a pattern much closer than X to ξ.

One of the common methods (Hopfield [4, 5]) is the *Hopfield* model of associative memory: each neuron "signals" its *state*, i.e., neuron j signals its coordinate X_j in X, the result of which is that neuron i receives a *field value* $f_i = \sum_{j=1}^{N} W_{ij} X_j$. Neuron i now changes X_i into sign(f_i), the first *association.* The process is repeated to produce successive associations, hopefully converging to a fixed point. To see a rationale for the method, consider the initial association, assuming all patterns to be generated by independent fair coin tossings. Then

$$f_i = \sum_{j=1}^{N} W_{ij} X_j = \sum_{\xi^{(\mu)} \neq \xi} \sum_{j=1}^{N} \xi_i^{(\mu)} \xi_j^{(\mu)} X_j + \xi_i \sum_{j=1}^{N} \xi_j X_j \tag{2}$$

$$\approx \sqrt{MN}Z + N\epsilon\xi_i = N\epsilon[\xi_i + \frac{1}{\epsilon}\sqrt{\frac{M}{N}}Z] ,$$

where Z has an asymptotic standard normal distribution. As an example, let $\epsilon = .5$ and $M = N/16$. Then the term $\frac{1}{\epsilon}\sqrt{\frac{M}{N}}$ equals .5, and while the initial pattern X had $(1-\epsilon)/2 = 25\%$ errors, the new association $\text{sign}(f)$ has $P(Z > 2) \approx 2.5\%$ errors. It is not inconceivable that further associations will improve similarity. Convergence and capacity have been extensively studied in more general models than introduced above (Komlos & Paturi [6]). Roughly speaking, M/N below $1/7$ affords reasonable recognition, at least in the weak sense that if ϵ is close enough to 1, similarity does not deteriorate. Komlos & Paturi develop in principle, for the purpose of presenting formal convergence proofs, lower bounds on the radius of the domain of very likely convergence to fixed points within a given tolerance from the pattern to be recalled.

In an attempt to discover dynamics with improved capacity or performance, we study implications of considering each neuron as a Bayesian decision maker that starts with a prior probability .5 of having a +1 in the pattern to be recalled (henceforth, "being" +1), updates it to $\lambda_i^{(0)} = (1+\epsilon X_i)/2$ upon receipt of the initial distorted pattern and proceeds then to update further posterior probabilities as a function of field values. These probabilities form the "personal history" of the individual neuron, as a function of which the neuron may determine its signal to the network in every iteration as well as its final decision on its own identity: Bayesian neuron i tests the hypothesis $H_0 : \xi_i = +1$ versus $H_1 : \xi_i = -1$ based on its personal history as observation. Technical details appear in Meilijson & Ruppin [7, 8].

3. The model. Each of the N neurons has synaptic connections from K of the other neurons. Let $I_{ij} = 1$ or 0 depending on whether there is a connection from neuron j to neuron i. The initial distorted pattern with $X_i = \pm 1$ is received by L neurons, while all others receive input $X_i = 0$. We assume

Assumption 3.

(i) The L active neurons are chosen by simple random sampling from among the set of N neurons, independently of everything else.

(ii) For each $1 \leq i \leq N$, the K synaptic connections into neuron i are chosen by simple random sampling from the set of $N - 1$ other neurons, independently of everything else.

With regard to Assumption 3(i), quiescent neurons are allowed in [7, 8] to receive the distorted pattern with a distortion rate δ replacing the distortion rate ϵ of the active neurons. The statement above that quiescent neurons

receive $X_i = 0$ is tantamount to assuming that $\delta = 0$. Assumption 3(ii) will be weakened in Section 6 by allowing more general connectivity architectures.

Letting $g(t) = \text{arctanh}(t) = \frac{1}{2}\log\frac{1+t}{1-t}$, we may define a notion of *generalized field* $g_i^{(0)} = g(\epsilon)X_i$ of neuron i, that conveniently represents its prior probability of being $+1$ as

$$\lambda_i^{(0)} = P(\xi_i = +1|X_i) = \frac{1}{1+\exp(-2g_i^{(0)})} \tag{3}$$

and its *prior belief* $O_i^{(0)} = \epsilon X_i$, or degree of preference for $+1$ over -1, as $O_i^{(0)} = \lambda_i^{(0)} - (1-\lambda_i^{(0)}) = 2\lambda_i^{(0)} - 1 = \tanh(g_i^{(0)})$. In this context, ϵ may be termed *belief coefficient*. As we shall see in (6), (7), (15) and (16), the posterior probabilities $\lambda_i^{(1)}$ and $\lambda_i^{(2)}$ that neuron i assigns to being $+1$ are also expressible as (3) for additively updated generalized field values $g_i^{(1)}$ and $g_i^{(2)}$. Using statistical terminology, generalized field values are (up to multiplication by 2) prior and posterior log odds in Bayesian hypothesis testing with normally distributed observations.

When each neuron signals its initial state X, the mean number of non-zero signals received by each neuron is $n_1 = LK/N$, and the *input field* $f_i^{(1)}$ of neuron i is (compare with (2))

$$f_i^{(1)} = \sum_{j=1}^{N} W_{ij}I_{ij}X_j \ , \tag{4}$$

with

$$\begin{aligned} E(f_i^{(1)}|\xi_i, X_i) &= n_1\epsilon\xi_i \\ \text{VAR}(f_i^{(1)}|\xi_i, X_i) &= n_1 M. \end{aligned} \tag{5}$$

Upon observing its input field, neuron i updates its generalized field from $g_i^{(0)}$ to

$$g_i^{(1)} = g_i^{(0)} + \epsilon f_i^{(1)}/M \tag{6}$$

and represents its posterior probability of being $+1$ as (see (3))

$$\lambda_i^{(1)} = P(\xi_i = +1|X_i, f_i^{(1)}) = \frac{1}{1+\exp(-2g_i^{(1)})} \ , \tag{7}$$

with the corresponding posterior belief $O_i^{(1)} = \tanh(g_i^{(1)})$.

If each neuron were to choose a sign on the basis of this single association, the optimal decision of neuron i would be $\text{sign}(g_i^{(1)})$. The *similarity*, or overall probability of correct decision, can be readily seen to be

$$S_1 = \frac{n_1}{K}Q^*(\frac{\epsilon}{\sqrt{\alpha}}, \epsilon) + (1 - \frac{n_1}{K})Q^*(\frac{\epsilon}{\sqrt{\alpha}}, 0) \ , \tag{8}$$

where

$$\alpha = M/n_1 \tag{9}$$

denotes the *initial memory load* and

$$Q^*(x,t) = \frac{1+t}{2}\Phi(x + \frac{g(t)}{x}) + \frac{1-t}{2}\Phi(x - \frac{g(t)}{x}) \ , \tag{10}$$

for $x > 0$ and $0 \leq t < 1$, expressed in terms of the standard normal cumulative distribution function Φ. $Q^*(x,t)$ is the probability of correct decision for a neuron that starts with a prior probability $(1+t)/2$ of correct decision and observes an input field with signal-to-noise ratio x. Expression (8) was derived by Englisch, Engel, Schutte & Stcherbina [2] for the case $n_1 = K = L = N$ by searching for an optimal re-definition of W_{ii}.

Let us now allow neuron i to signal the network some general *activation function* $h(g_i^{(1)}, X_i)$ depending on its personal history. The new input field received by neuron i is (compare with (4))

$$f_i^{(2)} = \sum_{j=1}^{N} W_{ij} I_{ij} h(g_j^{(1)}, X_j) \ . \tag{11}$$

In order to analyze the effect of this second association, it is necessary to identify the asymptotic joint distribution of the input fields $(f_i^{(1)}, f_i^{(2)})$ given (ξ^i, X_i). We have not yet proved an adequate version of the Central Limit Theorem for this problem, but proceed with the computations under a joint normal working paradigm. Under this model assumption, the conditional variance $\mathrm{VAR}(f_i^{(2)}|\xi^i, X_i, f_i^{(1)})$ should be constant in $f_i^{(1)}$ and the corresponding conditional mean should be linear in $f_i^{(1)}$, satisfying the usual Linear Regression formulas, in terms of three parameters to be identified, $E(f_i^{(2)}|\xi_i, X_i)$, $\mathrm{VAR}(f_i^{(2)}|\xi_i, X_i)$ and $\mathrm{COV}(f_i^{(1)}, f_i^{(2)}|\xi_i, X_i)$. This is done rigorously in [7] and [8]. As it turns out, the first parameter is linear and the other two are constant. Thus, the two-association dynamics and performance are fully described by four constants ϵ^*, b, a and τ^2, that model respectively the conditional mean (see (5)) via the *modified belief coefficient* ϵ^* and the *feedback parameter* b as

$$E(\frac{f_i^{(2)}}{N}|\xi_i, X_i) = \epsilon^* \xi_i + b X_i \ , \tag{12}$$

the regression coefficient

$$a = \frac{\mathrm{COV}(f_i^{(1)}, f_i^{(2)}|\xi_i, X_i)}{\mathrm{VAR}(f_i^{(1)}|\xi_i, X_i)} \tag{13}$$

and the residual variance

$$\tau^2 = \mathrm{VAR}(\frac{f_i^{(2)}}{N}|\xi_i, X_i, f_i^{(1)}) \ . \tag{14}$$

Neuron i now updates its probability of being $+1$ to (see (7))

$$\lambda_i^{(2)} = P(\xi_i = +1|X_i, f_i^{(1)}, f_i^{(2)}) = \frac{1}{1+\exp(-2g_i^{(2)})} \ , \tag{15}$$

where the new generalized field $g_i^{(2)}$ is given by (see (6))

$$g_i^{(2)} = g_i^{(1)} + \frac{\epsilon^* - a\epsilon}{\tau^2}(\frac{f_i^{(2)}}{N} - a\frac{f_i^{(1)}}{N} - bX_i) \ . \tag{16}$$

If each neuron were to choose a sign on the basis of two associations, the optimal decision of neuron i would be $\mathrm{sign}(g_i^{(2)})$. The overall probability of correct decision turns out to be expressible in terms of the parameter (compare with (9))

$$\alpha^* = \frac{M}{n_1 + M((\epsilon^*/\epsilon - a)/\tau)^2} = \frac{M}{n_1^*} \tag{17}$$

as (compare with (8))

$$S_2 = \frac{n_1}{K}Q^*(\frac{\epsilon}{\sqrt{\alpha^*}}, \epsilon) + (1 - \frac{n_1}{K})Q^*(\frac{\epsilon}{\sqrt{\alpha^*}}, 0) \ , \tag{18}$$

i.e., as if there was a single association with n_1^* rather than n_1 non-zero signals received by each neuron. S_2 is a decreasing function of α^*. Hence, the two-association performance is measured by the value of $|\epsilon^*/\epsilon - a|/\tau$.

4. Convergence in the Bayesian set-up. The analysis beyond two associations becomes very complicated. However, from a conceptual point of view, Bayesian dynamics provide a clear notion of convergence, as we shall now see. In contrast, convergence of the Hopfield associative iterations with a non-symmetric matrix $W_{ij}I_{ij}$ is not to be taken for granted.

Whatever the activation functions be, the posterior probability process

$$\lambda_i^{(t)} = P(\xi_i = +1|X_i, f_i^{(1)}, f_i^{(2)}, \cdots, f_i^{(t)}) \tag{19}$$

is a Martingale. Since $\psi(x) = \max(x, 1-x)$ is a bounded convex function on $[0,1]$, the stochastic process $\psi(\lambda_i^{(t)})$ is a bounded Sub-martingale. Hence, for Bayesian dynamics, the overall proportion $S_t = E(\psi(\lambda_i^{(t)}))$ of neurons with a correct decision on their identity, increases with every association. Furthermore, by Doob's Martingale Convergence Theorem, every neuron has

a limiting probabilistic opinion $\lambda_i^{(\infty)} = P(\xi_i = +1|X_i, f_i^{(1)}, f_i^{(2)}, \cdots)$ about its being $+1$.

Unlike the usual Bayesian updatings in which the next posterior probability in (19) is obtained as a function of the new data and the current prior, we see in (16) that the entire history of the neuron enters into play. This is so because neurons sample conditionally correlated rather than conditionally i.i.d. field values.

5. Optimal signaling under two associations. If the posterior probabilities $\lambda_i^{(1)}$ (see (7)) were independent, the activation function $h(g_i^{(1)}, X_i)$ could have been taken to be constant in X_i. In fact, the optimal function would have been the posterior belief or *sigmoid* $h(g_i^{(1)}, X_i) = \tanh(g_i^{(1)})$: neurons very certain of their sign should transmit this sign and neurons very uncertain as to their sign should be quiescent.

However, these beliefs are based on signals shared by all neurons, the signals emitted by the neurons themselves, and the dependence this information sharing induces on the input fields $f_i^{(1)}$, may in principle and does in practice influence signals and personal neuron's decisions. We optimized the activation function by maximizing $|\epsilon^*/\epsilon - a|/\tau$ (see the end of the previous section) in the Neyman-Pearson tradition and found that the best choice is not even monotone in the neuron's belief! The optimal activation function (as a function of the normalized input field $f_i^{(1)}/n_1$, for $X_i = 1$ or $X_i = 0$) is a *slanted sigmoid* (see Figure 1), the sum of the (bounded) sigmoid and an (unbounded) linear function with negative slope. Thus, neurons with very strong beliefs should play "Devil's advocate" and signal a sign opposite to the one they so strongly believe in. Non-monotone activation functions have been considered by Yoshizawa, Morita & Amari [10] and by De Felice, Marangi, Nardulli, Pasquariello & Tedesco [1].

A possible rationale for the use of the slanted sigmoid as activation function is provided by a technical Lemma (see [7]) claiming that the conditional covariance between the input fields $f_i^{(1)}$ and $f_j^{(1)}$ received by the two neurons i and j, given their states $\xi_i^{(\mu)}$ and $\xi_j^{(\mu)}$ in every pattern $1 \leq \mu \leq M$, is proportional to their synaptic weight W_{ij}. Suppose that neuron j receives an input field $f_j^{(1)}$ with $f_j^{(1)}/(n_1\epsilon) = 5$ and assume that the memory loads are such that $f_j^{(1)}/(n_1\epsilon)$ should have unit variance (see (5)). For neuron j, ξ_j is either $+1$ - in which case its Gaussian field is four standard deviations above the mean - or is -1, with a noise of six standard deviations. The latter is so unlikely relative to the former that neuron j is convinced of being $+1$, and would like to transmit that to the network. However, knowing that its field has a heavily exaggerated noise, neuron j infers that the other neurons, with input fields correlated with its own, have exaggerated field

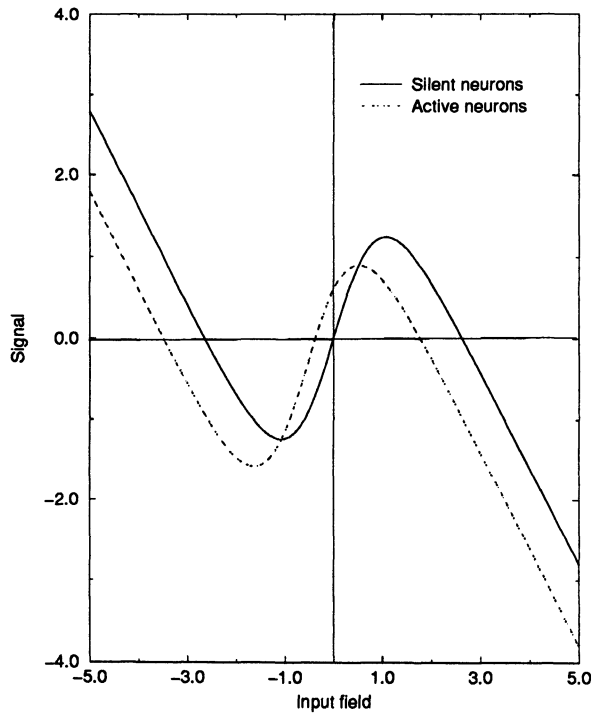

Figure 1: A typical plot of the slanted sigmoid as a function of $f^{(1)}/n_1$. Network parameters are $N = 5000$, $K = 3000$, $n_1 = 200$ and $M = 50$.

values as well. Neuron j should apply Regression Analysis and warn neuron i by telling the latter to subtract from its generalized field some multiple of the product of the synaptic weight W_{ij} and neuron j's input field bias. By the above quoted Lemma, neuron j can accomplish this task by simply signaling to the network some negative multiple of its field. These two conflicting roles, the excitatory communication of the neuron's belief on its identity and the inhibitory correction of the other neurons' biases, can be accomplished simultaneously by superimposing the decreasing linear correction on the increasing sigmoidal communication, giving rise to the slanted sigmoid.

For the sake of balance, it has been argued that this may be an artifact of the feeling of urgency we have imposed on the neurons by forcing them to guess their identity following one more association. It should be added, though, that this pressure justified itself. These excitatory-inhibitory Bayesian neurons achieve after two associations a similarity S_2 to the input pattern (see (18)) significantly closer than that achieved in the limit by Hopfield dynamics.

6. Performance under two associations. Performance under a single association as measured by the similarity S_1 (see (8)) depends on the

connectivity graph only via the total number K of synaptic connections into each neuron. Performance under two associations depends on the graph itself, via the probability r_2 that there is a synapse from neuron j to neuron i given that there is a synapse from i to j, the probability r_3 that there is a synapse from i to k given that there are synapses from i to j and from j to k, and the probability r_4 that there is a synapse from k to l given that there are synapses from i to each of j and k and from j to l. We should thus relax Assumption 3(ii) to allow for connectivity architectures other than a purely random graph.

For *fully connected* networks all three r_i's are equal to 1. For *multi-layered* networks, $r_2 = r_4 = 1$ and $r_3 = 0$. For d-dimensional *Gaussian connectivity*, where the probability of a synapse between neurons x apart is $p \exp\{-x^2/(2s^2)\}$, these coefficients are $r_k = p/(k^{d/2})$.

As it turns out, all three parameters r_2, r_3 and r_4 affect dynamics and decisions, but only the last two affect performance: Bayesian neurons are fully capable of filtering out from what they learn from other neurons in the second association the information they imparted to the network in the first association. Skipping technical details that can be found in [7, 8], the modified belief coefficient ϵ^* (see (12)) is independent of the connectivity architecture, the regression coefficient a (see (13)) is a linear function of r_3, the feedback parameter b (see (12)) is a multiple of r_2 and the residual variance τ^2 (see (14)) is a linear function of $r_4 - r_3^2$.

We see that the restriction of the analysis to two associations permitted us to expand its scope to a rather general model in terms of connectivity architecture and initial activity. While we have not yet pursued an organized study of the effects of activity and connectivity on optimal performance, the second example in Section 8 is a promising beginning. This example shows performance as a function of the network size N for a fixed number of synaptic connections per neuron (K) and initial memory load α_1. The final similarity S_2 is *not* monotone in N: a fully connected network ($N = K$) has high performance, but this performance deteriorates very rapidly as N increases until it doubles or triples K and then improves towards sparse connectivity. Given that nervous tissue regenerates poorly if at all, synaptic deletion is unavoidable. Could the U-shape of the performance function explain the evolutionary development of our big ... empty brains?

7. Qualitative comparison with Hopfield dynamics. It is commonly stated for (history-independent) Hopfield dynamics that *fresh* associations are always more effective than follow-up associations that start with the same similarity. This is not the case for Bayesian dynamics. We built examples in which the first association brings the network to a state that makes the second association much more valuable. This is especially the case for sparsely connected networks in which the initial distorted pattern is

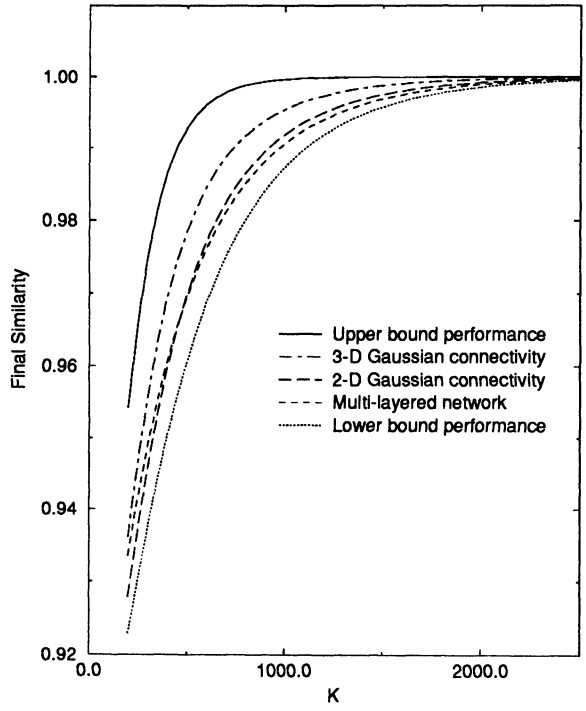

Figure 2: Two-iteration performance achieved with various network architectures, as a function of the network connectivity K. Network parameters are $N = 5000$, $L = 1000$, $M = 50$ and $\epsilon = 0.5$.

received and signaled by a small fraction of the neurons (see the first example in the next section).

For history-independent Hopfield dynamics, a single association with a sufficiently small number of active neurons outperforms two associations with the same total level of activity. Hence, signaling should be essentially *synchronous*, unlike the firing of real neurons. In contrast, under Bayesian history-dependent dynamics, two associations always outperform a single association with the same total level of activity. This conforms more adequately with the asynchronous nature of neuronal firing.

8. Illustration. Examples 1 and 2 illustrate performance, measured by the final similarity S_2 as presented in (18). Simulation results agree with the theoretical assessments.

Example 1. The similarity S_2 between pattern and decision is illustrated in Figure 2 for various connectivity architectures, as a function of the number K of synaptic connections into each neuron, for fixed network size N. In this example the input pattern is received and signaled by 20% of the network and the initial similarity of the neurons receiving the distorted input pattern is 75% (or, $\epsilon = 0.5$).

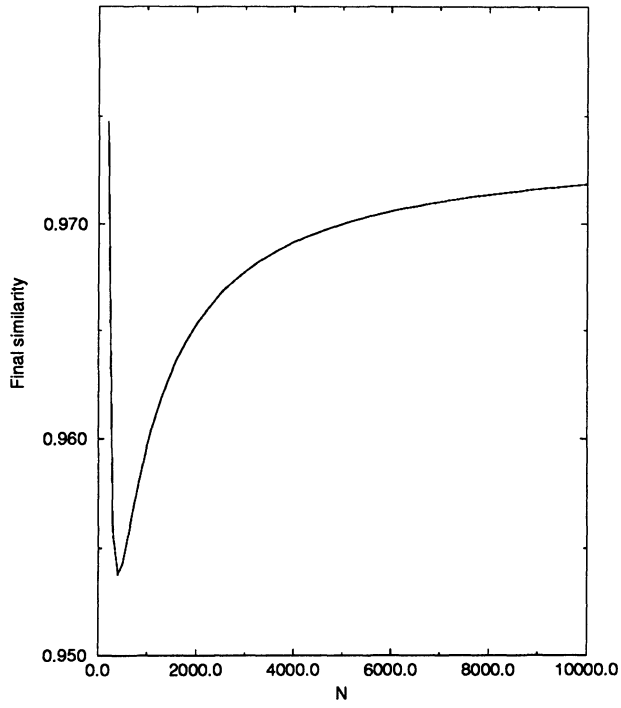

Figure 3: Two-iteration performance in a full-activity network with random connectivity, as a function of network size N. Network parameters are $n_1 = K = 200$, $M = 40$ and $\epsilon = 0.5$.

Example 2. The similarity S_2 between pattern and decision is illustrated in Figure 3 as a function of the network size N, for a fixed number K of synaptic connections per neuron, satisfying in this example $K = n_1 = 5M$. As in Example 1, $\epsilon = 0.5$. The non-monotonicity of this function was discussed in Section 6.

9. Acknowledgments. I wish to thank Eytan Ruppin for past, present and future cooperation and Boris Tsirelson for fruitful discussions.

REFERENCES

[1] De Felice P., Marangi C., Nardulli G., Pasquariello G. and Tedesco L. (1993). Dynamics of neural networks with non-monotone activation function. *Network* 4 1-9.

[2] Englisch H., Engel A., Schutte A. and Stcherbina M. (1990). Improved retrieval in nets of formal neurons with thresholds and non-linear synapses. *Studia Biophys* **137** 37-54.

[3] Hebb D. O. (1949). *The organization of behavior: A Neuropsychological theory.* New York: Wiley.

[4] HOPFIELD J. J. (1982). Neural networks and physical systems with emergent collective abilities. *Proc. Nat. Acad. Sci. USA* **79** 2554-2558.

[5] HOPFIELD J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Nat. Acad. Sci. USA* **81**, 3088.

[6] KOMLOS J. AND PATURI R. (1988). Convergence results in an associative memory model. *Neural Networks*, **1**, 239.

[7] MEILIJSON I. AND RUPPIN E. (1993). History-dependent attractor neural networks. *Network*, **4**, 195-221.

[8] MEILIJSON I. AND RUPPIN E. (1994). Optimal signalling in attractor neural networks. *Network*, **5**, 277-298.

[9] TSODYKS M.V. AND FEIGEL'MAN M.V. (1988). The enhanced storage capacity in neural networks with low activity level. *Europhys. Lett.*, **6**, 101-105.

[10] YOSHIZAWA S., MORITA M. AND AMARI S. (1993) Capacity of associative memory using a non-monotonic neuron model. *Neural Networks*, **6**, 167-176.

SCHOOL OF MATHEMATICAL SCIENCES
RAYMOND AND BEVERLY SACKLER FACULTY OF EXACT SCIENCES
TEL AVIV UNIVERSITY, TEL AVIV 69978, ISRAEL
isaco@math.tau.ac.il