# Analytic Study of the Memory Storage Capacity of a Neural Network*

W. A. LITTLE†
*Physics Department, Stanford University, Stanford, California 94305*

AND

GORDON L. SHAW
*Physics Department, University of California, Irvine, California 92717*

*Received 7 October 1977*

ABSTRACT

Previously, we developed a model of short and long term memory which was based on an analogy to the Ising spin system in a neural network. We assumed that the modification of the synaptic strengths was dependent upon the correlation of pre- and post-synaptic neuronal firing. This assumption we denote as the Hebb hypothesis. In this paper, we solve *exactly* a *linearized* version of the model and explicitly show that the capacity of the memory is related to the number of synapses rather than the much smaller number of neurons. In addition, we show that in order to utilize this large capacity, the network must store the major part of the information in the capability to generate patterns which evolve with time. We are also led to a modified Hebb hypothesis.

## I. INTRODUCTION

In an earlier series of papers [1–3] we developed a model of the brain which is based on an analogy to an Ising spin system [4]. This model describes the nature of short and long term memory in a neural network. We assumed that memory results from a form of synaptic modificaton which is dependent on the correlation of pre- and post-synaptic neuronal firing. This assumption is similar to that discussed by Hebb [5] and used by several other authors. We have shown that such a network can behave in a reliable way in spite of the presence of noise. This noise results from fluctuations in the number of neurochemical transmitter molecules released at the synapses [7] and the resultant fluctuations in the post-synaptic potentials. Reliable, well-defined behavior of the network can still occur,

*Supported in part by the Research Corporation.
†Address for all proofs and correspondence.



 0025-5564/78/0039-0281$0.175

however, because of the large number and complexity of the synaptic connections.

In any such model of memory one needs to examine the factors which determine the capacity of the memory and the independence or orthogonality of different memory "traces". The purpose of this paper is to address ourselves to these problems. First, we briefly review the essential features of the model and then consider the question of storage capacity. Although we have not been able to solve the full mathematical problem posed in our model, we are able to solve *exactly* a linearized version of the problem, and this clearly shows that the capacity of the memory is related to the number of synapses rather than to the much smaller number of neurons. In addition, it shows that in order to utilize this large capacity, the network must store the major part of the information in the capability to generate patterns which evolve with time. We are also led to a modified Hebb hypothesis. This might explain why previous experiments have failed to verify the original Hebb hypothesis. Finally, we make some comments on the close relation of our solution of the linear model to the solution of the full problem.

## II. ANALYTICAL MODEL

Consider a highly interconnected network of $N$ neurons. We describe [1–3] the system in terms of the pattern of neurons which have fired or not fired at a given time and look at the evolution of such patterns in discrete time steps $\tau$ of the order of a few msec, which is the order of the refractory period and also the decay time of a below threshold post-synaptic potential. At each fixed time $l\tau$, then, there are $2^N$ possible firing patterns of the $N$ neurons, denoted by $\alpha$:

$$\alpha = (S_{1\alpha}, S_{2\alpha}, \ldots, S_{N\alpha}), \tag{1}$$

where $S_{i\alpha} = 1$ if the $i$th neuron fires, $-1$ if the $i$th neuron does not fire for the pattern $\alpha$. We need to calculate the probability $P_{\beta\alpha}$ that a given firing pattern $\alpha$ at time $l\tau$ goes to $\beta = (S_{1\beta}, \ldots, S_{N\beta})$ at $t = (l+1)\tau$. Assuming that the $i$th neuron has a synaptic junction, either excitatory or inhibitory, at an ending of the axon from the $j$th neuron, then when the $j$th neuron fires there is an average change in polarization $V_{ij}$ induced at the hillock of neuron $i$. If the net depolarization in a time step from the sum of these post-synaptic potentials exceeds the threshold $V_i^T$, then neuron $i$ will fire. We included [2] two sources of statistical fluctuation [7] of $V_{ij}$: (a) the number of quanta of chemical transmitter released when an action potential reaches the synapse is given by a Poisson process, and (b) the quanta have a Gaussian distribution of transmitter molecules. We showed then that the probability $P_{\beta\alpha}$ as

defined above is given by

$$P_{\beta\alpha} = e^{\varepsilon M_{\beta\alpha}} / \sum_{\gamma} e^{\varepsilon M_{\gamma\alpha}}, \tag{2}$$

where

$$M_{\beta\alpha} = \sum_{i=1}^{N} S_{i\beta} \left\{ \sum_{j=1}^{N} \left[ V_{ij} \frac{S_{j\alpha}+1}{2} \right] - V_i^T \right\} \tag{3}$$

and $\varepsilon$ is a smearing factor explicitly determined from the details of (a) and (b).

In the real network, the patterns do not evolve by discrete time steps but rather in a continuous manner. However, we have argued [1] that the key features relating to the persistence of patterns in the network are common to the discrete and continuous systems. The discrete model is *much easier* to handle analytically, so we use it in the belief that it describes these essential features correctly.

We are interested in considering the consequences of the network having modifiable synapses of the type introduced by Hebb [5]. Many workers have shown that such an assumption leads to the network "learning". Here we investigate the storage capacity of the network. Our analysis is based on Hebb's neurophysiological postulate: "When an axon of cell $A$ is near enough to excite a cell $B$ and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that $A$'s efficiency as one of the cells firing $B$ is enhanced." Specifically, we assume that $V_{ij}$ is given by the time average $\overline{CS_iS_j}$ of the past correlated firing history of neurons $i$ and $j$ having the factored form

$$V_{ij} = \left( \sum_{\eta,\nu} C_{\eta\nu} S_{i\eta} S_{j\nu} \right) v_i v_j, \qquad C_{\eta\nu} \geqslant 0, \quad v_i \geqslant 0, \tag{4}$$

where the $C_{\eta\nu}$ are *complicated* functions of the long term past history. There are also possible short term facilitation changes in $V_{ij}$ which are crucial in understanding short term memory [6]. Likewise we expect such factors as alertness, attention, etc. will effect the $V_{ij}$. However, we ignore these factors in our analysis of our simplified model.

Now we examine the linearized version of $P_{\beta\alpha}$, which we are able to solve exactly. In the limit that the statistical fluctuations are large, corresponding to small $\varepsilon$ in (2), we have

$$P_{\beta\alpha} \approx \frac{1+\varepsilon M_{\beta\alpha}}{\sum_{\gamma} (1+\varepsilon M_{\gamma\alpha})}. \tag{5}$$

Now from (3), $\sum_\gamma M_{\gamma\alpha}=0$, since $\sum_\gamma S_{i\gamma}=0$; thus (5) becomes

$$P_{\beta\alpha}=\frac{1+\varepsilon M_{\beta\alpha}}{2^N}. \tag{6}$$

(Note that in the limit of very large noise, $\varepsilon\sim 0$, every firing pattern is equally probable and thus each neuron $i$ has an average firing $\overline{S_i}=0$ which corresponds to firing on the average once every two time steps.) We assume the particular conservation restriction

$$\sum_j V_{ij}=2V_i^T; \tag{7}$$

however, we will later show the effects of relaxing this condition. Then substituting (4) and (7) into (3), we have

$$M_{\beta\alpha}=\frac{1}{2}\sum_{\eta\nu} C_{\eta\nu}\Big(\sum_i v_i(S_{i\beta}S_{i\eta})\Big)\Big(\sum_j v_j(S_{j\alpha}S_{j\nu})\Big). \tag{8}$$

We define the $2^N\times 2^N$ matrix

$$T_{\gamma\eta}=\sum_i S_{i\gamma}S_{i\eta}v_i, \tag{9}$$

which occurs twice in (8). This matrix is symmetric and may be put in diagonal form by finding its eigenvalues $\lambda_u$ and eigenvectors $\Psi^u$ ($\sum_\eta T_{\gamma\eta}\Psi_\eta^u=\lambda_u\Psi_\gamma^u$). In the Appendix we show that the eigenvalue spectrum is extremely simple. Only $N$ of the $\Psi^u$'s (denoted by $\psi^a$) have eigenvalue $\lambda_u\neq 0$, and all the $2^N-N$ others (denoted by $\chi^k$) have $\lambda_u=0$:

$$\sum_\eta T_{\gamma\eta}\psi_\eta^a=v_a 2^N\psi_\gamma^a, \qquad a=1,\ldots,N, \tag{10}$$

$$\sum_{\eta=1}^{2^N} T_{\gamma\eta}\chi_\eta^k=0, \qquad k=1,\ldots,2^N-N, \tag{11}$$

where the $\Psi^u$ form an orthonormal basis:

$$\sum_\alpha \Psi_\alpha^u\Psi_\alpha^v=\delta_{uv}. \tag{12}$$

A general form for the $\psi$'s is

$$\psi_\alpha^a=2^{-N/2}\sum_l R_{al}S_{l\alpha}, \tag{13}$$

where the set $l$ have the same eigenvalue $v_a 2^N$ and $R$ is an orthogonal matrix. Now instead of the expansion (4) of $V_{ij}$ in terms of patterns $\eta$, we expand in terms of patterns weighted by the $\Psi^u$:

$$V_{ij} = \sum_{u,r} C_{ur} \sum_{\nu,\eta} S_{i\eta} \Psi^u_\eta S_{j\nu} \Psi^r_\nu v_i v_j. \tag{14}$$

Using (14), along with (10) and (11), we obtain for (3)

$$M_{\beta\alpha} = \frac{1}{2} \sum_{u,r} C_{ur} \sum_{\eta} \left[ \sum_i S_{i\beta} S_{i\eta} v_i \Psi^u_\eta \right] \sum_{\nu} \left[ \sum_j S_{j\alpha} S_{j\nu} v_j \Psi^r_\nu \right]$$

$$= \frac{(2^N)^2}{2} \sum_{ab} C_{ab} \psi^a_\beta \psi^b_\alpha v_a v_b. \tag{15}$$

The $\chi^k$, which have $\lambda_k = 0$, do not contribute to (15), and due to the orthogonality of the $\chi$ and the $\psi$ [and the specific form (13) for $\psi$], the $\chi$'s do not contribute to (14):

$$V_{ij} = \sum_{a,b} C_{ab} \sum_{\nu,\eta} v_i (S_{i\eta} \Psi^a_\eta) v_j (S_{j\nu} \Psi^b_\nu), \qquad C_{ab} > 0$$

$$= 2^N v_i v_j \sum_{a,b} R_{ai} R_{bj} C_{ab}. \tag{16}$$

Then using the specific form in the Appendix for the $\chi$'s, we obtain in the $\Psi$ basis the quantity $\sum_{\beta\alpha} \Psi^r_\beta P_{\beta\alpha} \Psi^u_\alpha \equiv P(\psi^r, \psi^u)$. We note that the term $1/2^N$ in (6) only has a contribution for the one eigenfunction $\chi^1$ which has equal components in each $\alpha$ [see Eq. (22)], since all the other $2^N - 1$ $\Psi$'s have $\sum_\alpha \Psi_\alpha = 0$.

Thus

$$P(\psi^e, \psi^d) = \varepsilon 2^{N-1} \sum_{a,b} v_a v_b C_{ab} \delta_{ae} \delta_{bd}, \tag{17}$$

and $P(\chi^1, \chi^1) = 1$. All the other elements are equal to zero. (Note that although the $C_{ab}$'s and the $v_a$'s must be positive, the synaptic potential strengths can be of either sign.) In our linear model, then, the only non-zero matrix elements aside from $P(\chi^1, \chi^1)$ are between states $\psi^a$ and $\psi^b$ where both $a$ and $b$ lie in the set of $N$ states with non-zero eigenvalues. There are thus $N^2$ independent factors $C_{ab}$ in (17) and (16) which describe how the patterns evolve with time. In our model with every neuron connected to every other neuron, this $N^2$ is the number of synapses. This is the basis of our statement that the memory storage capacity of the network is determined by the number of synapses. We use the term "capacity" in the sense of the number of independent parameters which determine the types

of firing patterns which can be excited and propagate in the network for the order of seconds [5]. We also note that the set of $N$ eigenvectors $\psi^a$ with non-zero eigenvalue form the basis for describing these patterns.

If we relax the conservation restriction (7) on the overall sum of the synaptic strength of a cell, then one of the $N$ non-zero $\lambda_i$ in (10) becomes much larger than the other $N-1$ non-zero ones. (See the Appendix.) This leads to the tendency of this one weighted pattern to dominate the firing activity of the network. There appears to be a need, therefore, for the real network to have some form of feedback to maintain the system near the symmetry point described by the condition (7).

## III. CONCLUSIONS

We have solved exactly a linearized version of our model [1–3] of a network of $N$ neurons having Hebb type modifiable synapses [5]. Previously, we focused on the $2^N$ "microscopic" firing patterns $\{\alpha\}$ describing the firing of each neuron at a given time step. An important result of our present analysis is that *only* $N$ specific (*orthogonal*) combinations, $\psi^a$, of these $2^N$ firing patterns $\alpha$ determine the firing behavior of the network. From these $N$ $\psi^a$'s we form $N^2$ transitions $\psi^a \to \psi^b$ which can be learned [Eq. (17)]. In order to make full use of this capacity, the network, in response to a stimulus, must use these transition elements and thus generate a time sequence of $\psi$'s ($\psi^a \to \psi^b \to \cdots$). Thus as the storage capacity is related to the number of synapses, we expect that the network could be excited into many different time sequences of the $\psi$'s. We are also led to a modified form of the Hebb hypothesis, Eq. (16), in which synaptic changes occur as the result of correlated pre–post-neuronal firing behavior of the *linear* combinations $\psi$.

Two important questions remain to be answered. First, one may ask to what extent our analysis of the linearized model is relevant to the full non-linear problem, and second, what physiological interpretation can be given to the $\psi$'s.

It is difficult to answer the first without a solution to the complete problem. This will be hard to come by, for the complete problem is very much more complicated even than the three dimensional Ising problem, which remains unsolved after almost thirty years of intense study [8]. On the other hand, certain qualitative arguments which will be presented elsewhere, together with computer simulations for a small number of neurons, indicate a close connection between our results for the linearized model and the behavior of the full model. These arguments seem to indicate that associated with the $N^2$ transitions of the linear model one has a set of $N^2$ *classes* of transitions for the general model. The full question, however, requires much more detailed study to determine if this is in fact generally true.

The second question relates to the physiological interpretation of the linear combinations of firing patterns $\psi^a$ which determine the statistical nature of the mathematical solution. A reasonable, but speculative, interpretation has been made by one of us [9], in which assemblies of neurons [10] play a crucial role. Experimental tests are suggested which involve obtaining firing correlations of neighboring neurons using two or more close spaced microelectrodes [11].

Finally, we note some kind of averaging process in the neural network is implied by the $\psi$'s. It is also implied by the well-known [12] fact that although the response of a given neuron for a single stimulus to the animal is not reproducible, the post-stimulus histogram (PSH), or average response of a single neuron in a network to many presentations of the stimulus to the animal, *is* reproducible. Furthermore, the behavioral response to a single presentation of a "meaningful" stimulus is in general reproducible. Thus, we expect that the nervous system does some averaging process for one stimulus, equivalent to that done by the experimenter to obtain the PSH. Whether this averaging is done by assemblies [9, 10] or some other mechanism must be determined by experiment. This averaging presumably must lead to statistically reliable behavior of the network and also limit the consequences of local damage to the network [9, 10].

## APPENDIX

First, we examine the eigenvalues $\lambda_u$ and eigenvectors $\Psi^u$ of the symmetric $2^N \times 2^N$ $T$ matrix, Eq. (9):

$$T_{\gamma\epsilon} = \sum_{i=1} S_{i\gamma} S_{i\epsilon} v_i,$$

where $S_{i\alpha} = \pm 1$ as defined in Eq. (1) and $v_i \geqslant 0$. As a solution of

$$\sum_{\eta=1}^{2^N} T_{\gamma\eta} \Psi_\eta^u = \lambda_u \Psi_\gamma^u, \tag{18}$$

$$\sum_{\eta} \Psi_\eta^u \Psi_\eta^r = \delta_{ur}, \tag{19}$$

consider

$$\psi_\eta^a = S_{a\eta}.$$

Then

$$\sum_{\eta} \sum_{i} S_{i\gamma} S_{i\eta} v_i S_{a\eta} = \sum_{i} S_{i\gamma} v_i \delta_{ai} 2^N = 2^N v_a S_{a\gamma},$$

since

$$\sum_{\eta} S_{i\eta} S_{a\eta} = \delta_{ai} 2^N.$$

Thus

$$\psi_{\eta}^{a} = 2^{-N/2} S_{a\eta}, \qquad a = 1, \ldots, N,$$
$$\lambda_a = 2^N v_a \tag{20}$$

are solutions of (18) and (19). A more general form can be obtained by rotating the $\psi$'s among the degenerate set:

$$\psi_{\eta}^{a} = 2^{-N/2} \sum_{b} R_{ab} S_{b\eta}, \tag{21}$$

the sum over $b$ having the same $v_b = v_a$, with $R$ an orthogonal matrix. We note that all the $2^N$ $\lambda$'s are $\geqslant 0$, since

$$\begin{aligned}
\lambda_u &= \lambda_u \sum_{\gamma} \Psi_{\gamma}^{u} \Psi_{\gamma}^{u} = \sum_{\gamma,\eta} \Psi_{\gamma}^{u} T_{\gamma\eta} \Psi_{\eta}^{u} \\
&= \sum_{i=1}^{N} v_i \left( \sum_{\gamma=1}^{2^N} \Psi_{\gamma}^{u} S_{i\gamma} \right) \left( \sum_{\eta=1}^{2^N} \Psi_{\eta}^{u} S_{i\eta} \right) \\
&= \sum_{i} v_i C_i^2 \geqslant 0
\end{aligned}$$

(with $C_i = \sum_{\eta} \Psi_{\eta}^{u} S_{i\eta}$). Now the trace of $T$,

$$\operatorname{tr} T = \sum_{\gamma} T_{\gamma\gamma} = \sum_{\gamma} \sum_{i} S_{i\gamma} S_{i\gamma} v_i = 2^N \sum_{i} v_i,$$

is equal to the sum of the $N$ eigenvalues (20). Thus the $2^N - N$ remaining eigenfunctions $\chi^k$ all have $\lambda_k = 0$. One $\chi$ has equal components in each pattern $\alpha$:

$$\chi^1 = 2^{-N/2} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix}, \tag{22}$$

whereas all the other satisfy

$$\sum_{\alpha=1}^{2^N} \chi_{\alpha}^{k} = 0, \qquad k = 2, \ldots, 2^N - N.$$

One may verify that all the $2^N$ $\Psi$'s can be written in the direct product form [8].

$$2^{-N/2}\theta_1 \times \theta_2 \times \cdots \times \theta_N,$$

$$\theta = A, B, \qquad A = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \tag{23}$$

The $N$ $\psi$'s, (20), correspond to having just one $B_i$, and the $\chi^1$ corresponds to no $B_i$.

Finally we examine the consequence of relaxing the conservation restrictions given by Eq. (8):

$$\sum_{j=1}^{N} V_{ij} = 2V_i^T.$$

We rewrite (3), adding and subtracting a term $b\Sigma_i S_{i\beta}\Sigma_j V_{ij}$:

$$2M_{\beta\alpha} = \sum_i S_{i\beta} \sum_j V_{ij}(S_{j\alpha} + b) - b \sum_i S_{i\beta}\left(\sum_j V_{ij}\right)$$

$$+ \sum_i S_{i\beta}\left[\left(\sum_j V_{ij}\right) - 2V_i^T\right]. \tag{24}$$

Now choose $b$ so that the second and third terms on the right hand side of (24) cancel. Suppose this value of $b$,

$$b = 1 - \frac{2V_i^T}{\sum_j V_{ij}},$$

is independent of $i$. Then

$$M_{\beta\alpha} = \frac{1}{2} \sum_i S_{i\beta} \sum_j V_{ij}(S_{j\alpha} + b). \tag{25}$$

To handle this case we consider a modified form of the Hebb learning hypothesis [5], i.e.,

$$V_{ij} = v_i v_j \sum_{\eta,\nu} S_{i\eta}(S_{j\nu} + b) C_{\eta\nu}, \tag{26}$$

in order to simplify the analysis of (25). Substituting (26) into (25) we have

$$M_{\beta\alpha} = \frac{1}{2} \sum_{\eta,\nu} C_{\eta\nu}\left[\sum_i v_i(S_{i\beta}S_{i\eta})\right]\left[\sum_j v_j(S_{j\alpha} + b)(S_{j\nu} + b)\right]. \tag{27}$$

The first factor involving the sum over $i$ is just our $T$ matrix considered above. We examine the second factor with all the $v_j$ the same. Defining

$$T'_{\alpha\nu} = \sum_j (S_{j\alpha} + b)(S_{j\nu} + b), \tag{28}$$

we can readily show that $T'$ has one very large eigenvalue $(b^2N+1)2^N$, $N-1$ eigenvalues $2^N$ and the $2^N - N$ remaining eigenvalues equal to zero. This would yield the tendency of the eigenfunction corresponding to the large eigenvalue to dominate the $N-1$ non-zero smaller ones in determining the firing activity of the network. Thus, there appears to be a need for some feedback to maintain the system near the symmetry point described by the condition (7).

## REFERENCES


1 W. A. Little, Existence of persistent states in the brain, *Math. Biosci.* **19**, 101 (1974).
2 G. L. Shaw and R. Vasudevan, Persistent states of neural networks and the random nature of synaptic transmission, *Math. Biosci.* **21**, 207 (1974).
3 W. A. Little and G. L. Shaw, A statistical theory of short and long term memory, *Behav. Biol.* **14**, 115 (1975).
4 J. Ashkin and W. E. Lamb, The propagation of order in crystal lattices, *Phys. Rev.* **64**, 159 (1943); G. F. Newell and E. W. Montroll, On the theory of the ising model of ferromagnetism, *Rev. Modern Phys.* **25**, 353 (1953).
5 D. O. Hebb, *The Organization of Behavior*, Wiley, New York, 1949.
6 J. L. McGaugh, Time dependent processes in memory storage, *Science* **153**, 1351 (1966).
7 B. Katz, *Nerve, Muscle and Synapse*, McGraw-Hill, New York, 1966.
8 K. Huang, *Statistical Mechanics*, Wiley, New York, 1963.
9 G. L. Shaw, Space-time correlations of neuronal firing related to memory storage capacity. *Brain Research Bulletin* **3**, 107(1978).
10. E. R. John, Switchboard versus statistical theories of memory and learning, *Science* **177**, 850 (1972).
11. M. Verzeano and K. Negishi, Neuronal activity in cortical and thalanic networks: A study with multiple microelectrodes, *J. General Physiol.* **43** (Suppl.), 177 (1960).
12 B. D. Burns, *The Uncertain Nervous System*, Edward Arnold, London, 1968; R. C. Dill, E. Vallecalle and M. Verzeano, Evoked potentials, neuronal activity and stimulus intensity in the visual system, *Physiol. and Behav.* **3**, 797 (1968); A. Ramos, E. L. Schwartz and E. R. John, Stable and plastic limit discharge patterns during behavioral generalization, *Science* **192**, 393 (1976); T. J. Tyler, R. A. Roemer and R. F. Thompson, Relations between gross and unit evoked activity in pericruciate cortex of cat, *Physiol. and Behav.* **6**, 375 (1971).