MATHEMATICAL BIOSCIENCES **19**, 101–120 (1974)

The Existence of Persistent States in the Brain

W. A. LITTLE
Department of Physics, Stanford University, Stanford, California

Communicated by S. M. Ulam

ABSTRACT

We show that given certain plausible assumptions the existence of persistent states in a neural network can occur only if a certain transfer matrix has degenerate maximum eigenvalues. The existence of such states of persistent order is directly analogous to the existence of long range order in an Ising spin system; while the transition to the state of persistent order is analogous to the transition to the ordered phase of the spin system. It is shown that the persistent state is also characterized by correlations between neurons throughout the brain. It is suggested that these persistent states are associated with short term memory while the eigenvectors of the transfer matrix are a representation of long term memory. A numerical example is given that illustrates certain of these features.

1. INTRODUCTION

In this paper we examine the long term behavior of a neuronal network such as the human brain. We will start from the assumption that the state of the brain at any time may be described by a configuration defined by the set of neurons that have fired within a certain specified recent interval of time and those that have not. We shall examine under what conditions a correlation can exist between states so defined, which are separated by a long period of time. In this context a long period of time is considered to be a time long compared to the refractory period of a neuron. The underlying reason for studying this problem is the belief that the states or configurations defined above are in some way related to the thought processes or experiences sensed by the individual and that in our own experience a long term correlation appears to exist in the latter. If our belief should prove to be valid it would imply a long term correlation between the neuronal configurations. We do not offer any proof that the thought processes and the neuronal configurations are representations of the same thing but suggest that this is a reasonable working hypothesis. Its proof or disproof lie beyond the scope of this paper.

B. Cabrera et al. (eds.), *From High-Temperature Superconductivity to Microminiature Refrigeration*

Starting from the state defined above and by using the known behavior of the neuron and the interneuronal connections, we will show how the state of the brain evolves with time. We find that a close analogy exists between this problem and the problem of an interacting spin system which has been studied extensively in statistical mechanics and in solid state physics. The existence of persistent states in the neuronal network corresponds to the occurrence of long range order in the spin problem and both are related to the existence of a degeneracy of the maximum eigenvalue of a certain matrix. This matrix in the neuronal system is determined by the topology of the neuronal network, the size and nature of the synaptic junctions and the various electrochemical potentials in the brain. While the problem of determining the detailed behavior of this matrix is formidable, certain general conclusions can be drawn from this analysis which are interesting. We find that while the number of possible states in the brain as defined above is enormous—of the order of 2^N where N is the number of neurons (of the order of 10^{10} in the human brain)—the number of states which determine the long term behavior is a very much smaller number. This represents a tremendous simplification. If these states could be identified it would provide great insight into the operation of the brain. A second feature that follows from this is that these persistent states are distinguished by the property that a coherence or correlation exists between the neurons *throughout* the entire brain or large portions of it. These states are thus a property of the brain as a whole rather than a localizable entity. Thirdly, one finds that the transformation from the uncorrelated to the correlated state in a portion of, or in the whole brain can occur by the variation of the mean biochemical concentrations in these regions, and that this transformation occurs in a manner closely similar to the phase transition in the analogous spin system. Our results suggest how items of memory stored, in our model, in the topology of the interneuronal connection and the properties of the synaptic junctions, may be recalled to active use as a pattern of firing neurons.

Our results are not exact but have been reached only after making certain simplifying assumptions. While the assumptions themselves appear to be somewhat innocuous we have not been able to dispense with them and, whether our conclusions would remain if we could, has still to be proven. This work should thus be viewed as suggestive rather than definitive.

2. SIGNIFICANCE OF PERSISTENT STATES

Our thesis is based on the argument that the existence of states or behavior in which a correlation exists for long periods of time, are of

prime importance to an understanding of the brain. There are many levels at which the significance of these persistent states can be appreciated. Clearly the ability of an animal to entertain one concept such as "flight from a predator" over a period of several minutes has high survival value. Likewise the capability of retaining the details of a recent attack over a longer period so as to analyze and review possible alternative defensive tactics in the future is similarly of survival value. In general, the ability to retain a theme or leit-motif while examining its many consequences over a period of time is of broader value at a more sophisticated level. The ability to meditate over long periods of time without external stimuli is another example of the existence of a long time correlated state of mind. At a much deeper level the conviction each of us has of our own existence is based to some extent on a certain internal continuity and coherence of behavior of ourselves over a long period of time.

On these admittedly imprecise, but suggestive grounds we argue that long term correlations exist within the brain. Assuming that these correlations imply correlations in the neuronal configurations, we ask what the conditions are in order for the latter to occur.

3. PHYSIOLOGICAL CONSIDERATIONS

The basic physiological structure of the brain has been studied extensively [1]. The important building blocks are the neurons which are bulbous nerve fibers illustrated in Fig. 1. Each neuron can be activated by

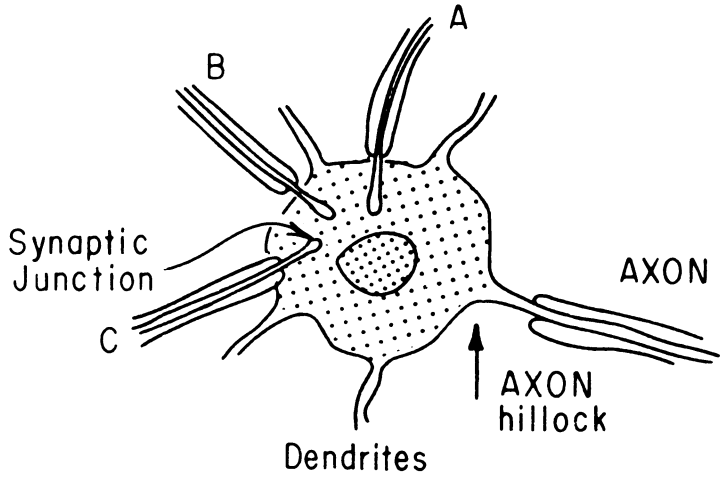

FIG. 1. Schematic view of typical neuron.

the flow of an activating chemical across the synaptic junctions from the axon of one neuron either onto the surface or onto the dendrites of another neuron. Alternately, the neuron may be inhibited by inhibitory synapses at which an inhibiting chemical is transmitted across the synaptic gap to the neuron. The transmission of these chemicals causes a change in the ionic concentration within the neuron and this results in a change of its electrochemical potential. These electrical effects are referred to as excitatory postsynaptic potentials, or inhibitory postsynaptic potentials, respectively. If the net potential at the axon hillock resulting from all the excitatory and inhibitory post synaptic potentials exceeds a certain threshold level the neuron "fires" and an action potential propagates down the elongated tail of the neuron termed the axon and normally terminates on the synaptic junction of another neuron. The arrival of the action potential at this synapse triggers the release of the activating or inhibiting chemical, thus activating or inhibiting the next neuron and so the process continues.

The potential of the neuron is determined by the integrated effect of all the excitatory and inhibitory post synaptic potentials delivered to it over an integrating period of several msec. This is the period of latent summation. If the threshold is reached and it fires, a sharp positive pulse appears followed by a negative going excursion [2]. The potential then returns to the resting potential after a few msec. During this latter refractory period the neuron is recovering and cannot fire again.

We note also that the velocity of propagation of the action potential along the axon is about a 10^2 cm/sec in the axons within the brain and the mean length of axon from one neuron to the synaptic junction of another is no more than about 10^{-2} cm. Thus the flight time of a signal from one neuron to the next ($\approx 10^{-1}$ msec) is appreciably less than the refractory period of a neuron.

Our model is based on using in simplified form these various facts in order to determine how the state of the brain evolves with time.

4. MODEL SYSTEM

First, let us consider a neural network in which there are no connections to nerve cells which lie outside the network itself. Thus we consider the network as isolated from external stimuli. Alternately, one may consider the network as part of the brain but situated in a deprived environment receiving no external stimuli. Later we shall consider how this restriction can be removed or relaxed.

Second, we shall suppose that the neurons are not permitted to fire at any random time but rather that they are synchronized such that they

can only fire at some integral multiple of a period τ which is of the order of the refractory period of the neuron. We suppose that the net neuron potential at the end of this period is determined by the sum of all the excitatory and inhibitory post synaptic potentials that occur during the period. The value of this potential at the end of the period then determines whether the neuron will fire or will not fire. Further we suppose that the influence of these potentials decays to a negligible value by the time the following opportunity to fire occurs. Thus within each period we may consider each neuron as starting with a clean slate, or in other words that these processes are Markoffian. Later we will consider how this limitation may be relaxed.

Third, we will assume that the connections between the neurons via the axons; and the properties of the synaptic junctions themselves are all fixed and do not change with time. We are not concerned here with learning behavior in which changes might be expected to occur in some one or other of these connections but rather we are interested in the behavior of the network in which these properties are assumed fixed. We believe and later will give arguments to bolster this belief that these given properties represent hereditary information or long term memory while the pattern of firing neurons defined by our "states" are related to short term memory involving the active state of the mind.

These assumptions appear to impose some rather artificial constraints upon the system. They are imposed for reasons of mathematical convenience. By so doing we are able to calculate certain properties of the system. However, we will argue on physical grounds by analogy with other related systems that the properties, which we calculate with these constraints, can be expected to remain even when certain of the constraints are relaxed.

With the above constraints we can define the "state of the brain" by the configuration determined by the set of neurons that have fired most recently and those that have not. It is convenient to write this using terminology borrowed from quantum mechanics [3]. We define that state of the brain at time t as the configuration

$$\psi(t) \equiv |s_1, s_2, \ldots, s_N\rangle, \tag{1}$$

where $s_i = +1$ if the ith neuron has just fired and $s_i = -1$ if the ith neuron has not fired, and N is the total number of neurons. There are thus 2^N states defined.

The neurons which have fired will then send a signal down their axons to their terminating synaptic junctions. These signals will trigger the release of the excitatory or inhibitory chemicals which in turn will raise or lower the potential of the neurons to which they are attached. Let us

define V_{ij} as the resulting change in potential of the ith neuron due to an activating signal arriving from the jth neuron at the synaptic junction connecting the axon of j to the neuron i. This incremental potential V_{ij} may be positive (excitatory), negative (inhibitory), or may be zero if no synaptic connection exists between j and i. It will also depend upon the size and structure of the synaptic junction and the size or volume of the neuron to which it is attached. We will assume further that the V_{ij} are fixed and do not change with time.

We make use of the fact that the various postsynaptic potentials sum within a period so that the net change of potential of the ith neuron will be given by the sum of the various contributions from the activated synapses. If the existing state of the brain is given by (1) then the sum of the post synaptic potentials of the ith neuron may be written as

$$\sum_j V_{ij}\left(\frac{s_j + 1}{2}\right).$$

We see that if $s_j = +1$ we get a contribution, V_{ij} to the sum but if $s_j = -1$ we get no contribution. If the total potential exceeds some threshold value V_0 the neuron will probably fire. It is convenient to express this mathematically as follows. Let $p(+1)$ be the probability that the ith neuron will fire then $p(+1)$ may be written as

$$p(+1) = \frac{1}{\exp -\beta\left\{\left[\sum_j V_{ij}\left(\frac{s_j + 1}{2}\right)\right] - V_0\right\} + 1}. \tag{2}$$

The behavior of this function is shown in Fig. 3. If the sum is appreciably less than V_0 the exponential is large and $p(+1)$ is small, i.e. the probability of firing is small. On the other hand if this sum is appreciably greater than V_0 then the exponential becomes small and $p(+1)$ approaches unity, i.e. the neuron almost certainly fires. The factor β gives a measure of the uncertainty in the width of the threshold region.

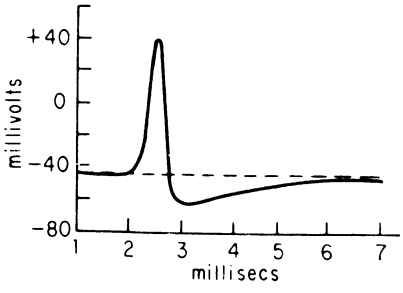

FIG. 2. Axon potential showing positive going signal and refractory period [2].

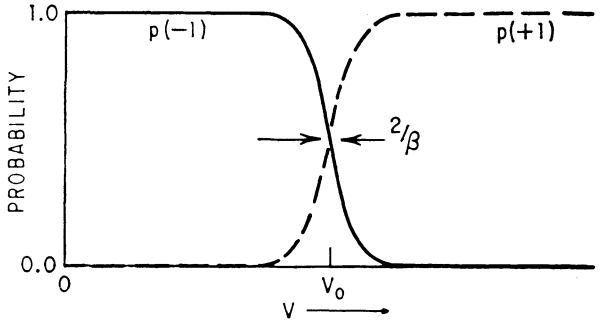

FIG. 3. Probability of neuron firing $p(+1)$ or not firing, $p(-1)$ as a function of the sum of the post-synaptic potentials V relative to the threshold, V_0.

The probability of *not* firing, $p(-1)$ is given by $1 - p(+1)$ and thus can be simplified to give

$$p(-1) = \frac{1}{\exp + \beta\left\{\left[\sum_j V_{ij}\left(\frac{s_j + 1}{2}\right)\right] - V_0\right\} + 1}. \tag{3}$$

Notice that both (2) and (3) may be expressed as

$$p(s_i') = \frac{1}{\exp - \beta s_i'\left\{\left[\sum_j V_{ij}\left(\frac{s_j + 1}{2}\right)\right] - V_0\right\} + 1}. \tag{4}$$

Strictly speaking β and V_0 should also be considered as dependent on i but to simplify the problem we will treat them as constants throughout the network.

We may use these expressions to compute the probability of obtaining a state $|s_1', s_2', \ldots, s_N'\rangle$ given a state $|s_1, s_2, \ldots, s_N\rangle$ immediately preceding it. Using the usual bra and ket notation of quantum mechanics [3] and defining an operator P which yields this probability, Eq. (4) then gives us the result that:

$$\langle s_1', \ldots, s_N'|P|s_1, \ldots, s_N\rangle = \prod_{i=1}^{N}\left(\frac{1}{\exp - \beta s_i'\left\{\left[\sum_j V_{ij}\left(\frac{s_j + 1}{2}\right)\right] - V_0\right\} + 1}\right). \tag{5}$$

This then defines a $2^N \times 2^N$ matrix whose elements give the probability of a particular state $|s_1, s_2, \ldots, s_N\rangle$ yielding after one cycle the new state $|s_1', s_2', \ldots, s_N'\rangle$. The primed set refer to the row, $\langle s_1', s_2', \ldots, s_N'|$ and the unprimed set to the column $|s_1, s_2, \ldots, s_N\rangle$ of the element of the matrix.

5. ANALOGY WITH A SPIN SYSTEM

Having expressed the problem in this form one can see immediately the close similarity between it and the problem of an Ising system. Kramers and Wannier [4] in their classic paper on the Ising problem showed how the partition function for that spin problem could be expressed in terms of a matrix. The similarity is made more apparent by rewriting (5) in the form:

$$\langle s_1', \ldots, s_N' | P | s_1, \ldots, s_N \rangle = \frac{\prod_{i=1}^{N} \exp \beta \frac{s_i'}{2} \left\{ \left[\sum_j V_{ij} \left(\frac{s_j + 1}{2} \right) \right] - V_0 \right\}}{\prod_{i=1}^{N} \sum_{s_i' = \pm 1} \exp \beta \frac{s_i'}{2} \left\{ \left[\sum_j V_{ij} \left(\frac{s_j + 1}{2} \right) \right] - V_0 \right\}} \tag{6}$$

This should be compared with Eq. 3.5 of the review of Ferromagnetism by Newell and Montroll [5].

The methods for handling this type of Ising problem have been discussed at length in many papers and texts. We refer the reader to the review of Newell and Montroll [5], Huang's text on Statistical Mechanics [6] and, for what follows, the paper by Ashkin and Lamb [7]. In the latter the question of the propagation of order in a crystal lattice is discussed. We shall study the analogous problem to this in the neural network. In the crystal problem one considers the configuration of atomic spins on a row of atoms with each atom having a spin of one half. One such configuration can be described by a state analogous to Eq. (1) i.e. $|s_1, s_2, \ldots, s_N\rangle$, where $s_i = +1$ means spin i is "up" and $s_i = -1$ that spin i is "down." The total, N refers to the number of atoms in the row. Due to their interaction with the spins on the next row the probability of a configuration $|s_1', s_2', \ldots, s_N'\rangle$ occurring in the next row can be calculated. An expression somewhat similar to (6) is then obtained. By using the same process again and again one can calculate the probability of obtaining a particular configuration in the mth row. In that problem a question of great importance is whether or not a correlation can exist between a configuration in row q, say, and row r where the distance between q and r becomes very large. When such a correlation does occur we say that long range order exists in the lattice. For a spin system which becomes ferromagnetic it is found that long range order sets in at the Curie point and exists at all temperatures below that. As shown by Lassettre and Howe [8] and discussed further by Ashkin and Lamb the onset of this long range order is intimately associated with the occurrence of a degeneracy of the maximum eigenvalue of the matrix analogous to (6).

In our problem a configuration determines the state of the brain at a particular instant of time. This corresponds to the configuration of spins

in *one row* of the lattice in the crystal problem. The configuration which describes the state of the brain after the *next cycle* corresponds to the spin configuration *in the next row*. Thus the existence of a correlation between two states of the brain which are separated by a long period of time is directly analogous to the occurrence of long range order in the corresponding spin problem. We will show that the occurrence of these persistent states is also related to the occurrence of a degeneracy of the maximum eigenvalue of the matrix P given in (5).

We draw the analogy between the neural network and the two dimensional Ising problem. The configuration of N spins in one row corresponding to the configuration of N neurons at one particular instant of time. An analogy could equally well be drawn between a three dimensional Ising problem and the neural network. To do this one would associate the N spins in one *layer* of the crystal with N neurons at one instant of time: the next layer being associated with the next instant of time. Either analogy is equally good, for we note that in the neural problem we have no interaction terms V_{ij} such that i and j are both in the primed set or both in the unprimed set of s_i's. This would correspond in the 2-D spin problem to an interaction between spins on adjacent rows only, with no interaction between spins on the same row. Thus the geometric arrangements of the spins in a row (as in the 2-D analogy) or a layer (as in the 3-D analogy) is irrelevant and we may thus use either.

A nontrivial difference between our problem and the spin problem is that in the spin problem the matrix corresponding to P is symmetric. It is thus diagonalizable. In our case we have no guarantee that the matrix P will be diagonalizable without knowing the neuron connections. Moreover, it is reasonably certain that the matrix P will not be symmetric because the signals very clearly propagate from one neuron down its axon to the synaptic junction of the next neuron and not in the reverse direction. Only by accident would one have an identical path also running in the reverse direction. However, we will make the assumption here that P is diagonalizable. Later we will show how our argument may be extended to the situation in which P is not diagonalizable.

The occurrence of persistent states in the neural network will now be examined using a similar approach to that used for the study of long range order in the spin systems.

6. LONG RANGE ORDER AND PERSISTENT STATES

First, it is useful to note that the probability of obtaining a configuration $|s_1', \ldots, s_N'\rangle$ after two cycles is

$$\sum_{s_1'', \ldots, s_N''} \langle s_1', \ldots, s_N'|P|s_1'', \ldots, s_N''\rangle\langle s_1'', \ldots, s_N''|P|s_1, \ldots, s_N\rangle, \qquad (7)$$

or in matrix notation

$$\langle s_1', \ldots, s_N' | P^2 | s_1, \ldots, s_N \rangle, \tag{8}$$

and thus after m cycles

$$\langle s_1', \ldots, s_N' | P^m | s_1, \ldots, s_N \rangle. \tag{9}$$

To contract our notation let us use $\psi(\alpha)$ to represent the state $|s_1, \ldots, s_N\rangle$ and $\psi(\alpha')$ for $|s_1', \ldots, s_N'\rangle$. Then it is useful to represent these in terms of the eigen vectors φ_r of the operator P. There are 2^N such eigen vectors each of which has 2^N components $\varphi_r(\alpha)$, one for each configuration α. Thus we have

$$\psi(\alpha) = \sum_r \phi_r(\alpha), \tag{10}$$

and assuming, we normalize $\varphi_r(\alpha)$ to unity, so

$$\sum_\alpha \phi_r(\alpha)\phi_s(\alpha) = \delta_{rs}, \tag{11}$$

we obtain

$$\langle s_1', \ldots, s_N' | P | s_1, \ldots, s_N \rangle = \sum_r \lambda_r \phi_r(\alpha')\phi_r(\alpha), \tag{12}$$

where λ_r, is the rth eigenvalue.

We wish to find now the probability of obtaining a particular configuration α'. In general we do not know the initial conditions so we will set up the problem in such a way that the initial conditions play no role. One way to do this is to allow the system to run for a total of M cycles where M is very large and ask for the probability of obtaining a configuration α after m cycles, and then average over all initial configurations and sum over all final configurations. A simpler procedure which gives the same result is to assume that after M cycles the system returns to the initial configuration and we average over all initial configurations. This corresponds to cyclic boundary conditions in the spin problem. Using (9) we obtain the probability $\Gamma(\alpha_1)$ of obtaining the configuration α_1 after m cycles.

$$\Gamma(\alpha_1) = \sum_\alpha \langle \alpha | P^{M-m} | \alpha_1 \rangle \langle \alpha_1 | P^m | \alpha \rangle \Big/ \sum_\alpha \langle \alpha | P^M | \alpha \rangle, \tag{13}$$

which from (12) gives

$$\Gamma(\alpha_1) = \sum_\alpha \sum_{r,u} \phi_u(\alpha)\lambda_u^{M-m}\phi_u(\alpha_1)\phi_r(\alpha_1)\lambda_r^m\phi_r(\alpha) \Big/ \sum_r \lambda_r^m. \tag{14}$$

Using (11)

$$\Gamma(\alpha_1) = \sum_r \lambda_r^M \phi_r^2(\alpha_1) \Big/ \sum_r \lambda_r^M. \tag{15}$$

We notice that this is independent of m and hence gives the probability at any time of obtaining α_1, assuming no constraints or knowledge of the system at an earlier time.

Next we ask what the probability is of obtaining a configuration α_2 after l cycles given that we know we have configuration α_1 after m cycles. This joint probability $\Gamma(\alpha_1, \alpha_2)$ is given by

$$\Gamma(\alpha_1, \alpha_2) = \sum_{\alpha} \langle \alpha | P^{M-l} | \alpha_2 \rangle \langle \alpha_2 | P^{l-m} | \alpha_1 \rangle \langle \alpha_1 | P^m | \alpha \rangle \Big/ \sum_{\alpha} \langle \alpha | P^M | \alpha \rangle, \tag{16}$$

$$= \sum_{\alpha} \sum_{r,u,v} \lambda_v^{M-l} \phi_v(\alpha) \phi_v(\alpha_2) \lambda_u^{l-m} \phi_u(\alpha_2) \phi_u(\alpha_1) \lambda_r^m \phi_r(\alpha_1) \phi_r(\alpha) \Big/ \sum_r \lambda_r^M, \tag{17}$$

which again, through the use of (11), gives

$$\Gamma(\alpha_1, \alpha_2) = \sum_{r,u} \lambda_r^{M-l+m} \lambda_u^{l-m} \phi_r(\alpha_2) \phi_u(\alpha_2) \phi_u(\alpha_1) \phi_r(\alpha_1) \Big/ \sum_r \lambda_r^M. \tag{18}$$

If M and $l - m$ are large numbers then the only significant contribution to (18) will come from the maximum eigenvalues. Let us assume first that these eigenvalues are nondegenerate then (18) gives

$$\Gamma(\alpha_1, \alpha_2) = \phi_{\max}^2(\alpha_2) \phi_{\max}^2(\alpha_1). \tag{19}$$

while the probability of obtaining α_1 is obtained from (15) giving

$$\Gamma(\alpha_1) = \phi_{\max}^2(\alpha_1). \tag{20}$$

Thus we have

$$\Gamma(\alpha_1, \alpha_2) = \Gamma(\alpha_2) \cdot \Gamma(\alpha_1). \tag{21}$$

In this situation we see that the joint probability is just the product of the probabilities of obtaining the two configurations independently. In this case the influence of configuration α_1, does not affect the probability of obtaining the configuration α_2. The network then does not have any persistent states.

On the other hand if the maximum eigenvalue of P is degenerate then the degenerate eigenvalues contribute in the sum of (18). Consider the simplest case when $\lambda_{\max}$ is doubly degenerate having eigen functions φ_1 and φ_2, then (18) becomes

$$\begin{aligned}\Gamma(\alpha_1, \alpha_2) = \{ & \lambda_1^M \phi_1^2(\alpha_2) \phi_1^2(\alpha_1) + \lambda_2^M \phi_2^2(\alpha_2) \phi_2^2(\alpha_1) \\ & + \lambda_1^{M-l+m} \lambda_2^{l-m} \phi_1(\alpha_2) \phi_2(\alpha_2) \phi_2(\alpha_1) \phi_1(\alpha_1) \\ & + \lambda_2^{M-l+m} \lambda_1^{l-m} \varphi_2(\alpha_2) \phi_1(\alpha_2) \phi_1(\alpha_1) \phi_2(\alpha_1) \} / (\lambda_1^M + \lambda_2^M),\end{aligned} \tag{22}$$

and

$$\Gamma(\alpha_1) = (\lambda_1^M \phi_1^2(\alpha_1) + \lambda_2^M \phi_2^2(\alpha_1)) / (\lambda_1^M + \lambda_2^M). \tag{23}$$

In this case $\Gamma(\alpha_1, \alpha_2)$ no longer factorizes as in (21). This result also holds even if λ_1 and λ_2 are not strictly degenerate but are sufficiently close in value for $|\lambda_1^M| \approx |\lambda_2^M|$. This is of some importance because of a

theorem of Frobenius [7] which shows that the maximum eigenvalue of a matrix whose elements are all positive is nondegenerate. P is such a matrix as can be seen from Eq. (6). However, if the elements are *sufficiently* small then a practical degeneracy such that $|\lambda_1| \approx |\lambda_2^M|$ can still occur and $\Gamma(\alpha_1, \alpha_2)$ will no longer factorize. The probability of obtaining a configuration α_2 is then dependent upon the configuration α_1, and thus the influence of α_1 persists for an arbitrarily long time. This influence results from the two last terms in Eq. (22) and these involve the two eigenvectors associated with the degenerate maximum eigenvalues. We thus have the possibility of states occurring within the brain which are correlated over arbitrarily long periods of time. It is worth noting too that the characteristics of the states which so persist are describable in terms of the eigenvectors associated only with the degenerate maximum eigenvalues. In this sense these persistent states are very much simpler to describe than an arbitrary state of the brain for they involve only that small set of eigenvectors associated with the degenerate maximum eigenvalues, whereas other states of the brain are describably, in general, in terms of the full set of 2^N eigenvectors. This represents in principle, a very beautiful simplification of the behavior of the brain.

We have assumed that the cycle time, τ is of the order of a few msec. A time period t then corresponds to t/τ cycles or powers to which we raise the operator P. A few seconds thus corresponds to about a thousand cycles. For a correlation to exist for even a few seconds then the maximum eigenvalues must be degenerate to within a small fraction of a percent.

Another consequence of a degeneracy of the maximum eigenvalue is that under these circumstances and only under these can a correlation exist between neurons that are widely separated in the brain. To show this we define the topological "distance" between two neurons as the integer n_{ij} equal to the smallest number of synaptic junctions one need cross to get from the one neuron i to the other, j moving always from axon to neuron and not vice-versa. In general we expect this integer to be large for neurons which are widely separated and $n_{ij} \neq n_{ji}$. In order for the firing of i to influence the state of j, at least n_{ij} cycles must occur, so in order for neurons i and j to be correlated we need to have a state which persists for a period of time at least as long as $n_{ij}\tau$. The condition then for large spatial correlations is identical to the condition for persistent states, i.e. a degeneracy of the maximum eigenvalue of P. We see thus that the persistent states are characterized by a coherent or correlated behavior of the neurons throughout the brain or at least within large portions of it. By analogy with the spin system the long range order exists not only from row to row (i.e. in time) but also down the rows themselves (i.e. in space).

Finally we note that whether or not persistent order exists is determined by the properties of the matrix P, which in turn are dependent upon the parameters β, V_{ij}, and V_0 given in (6). In the analogous spin system β would be equal to $1/kT$ where k is Boltzmann's constant and T, the temperature; V_{ij} the interaction energy of the ith and jth spin and $\{\Sigma_j(V_{ij}/2) - V_0\} = H$, the interaction energy of a spin with a static magnetic field. (This can be seen by comparing (6) with Eq. (3.5) of Ref. [5].) The transition to the state of long range order occurs at the Curie point which, in the absence of a magnetic field, is determined by the ratio of V_{ij}/kT. The presence of the field, $H = \{\Sigma_j(V_{ij}/2) - V_0\}$ causes a shift of the Curie point and one finds a phase boundary in the T, H plane separating the ordered phase from the disordered phase. We expect therefore that the transition to the persistent state in the neural network would likewise be determined by the relative magnitudes of V_{ij}, H, and $1/\beta$. In our choice of the simple expression (4) we have lumped all the spread in the uncertainty of firing of the neuron in the parameter β. In the actual system we would expect an uncertainty in the size of V_{ij} and some fluctuation in the magnitude of H, both of which would be dependent on the local physiochemical conditions at the neuron. In our model these are the sources of the fluctuations which give rise to the finite width of the threshold curve. In our approximation we represent it by the single parameter β. By analogy to the spin system we would expect that for fixed values of the set of $\{V_{ij}\}$ a phase boundary could be defined in the H_0, $1/\beta$ plane. Or, more generally, a surface separating the ordered from the disordered state could be defined in the H, $1/\beta$, $\{V_{ij}\}$ space. In the simplest model we may assume that the set of $\{V_{ij}\}$ are scaled by the same factor γ and that the phase boundary is thus described by a surface in the three dimensional space, H, $1/\beta$, γ. We expect therefore that a change in the general physiochemical environment of the neurons which give rise to a shift in H or γ, could thus drive the network or a portion of the network across the phase boundary, transforming that portion from the coherently ordered, persistent state to the disordered state, or vice versa. In the Appendix we illustrate this with a numerical example. Our analogy suggests that such a transition cannot occur continuously but must occur discontinuously just as for other phase transitions in the solid state or for the liquid-gas transition.

One additional point worth stressing is that the expression $\Sigma_\alpha \langle \alpha | P^M | \alpha \rangle$, is directly analogous to the partition function for a lattice of M rows of N atoms in the spin problem so that all the corresponding behavior of the neuronal network can be deduced from it. This describes the time averaged behavior of the network because it involves the sum over configurations during a period of time, $t = M\tau$.

7. DISCUSSIONS OF ASSUMPTIONS OF THE MODEL

Our results have been derived on the basis of five principal assumptions or approximations. These are first, that the network has no external stimuli, second, that the probability of the neurons firing is a Markov process, third, that the neurons are synchronized, fourth, that the transfer matrix, P is diagonalizable, and fifth, that the properties of the synaptic junctions are fixed in time. We will discuss these in turn.

ROLE OF EXTERNAL STIMULI

There are two obvious ways in which one can take into account the presence of external stimuli. If the number of synapses from external sensors is small compared to the number of synapses connected to neurons within the network one could use perturbation theory to calculate the changes in the eigenvectors and eigenvalues of the matrix due to external stimuli. The sum over $\Sigma_j V_{ij}((s_j + 1)/2)$ in (4) would be replaced by

$$\left(\sum_j V_{ij}\left(\frac{s_j + 1}{2}\right) + \sum_k W_{ik}\left(\frac{e_k + 1}{2}\right)\right)$$

where W_{ik} is the postsynaptic potential of the ith neuron arising from a signal at the synapse from the kth external source, and $e_k = \pm 1$ depending whether such a signal is present or not. If we treat the sum over W_{ik} as a perturbation when we can show from standard perturbation theory [3] that to first order these terms cause a shift in the eigenvalues, with the eigenvectors remaining unchanged. For larger W_{ik} changes will occur in the eigenvectors as well. Thus, in principle, one could take these effects into account in this way, however, this procedure does not cast much light on the role the sensory inputs would play and we propose a second way of looking at this aspect of the problem.

We suggest that the input signals play a somewhat different role from the interneuronal signals. We know that a strong external signal results in a rapid series of nerve pulses at the synaptic junction. The effect of such a barrage of signals would be to generate a fairly constant average value for the term $(\overline{\Sigma_k W_{ik}((e_k + 1)/2)})$. Adding this average to V_0 transforms the effective threshold V_0 to a new threshold $(V_0 - \overline{\Sigma_k W_{ik}((e_k + 1)/2)})$ where the bar represents a time average. We suggest that this shift could drive the network or parts of the network across the phase boundary from the ordered to the disordered state or vice versa. Thus the external stimuli could play the role of initiating the onset of this persistent state and similarly other stimuli could terminate this state. If this view is correct our model suggests that the persistent state is a representation of long term memory. Which particular memory trace is uppermost would be

determined by the eigenvalues which become the degenerate maximum eigenvalues under a particular form of the external stimuli given by the set of $\{e_k\}$.

We suggest that the different eigenvectors of the matrix P represent certain memories. Under external stimulus or stimulus from some other portion of the brain the eigenvalues are perturbed so two or more become degenerate and larger than any others. A new persistent state will then evolve dominated by the structure of the eigenvectors corresponding to these maximum eigenvalues and with initial conditions determined by the external stimuli, and thus the active state of the brain carry the information contained in the eigenvectors of P. As we pointed out earlier these eigenvectors are determined by the interneuronal connections and the strength of the synaptic junctions as given by the set of V_{ij}'s. Changes in these would change the properties of the eigenvectors. So in this model the process of learning would be any process which resulted in changes in V_{ij} and thus in the eigenvectors of P.

The above model of memory would then require that the sites where information is stored would be highly delocalized. This follows because the eigenfunctions of P are built up of contributions from neurons throughout the entire network. This can be seen from the inverse relationship to (10),

$$\phi_r = \sum_{\alpha} \psi_r(\alpha), \tag{24}$$

where, in general, contributions to φ_r come from all configurations, α.

MARKOV ASSUMPTION

We have made the assumption that only the most recent signals which reach a particular neuron determine whether it is to fire or not. We may relax this at the cost of greater mathematical complexity. This may be done as follows. Instead of describing the state of the brain by the configuration of neurons which have fired on the last cycle only we can describe it by the combination of the last two or more cycles. The transfer matrix then becomes correspondingly larger but can be handled in exactly the same way so that all our arguments go through as before. This generalization in the spin problem corresponds to including both near neighbors and next nearest neighbors, next next neighbors, etc. This can be handled in the standard way [5] using an expanded transfer matrix.

SYNCHRONIZATION OF THE NEURONS

Our third principal assumption was that the neurons could fire only at certain prescribed times. Clearly the method we have used hangs heavily upon this assumption requiring as it does an evolution of the neuronal configuration in a discontinuous manner. Our method cannot

simply be modified to take into account a continuous evolution with time. To do this some other method would need to be devised to determine the nature of the long term correlations between the neuronal configurations. This is a formidable task. However, we suggest that the essential feature derived above, i.e. that a sharp distinction can be drawn between states of persistent order and those without this characteristic, will remain in a model in which the neurons are permitted to fire at arbitrary times. We base this conjecture on the following argument.

We have repeatedly invoked the analogy between the neural network and a two dimensional spin system. The model spin system in turn gives a remarkably good description of the phase transition from the paramagnetic to the ferromagnetic phase and of the liquid-gas phase transition. The two dimensional Ising system has been solved exactly using the matrix method. For its solution by this method one requires a strictly regular array of spins. The problem of a disordered lattice of spins cannot be solved by the matrix method because the interaction of one row of spins with the next cannot be uniquely defined where the concept of the row itself is lost as a result of the disorder. Yet we know from physical measurements that the occurrence of a ferromagnetic phase transition is not strongly dependent upon a high degree of order in the crystalline lattice. Indeed in an amorphous material such transition can still occur. Likewise the actual thermodynamic behavior of the liquid-gas transition is quite well described by the behavior of the analogous transition of the lattice gas [9]. In the lattice gas model the particles are only allowed to occupy mesh points on a regular lattice while in the real gas a particle can, of course, occupy any position. In spite of this difference the phase transition of the lattice gas is remarkably similar to that of a real gas [10]. This shows that the regularity of the lattice is not essential for the occurrence of the phase transition, it merely provides a mathematically convenient way of handling the problem.

In our model our assumption of the strictly regular synchronism of the firing of the neurons corresponds to a strictly regular crystalline array. By analogy with the above we suggest that just as the regularity of the lattice is not essential for the occurrence of the phase transition of the spin system or of the lattice gas, so the regularity in the firing is not an essential requirement for the occurrence of a transition to an ordered persistent state in the neural network. We believe therefore that our conclusions should remain even in a more realistic model.

DIAGONALIZABILITY OF THE CHARACTERISTIC MATRIX

Our fourth principal assumption is that the matrix P is diagonalizable. Without a knowledge of the topology and strength of the various terms

in P we cannot tell a priori whether or not P can be diagonalized. We can show, however, that our results can be generalized to an arbitrary matrix, for while a general matrix cannot always be diagonalized it can be reduced to, so called, Jordan Canonical form [11]. In this form eigenvalues occur along the main diagonal with ones or zeros on the diagonal immediately below it. For example

$$P = \begin{vmatrix} \Lambda_1 & 0 & 0 & 0 & 0 & . \\ 0 & \Lambda_2 & 0 & 0 & 0 & . \\ 0 & 0 & \Lambda_3 & 0 & 0 & . \\ 0 & 0 & 0 & \Lambda_4 & 0 & . \\ 0 & 0 & 0 & 0 & \Lambda_5 & . \\ . & . & . & . & . & . \end{vmatrix} \quad \text{where} \quad \Lambda_i = \begin{vmatrix} \lambda_i & 0 & 0 & 0 & 0 & . \\ * & \lambda_i & 0 & 0 & 0 & . \\ 0 & * & \lambda_i & 0 & 0 & . \\ 0 & 0 & * & \lambda_i & 0 & . \\ 0 & 0 & 0 & * & \lambda_i & . \\ & . & . & . & . & . \end{vmatrix} \tag{25}$$

where * = 0 or 1.

A representation of $\psi(\alpha)$ can now no longer be made in terms of eigenvectors alone but must be made in terms of principal vectors [11], $p(\alpha)$ which satisfy the matrix equation:

$$(P - \lambda_r I)^g p_r(\alpha) = 0, \tag{26}$$

where P is the matrix, g, an integer, is the grade of the principal vector, λ an eigenvalue, and I the identity matrix. If $g = 1$, $p_r(\alpha)$ is simply an eigenvector. For $g > 1$, this equation defines the principal vectors. An eigenvector $\psi(\alpha)$ can be derived from (26) as follows:

$$\psi(\alpha) = \frac{1}{(g-1)!}(P - \lambda I)^{g-1} p_r(\alpha). \tag{27}$$

It can be shown [12] that for large m the asymptotic form of

$$P^m p(\alpha) = m^{g-1}\lambda^m \psi(\alpha) + r^{(m)}, \tag{28}$$

where the remainder, $r^{(m)}$ is of order $m^{g-2}|\lambda|^m$. For the particular case of $g = 1$, $r^{(m)}$ is zero. Then we obtain the results used in (12) and (14). For the general case we must use the above asymptotic form to evaluate Eq. (13) and (16). This gives us

$$\langle \alpha' | P^m | \alpha \rangle \simeq \sum_{r,g} m^{g-1} \lambda_r^m \psi_{r,g}(\alpha') \psi_{r,g}(\alpha)$$

where $\psi_{r,g}(\alpha)$ is the eigenvector of eigenvalue, λ_r for the principal vector $p_r(\alpha)$ defined in (27). The conditions for persistent order then are that the maximum eigenvalues must be degenerate ($\lambda_1^M \approx \lambda_2^M$), *and* that their principal vectors must be of the same grade.

TIME-INDEPENDENCE OF MODEL PARAMETERS

We have assumed that V_{ij}, β, and V_0 are parameters which are fixed in time. It is reasonable to suppose that V_{ij}, in particular, might be influenced by learning. We might suppose that repeated firing of a given

synaptic junction might result in a permanent change in its physical and chemical properties and thus in the corresponding value of V_{ij}. Our model has neglected this, however, it appears as if one could extend without great difficulty the model to include such nonlinearities. The basis of this hope is that in humans at least, in order to learn something new so that it becomes part of long term memory one needs to concentrate for at least several seconds. We expect therefore that changes in V_{ij} take a time of this order to occur. On the other hand we have shown that this corresponds to something of the order of a thousand operations of the matrix. So we see that the changes in V_{ij} are likely to be small between each operation of the matrix operator and thus the nonlinear behavior might be approximated by the time-averaged quantities determined in the linear approximation.

8. CONCLUSION

We have argued that in a neural network the occurrence of states in which a correlation persists between neuronal configurations separated by long periods of time can occur if and only if the maximum eigenvalues of a certain transfer matrix are degenerate. By analogy with other systems which show long range order we show that a transition to such a persistent ordered state is analogous to a phase transition. We also show that in the ordered state a correlation occurs between neurons widely separated in the network.

If such persistent states can be identified in the brain, their presence must surely be of considerable significance, for their presence would dominate the average values of any quantity determined by the neuronal configuration just as the crystalline order dominates the average properties of a crystalline solid. It is of some interest to note too that, in general, the degeneracy of the eigenvalues of a matrix reflect some symmetry in the system which it represents, in fact, this is illustrated by a numerical example in the Appendix. This suggests that the capability of having persistent states in a neural network should be shown by some symmetry of the interneuronal network and the properties of the synaptic junctions. It would be interesting to know if this could be seen in the general anatomy of the brain.

APPENDIX

A better appreciation of some of the features of the model can be obtained by a numerical analysis of a simple network. We have considered just four neurons connected in various ways, have computed the 16×16 P-matrix for various values of β, V_{ij}, and V_0, and studied the asymptotic behavior of P^M for large M.

As a simple example we let $V_{ij} = 1.0$ for all i and j and considered powers of the matrix up to P^{32}. For $\beta = 0.2$ and $V_0 = 2.0$ we find that the matrix P^M, for large M, is such as to have identical columns, a characteristic feature of a matrix normalized as P is, with a nondegenerate maximum eigenvalue. On the other hand, for $\beta = 5.0$ and $V_0 = 2.0$ this feature is lost making it possible for P^M to transfer information of the structure of the initial state to the final state. This is the condition for the existence of the persistent state. For these values of V_{ij} and V_0 the transition to the persistent state appears to occur at about $\beta \simeq 1.0$. A sharp transition is not expected, however, because the number of neurons is so small.

For an arbitrary choice of the V_{ij}'s and V_0 we find that we do not always obtain a persistent state even at $\beta = 5.0$. This suggested that in order to obtain a degeneracy in the maximum eigenvalue the matrix P must have some special symmetry. Upon examination of (5) we notice that for the particular choice of conditions such that $\Sigma_j(V_{ij}/2) - V_0 = 0$, the matrix P would be invariant under the operation which changes the sign of all s_i' and all s_j. The choice of parameters of our first example satisfied this symmetry condition. On the other hand we find that for either $V_0 = 4.0$ or $V_0 = 0.0$, both of which violate this condition, no persistent state was found for $\beta = 5.0$. In the vicinity of $V_0 = 2.0$, however, the persistent state is found. The phase boundary in the β, V_0 plane can thus be located.

Having recognized this symmetry principle we examined a more complicated situation with both inhibitory and excitatory synapses described by the values of the V_{ij} given in Table 1. These values of V_{ij} were chosen again to satisfy the above symmetry principle. Again we found a persistent state for $\beta = 5.0$ and $V_0 = 2.0$ but with a structure different from that of our earlier choice of V_{ij}. For $V_0 = 4.0$ and 0.0 at $\beta = 5.0$ the nonpersistent state was found.

TABLE 1

	i			
j	1	2	3	4
1	−1.0	−1.0	+1.0	+1.0
2	−1.0	+1.0	+1.0	+1.0
3	+4.0	+2.0	+1.0	+1.0
4	+2.0	+2.0	+1.0	+1.0

These results illustrate some of the essential features of the model: first, the existence of the persistent state, second, the existence of a phase

boundary in the β, V_0 plane, and third, the existence of a symmetry principle for determining regions in which degenerate eigenvalues occur. One might expect the matrix to be invariant under certain other operations for other choices of V_{ij} and V_0. We expect that these would give rise to other types of persistent states.

REFERENCES

1 C. M. Smith, *The Brain*, G. P. Putmanns, New York (1970).
2 A. L. Hodgkin and A. F. Huxley, *Nature* **144,** 710 (1939).
3 L. I. Schiff, *Quantum Mechanics*, 3rd ed., McGraw-Hill, New York (1968).
4 H. A. Kramers and G. H. Wannier, *Phys. Rev.* **60,** 252 (1941).
5 G. F. Newell and E. W. Montroll, *Rev. Mod. Phys.* **25,** 353 (1953).
6 K. Huang, *Statistical Mechanics*, Wiley, New York (1963).
7 J. Ashkin and W. E. Lamb, Jr., *Phys. Rev.* **64,** 159 (1943).
8 E. N. Lassettre and J. P. Howe, *J. Chem. Phys.* **9,** 747, 801 (1941).
9 T. D. Lee and C. N. Yang, *Phys. Rev.* **87,** 410 (1952).
10 M. R. Moldover and W. A. Little, *Phys. Rev. Letters* **15,** 54 (1965).
11 J. N. Franklin, *Matrix Theory*, Prentice-Hall, Engelwood Cliffs, N.J. (1968).
12 Ibid., p. 275.