The Annals of Applied Probability
1991, Vol. 1, No. 2, 173–188

NEYMAN LECTURE

NEURAL NETS AND IMPLICIT INFERENCE

BY P. WHITTLE

University of Cambridge

The contention that artificial neural nets operate by a process of distributed hypothesis testing is supported by analysis of the antiphon, a model which is close to standard associative-memory models but is intended primarily to represent memory storage under stochastic disturbance of the system. The memory capacity of the antiphon under "neuronal" inference rules has been evaluated elsewhere; here it is evaluated under the supposition of efficient inference procedures.

1. Memory: Association and storage. It is a theme of this paper that neural networks operate by a process of continual hypothesis testing, distributed in time and space. Every time a signal is submitted to the nonlinear operation approximating a threshold response which (in some variant) is found so necessary in this area, such a test is being performed.

By "neural networks" is meant what are more explicitly termed "artificial neural networks" or [Whittle (1989)] "neuroidal networks," although the assertion above may well hold for the natural variety as well. The abbreviation ANN has become current, and seems useful.

We shall concentrate on a particular type of ANN: that designed to act as a memory. The version usually considered is that which acts as an *associative memory*, but we shall also consider a version which acts as what one might term a *memory store*. Let us define these terms.

An associative memory stores a number of "memory traces," and its function is to retrieve the correct trace when presented with a "cue," that is, with a partial or noise-corrupted version of the trace. Specifically, we suppose that there are M traces $a^{(j)}$, $j = 1, 2, \dots, M$, which we suppose to be binary N-vectors: $a^{(j)\prime} = (a_{j1}, a_{j2}, \dots, a_{jN})$, that is, column N-vectors whose components take the values 0 or 1. The cue presented is also an N-vector, $Y' = (y_1, y_2, \dots, y_N)$, which is assumed to be a distorted version of one of the traces. (However, the distortion may be such that Y is no longer binary; e.g., one may have a component value corresponding to "missing observation.") If the distortion is expressed statistically, then the problem is clearly one of statistical inference: that of deciding between the hypotheses $W = a^{(j)}$, $j = 1, 2, \dots, M$, on the basis of the data Y, where W is the "intended" trace behind the cue. If the M memory traces are equally likely to be presented,

Received November 1989; revised September 1990.

AMS 1980 *subject classifications*. 62H15, 93A15, 94A24

Key words and phrases. Neural nets, association, memory, Hopfield net, antiphon, channel coding.

then the inference procedure which minimises the probability of error is that of maximum likelihood: to decide for the value of j maximising $P(Y|W = a^{(j)})$. If the distortion is such that Y is binary and errors in different components of Y are independent with a constant error probability $\delta < \frac{1}{2}$, then the maximum likelihood criterion is equivalent to the minimisation of the Hamming distance $D(Y, a^{(j)})$; the number of places in which Y and $a^{(j)}$ differ.

In the ANN literature statistical hypotheses are often not very explicit, and the procedure adopted is rather to calculate the quantities

$$(1) \qquad \xi_j = f\Bigl(\sum_k a_{jk} y_k\Bigr), \qquad j = 1, 2, \dots, M.$$

Here f is a sigmoidal response function, frequently (and here) supposed to be of the simple threshold form

$$(2) \qquad f(x) = \begin{cases} 0, & x < d, \\ 1, & x \geq d, \end{cases}$$

for some threshold value d. We see here a form of statistical test: One tests for the validity of the hypothesis $W = a^{(j)}$ by testing whether the inner product of Y and $a^{(j)}$ exceeds the threshold d. The linear/threshold operation (1) has been regarded as the natural specification of an "artificial neuron" from the days of McCulloch and Pitts (1943); its function as an effective hypothesis test is apparent. It represents a simple fundamental operation which can be adapted to allow "learning" on the part of the network.

One would wish the threshold d to be chosen so that, with high probability, this threshold will be exceeded (and so $\xi_j = 1$) for and only for the "correct" value of j. This outcome is not certain, however, a point to which we return later. It is possible to implement neuronally a "winner-takes-all" rule which adjusts the threshold so that there is only a single threshold exceedance [Lippmann (1987) and Winters and Rose (1989)], although of course at some cost in additional circuitry.

We shall write relation (1) in the vector form as

$$(3) \qquad \xi = f(AY),$$

where A is the $M \times N$ matrix (a_{jk}), the understanding being then that the threshold operation is applied separately to each component of the vector AY.

Suppose that, as stated earlier, the aim of the network is to actually evoke (retrieve) the intended trace. This is the so-called *auto-associative* case. [For the more general *hetero-associative* case, in which the output is required to be a vector associated with the trace, see Baum, Moody and Wilczek (1988).] The network will then yield an output

$$(4) \qquad \sum_j \xi_j a^{(j)} = A' f(AY),$$

so that $A'f(AY)$ is the output of the network for input Y. A possibility often considered is that this output is fed back as a new input, so that several passes of the network are allowed for the output to home in on a memory trace. If

$Y(t)$, $\xi(t)$ refer to values of Y and ξ at the tth pass, with $Y(0) = Y$, we would then replace relations (3) and (4) by

$$\begin{aligned} \xi(t+1) &= f(AY(t)), \\ Y(t+1) &= A'\xi(t+1), \end{aligned} \qquad t = 0, 1, 2, \dots, \tag{5}$$

a pair of recurrences from which we can derive either of the single recurrences

$$\xi(t+1) = f(AA'\xi(t)), \tag{6}$$

$$Y(t+1) = A'f(AY(t)). \tag{7}$$

The celebrated Hopfield net [Hopfield (1982) and presaged in Kohonen (1977)] is also such a dynamic network and is also intended to act as an associative memory with traces $a^{(j)}$. It obeys the recursion

$$Y(t+1) = f(A'AY(t)) \tag{8}$$

(although sometimes the diagonal elements of the connection matrix $A'A$ are replaced by 0's).

In comparing (8) with relations (5)–(7), we notice two points. For one, as compared with (7), the threshold operation f seems to be misplaced. For another, relation (8), like (6) and (7), would appear to be better seen as the reduced form of a double recurrence analogous to (5).

This second point has considerable implications, both for conception and for physical realisation. If most of the elements of A are 0, then the network (5), with connection matrix A, has the desirable property of being sparse, whereas the network with connection matrix AA' or $A'A$ is in general fully connected. As far as conception goes, it is only by seeing the network in the nonreduced form (5) that one can appreciate the two phases: of an effective hypothesis test followed by production of the appropriate trace. The understanding of this two-phase operation is clear in the work of several authors [e.g., Baum, Moody and Wilczek (1988), whose exposition we have partly followed, Moopenn, Lambe and Thakoor (1987), Whittle (1989) and certainly others]. In the auto-associative case there is a particular economy, in that the same network is used for the two phases of the operation: once in the forward direction (with connection matrix A) and once backwards (implying the connection matrix A').

The notion of an associative memory thus centres on the assumption that the cue presented is subject to statistical disturbance. The idea of a memory store is prompted rather by the possibility of statistical disturbance in the operation of the network itself. The goal of achieving reliable memory under these conditions is a special case of the general problem of trying to achieve reliable operation from a system constructed of unreliable components. ("Unreliable," not in the sense that they may fail, but rather in that they follow stochastic rules and so are imperfectly predictable.) We demand of a memory store simply that it have a number of "memory states," and that, when set in one of these, it will hold that state reliably, that is, with prescribed probability over a prescribed time. If the system is stochastic its *physical* state will in fact drift; the memory states will correspond to domains of attraction in physical state space from which escape is difficult. So, memory states correspond to what might be variously termed metastable equilibria, quasistationary regimes or domains of attraction in state space. Reliability demands that these domains be sufficiently retentive; efficiency demands that there be as many as possible. These demands are of course conflicting.

In Whittle (1989) the author proposed a device for realising reliable memory from unreliable components: the *antiphon*. The antiphon is based upon the analogy of circulating a message, one of several possible, around a noisy communication loop, the noise representing the stochasticity (and so unreliability) of the system. It consists of a network which contains two sets of nodes: M α-nodes and N β-nodes. The jth α-node will be denoted α_j and the kth β-node β_k. At a given time an excitation pattern $\xi = (\xi_1, \xi_2, \dots, \xi_M)$ is defined on the α-nodes, the individual excitation (ξ_j at α_j) taking values 0 or 1. These patterns represent the possible memory values. The excitation pattern is transmitted to the β-nodes and then retrieved by the following rules.

If $\xi(t)$ is the pattern at time t, then β_k receives an input

$$u_k(t) = \sum_j \xi_j(t) a_{jk}. \tag{9}$$

Here a_{jk} indicates the capacity of the arc between α_k and β_k; we shall suppose that a_{jk} can adopt only the values 0 or 1 (corresponding to the simple absence or presence of an arc). We suppose that β_k then produces a scalar-valued output $y_k(t)$, these outputs being independent conditional on the inputs and conditioned only by the corresponding input. That is,

$$P(Y(t) = Y | U(t) = U) = \prod_k p(y_k | u_k), \tag{10}$$

where Y is the vector $(y_1, y_2, \dots, y_N)$ and so on, and $p(y|u)$ is a prescribed function of input and output. In the multiple-excitation case (when more than one ξ_j can be nonzero), the variables u_k will not be binary valued (i.e., restricted to the values 0 and 1). Correspondingly, there is no need to assume the y_k are binary valued, and distribution (10) may be such as to allow these variables to take values in a general space $\mathscr{Y}$.

Time proceeds in unit steps; we suppose that at the next instant of time α_j receives an input

$$x_j(t+1) = \sum_k a_{jk} y_k(t), \tag{11}$$

which is then normalised to an excitation variable by the threshold test

$$\xi_j(t+1) = f\big(x_j(t+1)\big). \tag{12}$$

One can regard the action as the alternation between the holding of a pattern in a local and reliable fashion (in the α-nodes) and in a distributed and unreliable fashion (in the β-nodes), the permanent holding of the pattern in the α-nodes being regarded as physically infeasible.

In the communication analogue the choice $\xi_j = 1$ indicates that the jth message value is to be transmitted. The vector $a^{(j)} = (a_{j1}, a_{j2}, \dots, a_{jN})$ represents the transmitted "word" W which encodes this message value. The vector Y represents the signal received at the other end of the channel, and the rule (10) represents the channel statistics. The operations (11) and (12) then amount to a statistical test of whether the signal $a^{(j)}$ has been transmitted or not. In fact, if the channel had been Gaussian, the M message values equiprobable and $\sum_k a_{jk}^2$ independent of j, then the efficient (and maximum likelihood) test for which message value had been intended would be to choose the value of j maximising x_j. The simple threshold rule (12) with appropriate threshold value d will achieve the same effect with high probability if the channel is not too noisy.

In the communication context one would send only a single message (i.e., only one of the ξ_j would be nonzero). It is a feature of the memory device that it is natural to allow the possibility of sending several superimposed messages at once. Indeed, it is only by this means that the system can be fully utilised; see Sections 4 and 5. Actually, it is not the compounding of messages which would be unconventional in the communication context (for this yields what can be regarded as a single compound message), but rather the representation of the compound message by the superposition of the corresponding words, which then constitutes a rather special form of coding.

It is not claimed that the antiphon represents any biological reality. Nevertheless, if one were to try to give a biological picture, one might interpret the β-nodes as sensory nodes and the α-nodes as centres which test the sensory information for the presence of particular memory traces. A memory is then held by allowing an abstract memory to evoke its sensory counterpart, this sensory evocation then to re-evoke the abstract memory, and so on. While still making no biological claim, we can give the mechanism more life by sometimes referring to the α- and β-nodes as "processing nodes" and "sensory nodes" respectively.

Despite its independent motivation from the communication context, the antiphon is nothing but a stochastic version of the associative memory considered earlier. Indeed, if the β-nodes are assumed reliable in that $Y(t) \equiv U(t)$, then relations (9), (11) and (12) amount exactly to relations (5). This agreement is welcome rather than disappointing in that it indicates a wider basis of support for the model which now turns out to be common. Much of the analysis we now give for the antiphon is also valid for the associative memory. The two models should be distinguished, however; the associative memory aims to optimise trace retrieval from an imperfect cue, the memory store aims to optimise trace holding despite imperfect internal operation. The random coding of communication theory is paralleled by the random choice of the connection matrix for the antiphon and the random choice of the memory traces for an associative memory. This coding aspect is made explicit in Baum, Moody and Wilczek (1988); Bruck and Blaum (1989) consider algebraic rather than random codings.

Note that there is an extensive literature on a stochastic (and so "unreliable") form of the Hopfield model: the "spin-glass" or "thermodynamic" version of this model [for extensive accounts see Amit (1989) and Kamp and Hasler (1990)]. This differs from the antiphon in at least two respects: The equivalent of the threshold operation is inserted at a different point, as we noted after formula (8), and the stochastic transitions are assumed to take a time-reversible form.

2. Reliability and capacity. Let us take N as a measure of the *size* of the network, i.e., the size of the input vector in the associative case and the number of nodes involved in the distributed representation in the storage case. One should be aware, however, that there are other size parameters which can affect the cost of realising the system, e.g., the number M of α-nodes for the antiphon, or the effective size of the structure introduced if one tries to realise more sophisticated calculations. The relation of these size parameters to N, and so the real cost of realisation, is something that has to be monitored.

The *probability of error* P_e for an associative memory is the probability that the cue is incorrectly recognised, averaged over random choices of the memory traces and equiprobable choice of which trace the cue represents. The corresponding probability for a memory store would be the probability that the system escapes after one time unit from the domain of attraction in which it was initially located. For the antiphon this is simply the probability that $\xi(t+1) \neq \xi(t)$.

Suppose that for either type of memory one considers a mode of construction and operation defined for varying N which is *reliable* in that the probability of error tends to 0 with increasing N. Let $R_A(N)$ be the number of traces which can be retrieved by this procedure (in the associative case) and $R_S(N)$ the number which can be stored (in the storage case).

Now, one would conjecture that there are procedures in the storage case for which $R_S(N)$ is exponentially large in N, because one expects that reliable storage of a positive number of bits of information per node should be possible. More explicitly, one expects that the limit

$$C[\pi] = \lim_{N\to\infty} \frac{R(N)}{N} \tag{13}$$

will exist and be positive, where $C[\pi]$ represents a *storage rate* in nats per node which is dependent upon the *policy* π, the procedure of construction and operation which is followed. (We shall in general take natural logarithms, and so measure information storage in nats rather than in bits.) One can then define a *capacity*

$$C = \sup_{\pi} C[\pi], \tag{14}$$

where the supremum is over reliable procedures. The capacity will of course depend upon the statistics assumed for the components of the system.

In Whittle (1990) the capacity is in fact evaluated for the antiphon under various restrictions on the mode of operation. However, the antiphon rules (11) and (12) represent a neuronal rule for estimating the value of $\xi(t)$ from that of $Y(t)$ which is in general statistically inefficient. In this paper we shall see how capacity evaluation changes if statistically efficient methods of inference are used.

In the associative case it is found that, at least for the Hopfield model, $R_A(N)$ can grow only linearly if reliability is demanded: The approximate estimate

$$R_A(N) \sim 0.15N$$

is supported by both simulation and asymptotic analysis [Hopfield (1982) and Amit, Gutfreund and Sompolinsky (1985)]. Part of the reason for this is that the $R_A(N)$ memory traces which are to be recognised are considered to be generated by random independent draws from a common distribution (of binary N-vectors). Performance can be greatly improved if this restriction is relaxed. The Hopfield net allows what are termed "spurious" or "parasitic" solutions in that, as well as recognising each of suitable small sets of randomly generated memory traces, it can also recognise some superpositions of these, that is, some vectors of the form $W = \sum_j \xi_j a^{(j)}$, where the coefficients ξ_j take values 0 or 1. If one includes such composites among the traces which are to be recognised, then one can indeed achieve an $R_A(N)$ which is exponentially large in N, and so a positive "association capacity." [That the Hopfield net allows exponentially many "solutions" if one includes the "spurious" solutions is established in, e.g., Bruce, Gardner and Wallace (1986), Gardner (1986) and Amit, Gutfreund and Sompolinsky (1987). However, the spurious solutions are conventionally regarded as meaningless and a nuisance in the associative-memory context. In the case of a memory store, we regard them as "memory states" which are as valid as any other.]

One can also achieve a positive association capacity, even with independently generated traces, by the use of nets which, compared with the Hopfield net, are large and sparse; see Section 3. The unwelcome feature here is that the cost of realisation can then also be exponentially large in N.

The capacities we calculate will be capacities averaged over a random choice of the antiphon structure, in which the a_{jk} are independently and identically distributed, taking values 1 or 0 with respective probabilities θ and $\phi = 1 - \theta$. This is completely analogous to the choice of a random coding in communication theory. Capacities for a given value of θ will be denoted by $C(\theta)$. More explicitly, we shall use $C_s^A(\theta)$ to denote capacity under the neuronal rules (11) and (12) and under s-fold excitation; that is, when one allows $\binom{M}{s}$ equiprobable memory traces, corresponding to exactly s of the ξ_j being nonzero.

We shall use $C_s^E(\theta)$ to denote the corresponding capacity if rules (11) and (12) are replaced by statistically efficient procedures for the estimation of $\xi(t)$ from $Y(t)$. The aim of this paper is to evaluate this capacity; for $s > 1$ this poses an interesting variant of standard situations in hypothesis testing and statistical communication theory.

The aspects of policy which remain at one's disposition are the choice of θ and s and the choice of inference rules (that is, either the simplest neuronal rules or an attempt at more efficient procedures). The aim under all circumstances is to find a procedure for which the capacity evaluation is genuine, in that the cost of realisation of the calculations is itself of no greater order than N.

3. Capacity under single excitation. Consider first the case of single excitation, when $\xi_j(t)$ is 0 for only a single value of j, so that there are just M possible memory values.

But this mode of operation, with efficient inference on ξ from Y, is nothing but the operation of a classical communication channel. In the communication analogue, there are just M possible message values, corresponding to the value of j for which ξ_j is nonzero. The jth message value is represented by the transmitted "word" $W = a^{(j)}$; the word sent out is to be inferred from the received signal Y. Assumption (10) implies that the channel is memoryless; let us rewrite this as

$$P(Y(t) = Y|U(t) = W) = \prod_k p(y_k|w_k), \tag{15}$$

where w_k is the kth letter of W. Moreover, the network statistics postulated imply a random coding in which the letters w_k of all codewords are independent, taking values 0 and 1 with probabilities ϕ and θ.

It follows then by standard theorems [see, e.g., Blahut (1987)] and also by the treatment of the more general s-excitation case in the next section that the capacity of the channel under this coding equals the mutual letter entropy

$$C_1^E(\theta) = i(w,y) = E\log\frac{P(y|w)}{P(y)}. \tag{16}$$

Here w and y are a pair of corresponding letters (w_k, y_k) with joint distribution specified by the random coding and (15). This evaluation is valid for any y-distribution for which the expectation (16) exists and is finite.

Let us use the notation

$$H[p(\cdot)] = -\sum_y p(y)\log p(y)$$

to denote the Shannon information measure for a discrete distribution $p(y)$. We shall denote this by $h(y)$ if we wish to emphasise the random variable to which it pertains rather than the distribution of which it is a function.

THEOREM 1. *The memory capacity of the antiphon under single excitation and efficient inference is*

$$C_1^E(\theta) = H[\phi p(\cdot|0) + \theta p(\cdot|1)] - \phi H[p(\cdot|0)] - \theta H[p(\cdot|1)]. \tag{17}$$

This follows immediately as the evaluation of expression (16), with $P(y|w) = p(y|w)$ and $P(y) = \phi p(y|0) + \theta p(y|1)$. The difference of information measures (17) has an obvious version even in the case when the distribution of y is not discrete.

Capacity evaluation in this case is thus very straightforward. However, we shall realise this capacity only by taking M of order $\exp(C_1 N)$ [which is shorthand for saying that we must take $\log M = (C_1^E(\theta) - \varepsilon)N + o(N)$ for arbitrarily small ε]. This is not acceptable if the processing nodes also have to be realised physically, and so contribute to the physical size and cost of the system. The problem is one that can only be surmounted by allowing multiple excitations, that is, by allowing several of the ξ_j to be nonzero. In the communication analogue this would amount to the coding of a compound message by superposition of the corresponding individual codewords, which is no longer the classical situation.

Neither does classical statistical communication theory provide an evaluation of capacity under the neuronal inference rules (11) and (12) of the strict antiphon. Quite different methods are adopted in Whittle (1990) to show that

$$(18)\quad C_1^A(\theta) = \sup_{z \geq 1} \sum_y \left[\theta y p(y|1) \log z - (\phi p(y|0) + \theta p(y|1)) \log(\phi + \theta z^y)\right].$$

This expression simplifies in the binary case, for which y can take only the values 0 and 1, say (corresponding to a β-node's not firing or firing). Let us in this case write $p(1|u)$, the probability of firing conditional on input u, as $p_u = 1 - q_u$. Then the maximising z in (18) is p_1/p_0 and we obtain

$$(19)\quad C_1^A(\theta) = -(\phi p_0 + \theta p_1) \log(\phi p_0 + \theta p_1) + \phi p_0 \log p_0 + \theta p_1 \log p_1.$$

Comparing this with expression (18) in the same case, we see that

$$(20)\quad \begin{aligned} C_1^E(\theta) = C_1^A(\theta) &- (\phi q_0 + \theta q_1) \log(\phi q_0 + \theta q_1) \\ &+ \phi q_0 \log q_0 + \theta q_1 \log q_1. \end{aligned}$$

The effect of coarsening the inference procedure to the antiphon rules is then to strip the q-terms from (20), leaving just the p-terms (19). It is somewhat as though one could derive information when a β-node fires, but not when it does not.

Expressions (17) and (18) equally provide evaluations of the capacity for an associative memory under the assumptions that the components a_{jk} of the M memory traces independently take values 1 or 0 with respective probabilities θ and ϕ and that the distribution of Y conditional on an intended trace W is given by (15). Expression (18) gives the capacity under the "neuronal" inference rule (3); expression (17) gives the capacity under efficient (maximum likelihood) inference rules.

Note that in all cases (association or storage, neuronal or efficient inference rules) a positive capacity is achieved, in that the number M of traces which can be reliably retrieved or stored is exponentially large in N. On the other hand, the capacity evaluations must also be said to be illusory, in that calculations are involved (by the mediation of the α-nodes or by importation of efficient procedures) which are themselves of a magnitude exponentially large in N. This is a difficulty which has to be surmounted.

In the binary symmetric case, when $\theta = \frac{1}{2}$ and $p_0 = q_1 = \psi$ (say), expression (17) reduces to

$$C_1^E = \log 2 + \psi \log \psi + (1 - \psi)\log(1 - \psi),$$

the familiar expression for the capacity of a binary symmetric channel. Chou (1989) has shown that this capacity can be realised for the associative case by the Kanerva memory; a network which involves only neuronal operations, although a large number of them [Kanerva (1988)]. Chou's calculations are involved, but one can see quite easily how this performance is achieved. The Kanerva memory essentially generates a large number of iid random binary N-vectors c and tests whether the proportion of these for which both $c'Y$ and $c'a^{(j)}$ exceed a threshold value Nd_1 itself exceeds a threshold value d_2, $j = 1, 2, \dots, M$. By this means one can effectively evaluate whether or not the probability $P(c'Y \geq Nd_1,\ c'a^{(j)} \geq Nd_1)$ (for random c and fixed Y and $a^{(j)}$) exceeds the threshold d_2 for each j. But if d_1 is taken in the range for which this probability can be evaluated by the central limit theorem for large N and if the binary vectors have elements ± 1 rather than 0 and 1, then one finds that, for large N, this probability depends upon Y and $a^{(j)}$ only in that it is a decreasing function of the Hamming distance $D(Y, a^{(j)})$. The Kanerva memory thus reveals itself as a technique for calculating these distances, which we know are exactly what is needed for efficient inference in the binary case.

The price of this improvement in performance is, as we might have guessed and as Chou proves, that one requires a network of size exponential in N.

4. Capacity under s-fold excitation. It is necessary to allow multiple excitation, that is, to allow more than one element of ξ to be nonzero, if the rate of increase of M with N is to be decreased. Suppose we allow exactly s nonzero values. In the communication analogue this corresponds to the idea that the transmitted signal representing a message value is a compound word

$$W = (a^{(j_1)}, a^{(j_2)}, \dots, a^{(j_s)}), \tag{21}$$

where $j_1, j_2, \dots, j_s$ is a selection of s distinct values from $(1, 2, \dots, M)$. There are thus $\binom{M}{s}$ possible message values (or memory patterns).

The conventional random coding theory then requires supplementation, because, instead of the $\binom{M}{s}$ values of W being generated by independent draws from a common distribution, it is the M individual words $a^{(j)}$ which are thus chosen, and compound words formed from them by the concatenation (21).

As we have noted, the idea of multiple excitation is one that has been largely rejected for associative models of memory [see, e.g., Amit (1989) and Kamp and Hasler (1990)], where the situation in which several memory traces are simultaneously evoked is regarded as a "spurious solution," biologically meaningless and of insufficient stability to be useful. At least for the antiphon model, however, multiply excited states can be stable, and so consistent with reliability.

Let us denote the collection of the first r codewords $(a^{(1)}, a^{(2)}, \dots, a^{(r)})$ by $a(r)$. For concreteness, let us suppose, unless otherwise stated, that it is the first s α-nodes which are excited, so that probabilities are to be calculated on the assumption that $W = a(s)$. Thus

$$P(Y|W) = \prod_k p\left(y_k | \sum_{j=1}^{s} a_{jk}\right). \tag{22}$$

The joint distribution of the y_k and the a_{jk} are thus specified by the random network assumptions and expression (22), also interpetable as $P(Y|a(s))$.

Define $\bar{a}(r) = (a^{(r+1)}, a^{(r+2)}, \dots, a^{(s)})$, the complement of $a(r)$ in $a(s)$. Define the random variable

$$\zeta_r = N^{-1} \log \frac{P(Y|a(s))}{P(Y|\bar{a}(r))}$$

whose expectation is the conditional mutual information rate

$$\gamma_r = E\zeta_r = N^{-1} i(a(r), Y|\bar{a}(r)).$$

Each of the random variables ζ_r will converge in probability to its expectation γ_r as N increases, because the random variables $\eta_k = (a_{1k}, a_{2k}, \dots, a_{sk}, y_k)$ are iid.

LEMMA 1. *The capacity under s-fold excitation and efficient inference has the lower bound*

$$C_s^E(\theta) \geq \min_{1 \leq r \leq s} (s\gamma_r/r). \tag{23}$$

PROOF. We continue to suppose that $W = a(s)$, that is, that it is the first s α-nodes which are excited. Then of the $\binom{M}{s}$ possible excitation patterns there are

$$m_r = \binom{s}{s-r}\binom{M-s}{r} = O(M^r), \tag{24}$$

which overlap this pattern by $s - r$, that is, which excite $s - r$ of the first s α-nodes and r of the last $M - s$. As a typical example of such a case, we can take $W = (b(r), \bar{a}(r))$, where $b(r)$ is a set of r codewords disjoint from $a(s)$. Let Q_r be the probability that such an excitation pattern is mistakenly accepted; this is not greater than the probability that

$$P(Y|W = (b(r), \bar{a}(r))) \geq P(Y|W = a(s)), \tag{25}$$

calculated on the assumption that $W = a(s)$. Then an upper bound for the probability of error is

$$P_e \leq \sum_{r=1}^{s} m_r Q_r.$$

A better bound is

$$(26)\qquad P_e \leq P(\bar{K}) + \sum_{r=1}^{s} m_r E(Q_r'),$$

where K is the event $\{\zeta_r \geq \gamma_r - \varepsilon_r;\ r = 1, 2, \dots, s\}$ and Q_r' is the probability, conditional on K and the values of Y and $a(s)$, that $b(r)$ adopts a value such that (25) holds.

Let us denote the set of $b(r)$ consistent with (25) for given Y and $a(s)$ by $B(Y, a(s))$, and let us denote probabilities P calculated on the assumption that $W = (b(r), \bar{a}(r))$ by P^b. Note then that $b(r)$ is independent of $\bar{a}(r)$ under any hypothesis, and that $P^b(Y|\bar{a}(r)) = P(Y|\bar{a}(r))$. For values of Y and $a(s)$ consistent with K and for $b(r)$ in $B(Y, a(s))$, we have then

$$P^b(Y|b(r), \bar{a}(r)) \geq P(Y|a(s)) \geq P(Y|\bar{a}(r))\exp[N(\gamma_r - \varepsilon_r)]$$

or

$$P^b(Y, b(r)|\bar{a}(r)) \geq P(b(r))P(Y|\bar{a}(r))\exp[N(\gamma_r - \varepsilon_r)]$$

or

$$P^b(b(r)|\bar{a}(r), Y) \geq P(b(r))\exp[N(\gamma_r - \varepsilon_r)].$$

Summing this last inequality over $b(r)$ in $B(Y, a(s))$, we deduce that

$$1 \geq Q_r' \exp[N(\gamma_r - \varepsilon_r)],$$

which, with (26), implies that

$$(27)\qquad P_e \leq P(\bar{K}) + \sum_{r=1}^{s} m_r \exp[N(\gamma_r - \varepsilon_r)].$$

Now, by the convergence of the ζ_r mentioned above, $P(\bar{K})$ will converge to 0 with increasing N for prescribed positive ε-values. Since $m_r = O(M^r)$ then expression (27) will tend to 0 with increasing N if $r \log M \leq N(\gamma_r - 2\varepsilon_r)$ for $r = 1, 2, \dots, s$, that is, if

$$N^{-1} \log M \leq \min_{1 \leq r \leq s} [(\gamma_r - 2\varepsilon_r)/r].$$

Since the ε_r are arbitrarily small and the memory size is $\log\binom{M}{s} \sim s \log M$, we see that a reliable rate arbitrarily close to the bound in (23) can be attained. □

We shall see that equality indeed holds in (23). One can then ask which value of r is minimising in this relation. If the minimising value of r were small, this would mean that the limiting factor on memory storage was the inability to distinguish the excitation pattern ξ from those differing from it in only a few places (that is, to distinguish the true hypothesis from those close to it). If the minimising value were near s, then this would mean that the limiting factor is the possibility of confusing ξ with one of the (many more) totally dissimilar patterns (i.e., of confusing the true hypothesis with one of the many totally dissimilar to it). In fact, the second is the case.

Lemma 2. *The quantity γ_r increases with r and γ_r/r decreases.*

Proof. We appeal to the facts that the random vectors $a^{(1)}, a^{(2)}, \ldots, a^{(s)}$ are iid when unconditioned and exchangeable when conditioned by Y. We have

$$N\gamma_r = E \log \frac{P(Y, a(s))}{P(a(s))P(Y|\bar{a}(r))} = \text{const.} + h(Y|\bar{a}(r)).$$

Since arguments are lost from $\bar{a}(r)$ as r increases, the first assertion follows. To establish the second, note that we can also write

$$N\gamma_r = E \log \frac{P(Y, a(s))P(\bar{a}(r))}{P(a(s))P(Y, \bar{a}(r))} = E \log \frac{P(a(s)|Y)}{P(a(r))P(\bar{a}(r)|Y)}.$$

Thus

$$N(\gamma_r - \gamma_{r-1}) = E \log \frac{P(\bar{a}(r-1)|Y)}{P(\bar{a}(r)|Y)P(a^{(r)})}$$

and

$$\begin{aligned} N(\gamma_{r+1} - 2\gamma_r + \gamma_{r-1}) &= E \log \frac{P(\bar{a}(r)|Y)^2}{P(\bar{a}(r-1)|Y)P(\bar{a}(r+1)|Y)} \\ &= E \log \frac{P(a^{(r)}|Y)^2}{P(a^{(r)}, a^{(r+1)}|\bar{a}(r+1), Y)} \\ &= E \log \frac{P(a^{(r)}|\bar{a}(r+1), Y)P(a^{(r+1)}|\bar{a}(r+1), Y)}{P(a^{(r)}, a^{(r+1)}|\bar{a}(r+1), Y)} \\ &= -i(a^{(r)}, a^{(r+1)}|\bar{a}(r+1), Y) \le 0. \end{aligned}$$

Thus γ_r is concave and so γ_r/r decreasing. □

Theorem 2. *The memory capacity of the antiphon under s-fold excitation and efficient inference is*

$$C_s^E(\theta) = H[Ep(\cdot|u)] - EH[p(\cdot|u)], \tag{28}$$

where the expectation is with respect to u, considered to have a binomial distribution with parameters s and θ.

That is, for a function $g(u)$,

$$Eg(u) = \sum_u \binom{s}{u} \phi^{s-u}\theta^u g(u).$$

Proof. By Lemmas 1 and 2 we have

$$C_s^E(\theta) \ge \gamma_s = N^{-1}i(a(s), Y) = i((a_{1k}, a_{2k}, \ldots, a_{sk}), y_k).$$

This expression is readily found to have the evaluation (28). In order to establish the reverse inequality, we again have to adapt standard arguments, although by much less. It follows by the extended Fano inequality [see, e.g., Blahut (1987)] that a bound on the probability of error for a fixed choice of codewords (21) is

$$P_e \geq 1 - \frac{1 + i(W, Y)}{\log\binom{M}{s}},$$

where $i(W, Y)$ is the mutual entropy between W and Y under the assumption that W takes each of its $\binom{M}{s}$ values (21) with equal probability. For the error probability averaged over codings we thus have

$$P_e \geq 1 - \frac{1 + E_a i(W, Y)}{\log\binom{M}{s}}$$

and so

$$C_s^E(\theta) \leq \limsup_{N \to \infty} N^{-1} E_a i(W, Y),$$

where E_a denotes an expectation over random codings a. Now, since the channel is memoryless

$$i(W, Y) \leq \sum_k i(w_k, y_k)$$

and so

$$N^{-1} E_a i(W, Y) \leq E_a i(w_k, y_k),$$

where w_k has the distribution in which it takes each of the $\binom{M}{s}$ values $(a_{j_1 k}, a_{j_2 k}, \dots, a_{j_s k})$ with equal probability. But with increasing N this distribution (itself a random variable under variation of the network a) converges in distribution to a distribution in which the elements of w_k are distributed independently, each taking values 0 or 1 with probabilities ϕ and θ. Since $i(w_k, y_k)$ is a bounded continuous function of the w_k-distribution, we then have $E_a i(w_k, y_k) \to \gamma_s$, so that $C_s^E(\theta) \leq \gamma_s$ and the proof of the theorem is complete. □

Expression (28) is exactly the capacity of the memoryless channel with input/output transition probability $p(y|u)$ under the random coding in which the letters w_k are independently binomially distributed with parameters s and θ; the expression is meaningful even if the distribution of y is not discrete. So, the conclusion is the simple but unobvious one: The constrained random coding in which one generates $\binom{M}{s}$ words by taking superpositions of s of M independent binary N-vectors realises the same information rate as that for which one takes the words as $\binom{M}{s}$ independent N-vectors with independent binomially distributed components.

5. Structural economies. The capacity $C_s = C_s^E(\theta)$ is a function of the two variables s and θ; these are the only parameters disposable for the maximisation of capacity. However, there is the consideration which we have already emphasised. The quantity M also reflects the size of the system, in that it represents the number of processing nodes. We should try and realise a given capacity with M as small as possible; that is, growing with N as slowly as possible.

Under s-fold excitation and full capacity working, we have $(s/N)\log M = C_s + o(1)$, so that, in this sense, $M \sim \exp[NC_s/s]$. That is, M grows exponentially fast with N, which is unacceptable if M and N are both to be regarded as size parameters. How C_s varies with s has yet to be determined. However, if C_s/s decreases with s then M grows more slowly with N as s is increased, although still at an exponential rate for any given s. One can then ask whether a slower rate of growth than exponential could be achieved by making s an increasing function $s(N)$ of N. An ideal would be if M could be made of order N, when the two size parameters would have been brought into correspondence.

Consider first the behaviour of $C_s^E(\theta)$ with increasing s (which means that we evaluate the limit as $N \to \infty$ for fixed s and then consider s-dependence of the limit). Suppose that the β-units *saturate* with increasing input, in that $p(y|u)$ has a limit $p(y|\infty)$ which is a distribution: $\sum_y p(y|\infty) = 1$. Then one readily verifies that, for fixed θ, expression (28) tends to 0 with increasing s. This simply reflects the fact that the input to the β-nodes becomes so high that their output is almost independent of the set of α-nodes which originated the input. To stabilise input to a fixed expected value λ with increasing s, one must make θ decrease as λ/s. One will then achieve a capacity, which we shall denote by $C_\infty^E(\lambda)$, given by expression (28) but with u following a Poisson distribution of parameter λ.

One may now ask whether the capacity $C_\infty = C_\infty^E(\lambda)$ can be realised by a mode of operation in which s increases with N at some appropriate rate, θ then decreasing as $\lambda/s(N)$. One would like $s(N)$ to increase as quickly as possible, so that $M(N)$ may increase as slowly as possible. Application of the methods of Section 4 suggests that one can attain the capacity C_∞ with s increasing as $N^{1/3}$ and so M increasing as $\exp(kN^{2/3})$ for some k. This is a far cry from the $M = O(N)$ behaviour we seek.

In fact, a positive capacity can be achieved with M of order N, although the methods of Section 4 are not strong enough to demonstrate this as they stand. The cruder neuronal inference rules are easier to analyse; it was demonstrated in Whittle (1990) that a positive reliable rate could be achieved under the constraint $M = O(N)$, although short of $C_\infty^A(\lambda)$. This positive rate was achieved with $s(N)$ of order $N/\log N$ and so the number $N\lambda/s(N)$ of β-nodes linked to a given α-node of order $\log N$.

One certainly has $C^E \geq C^A$ for any given mode of operation (that is, for any choice of dependence of s, θ and M upon N). The results quoted thus imply that a positive reliable rate can be achieved in the case of efficient estimation with M or order N. Specifically, if one chooses $M = LN$, $s = KN/\log N$ and $\theta = \lambda/s = \lambda \log N/KN$ for constant L and K, then there are positive values of K consistent with reliability, the supremum of such values being dependent on λ but independent of L. But K is itself the memory rate, since $(s/N)\log M = K + o(1)$, and so positive reliable rates exist.

Acknowledgment. This work was completed during the author's tenure of a Senior Fellowship awarded by the Science and Engineering Research Council, U.K.

REFERENCES

AMIT, D. J. (1989). *Modeling Brain Function*. Cambridge Univ. Press.

AMIT, D. J., GUTFREUND, H. and SOMPOLINSKY, H. (1985). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.* **55** 1530–1533.

AMIT, D. J., GUTFREUND, H. and SOMPOLINSKY, H. (1987). Statistical mechanics of neural networks near saturation. *Ann. Phys.* **173** 30–67.

BAUM, E. B., MOODY, J. and WILCZEK, F. (1988). Internal representations for associative memory. *Biol. Cybernet.* **59** 217–228.

BLAHUT, R. E. (1987). *Principles and Practice of Information Theory*. Addison-Wesley, Reading, Mass.

BRUCE, A. D., GARDNER, E. and WALLACE, D. J. (1986). Dynamical and statistical mechanics of the Hopfield model. Edinburgh preprint 387.

BRUCK, J. and BLAUM, M. (1989). Neural networks, error-correcting codes and polynomials over the binary n-cube. *IEEE Trans. Inform. Theory* **35** 976–987.

CHOU, P. A. (1989). The capacity of the Kanerva associative memory. *IEEE Trans. Inform. Theory* **35** 281–298.

GARDNER, E. (1986). Structure of metastable states in the Hopfield model. *J. Phys. A* **19** L1047–L1052.

HOPFIELD, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci. U.S.A.* **79** 2554–2558.

KAMP, Y and HASLER, M. (1990). *Associative Memory Neural Networks*. Wiley, Chichester.

KANERVA, P. (1988). *Sparse Distributed Memory*. MIT Press.

KOHONEN, T. (1977). *Associative Memory: A Systems-Theoretic Approach*. Springer, Berlin.

LIPPMANN, R. P. (1987). An introduction to computing with neural nets. *IEEE ASSP Mag.* April 4–22.

MCCULLOCH, W. S. and PITTS, W. (1943). A logical calculus of the ideas imminent in nervous activity. *Bull. Math. Biophys.* **5** 115–133.

MOOPENN, A., LAMBE, J. and THAKOOR, A. P. (1987). Electronic implementation of associative memory based on neural network models. *IEEE Trans. Systems Man Cybernet.* **17** 325–331.

NEWMAN, C. M. (1988). Memory capacity in neural network models: Rigorous lower bounds. *Neural Networks* **1** 223–238.

WHITTLE, P. (1989). The antiphon: A device for reliable memory from unreliable elements. *Proc. Roy. Soc. London Ser. A* **423** 201–218.

WHITTLE, P. (1990). The antiphon: The exact evaluation of memory capacity. II. *Proc. Roy. Soc. London A* **429** 45–60.

WINTERS, J. H. and ROSE, C. (1989). Minimum distance automata in parallel networks for optimum classification. *Neural Networks* **2** 127–132.

STATISTICAL LABORATORY
DEPARTMENT OF PURE MATHEMATICS
AND MATHEMATICAL STATISTICS
UNIVERSITY OF CAMBRIDGE
16 MILL LANE
CAMBRIDGE CB2 1SB
ENGLAND