Information Capacity of the Hopfield Model

YASER S. ABU-MOSTAFA AND JEANNINE-MARIE ST. JACQUES

Abstract—**The information capacity of general forms of memory is formalized. The number of bits of information that can be stored in the Hopfield model of associative memory is estimated. It is found that the asymptotic information capacity of a Hopfield network of N neurons is of the order N^3 b. The number of arbitrary state vectors that can be made stable in a Hopfield network of N neurons is proved to be bounded above by N.**

I. Introduction

IN CONTRAST to the standard model for memory, where the amount of information storage is an explicit quantity, the information capacity of certain models of associative memory is a debatable issue. Associative memory is a plausible model for biological memory, where a large number of simple connected building blocks (neurons) act individually in an apparently random way, yet collectively constitute an organ that does a specific complicated task in a robust manner. Apart from this biological interpretation, the ability to carry out collective computation in a distributed system of flexible structure without global synchronization has become a recognized engineering objective.

An important step in understanding collective systems is to quantify their ability to store information and carry out computation. The Hopfield neural network [2] is a model of associative content-addressable memory with a simple flexible structure. Being a content-addressable memory, it is capable of storing information, as well as carrying out certain computational tasks such as error correction and nearest neighbor search.

In this work, we introduce a definition of information capacity that is applicable to general forms of memory. We apply this definition to the Hopfield neural network and obtain tight upper and lower bounds for the number of bits that can be stored in a network of N neurons. We then restrict the format of information storage to stable states and obtain a linear upper bound for the number of vectors that can be made stable in the model, for every N. These results are equally valid in a completely different application that has the same mathematical formulation, namely the stable states of spin glasses [5], [8].

In Section II, we introduce the Hopfield model of associative memory and explain the function of the neural network. The concept of information capacity is formalized in Section III, and the definition is applied to get a tight asymptotic estimate for the information capacity of a network of N neurons. In Section IV, the linear upper bound for the number of stable states is derived; this constitutes a measure of the useful information capacity of the model. The Appendix discusses some background material about threshold functions.

Manuscript received August 27, 1984; revised December 19, 1984.

The authors are with the California Institute of Technology, Pasadena, CA 91125, USA.

II. The Hopfield Model

Complicated electronic circuits using neuron-like architectures can be made in an attempt to produce aspects of biological memory. However, these circuits are quite complex and highly ordered. It seems highly improbable that such mechanisms would arise naturally and be used as basic building blocks for biological memory. Instead, if a large number of neurons had computationally useful collective properties, arising simply due to their number, chance would favor the use of the building block that is the simplest and the least ordered. Hopfield [2] has shown that a large number of highly stylized neurons do have collective properties. He has found that a set of asynchronously operating nonlinear neurons can store information with stability and efficiency, recall it with some error-correcting capability, and exhibit a sense of time order. Also, his model is quite robust and should work even when more neurological details are added.

A neural network consists of N pairwise connected neurons. The ith neuron can be in one of two states: $u_i = -1$ (off) or $u_i = +1$ (on). The (synaptic) connections are undirected and have strengths that are fixed real numbers. Define the state vector $\boldsymbol{u}$ to be a binary vector (± 1) whose ith component corresponds to the state of the ith neuron. Randomly and asynchronously, each neuron examines its inputs and decides whether to turn itself on or off. It does this in the following manner. Let w_{ij} be the strength (which may be negative) of the synaptic connection from neuron j to neuron i. ($w_{ij} = w_{ji}$ and $w_{ii} = 0$). Let t_i be the threshold voltage of the ith neuron. If the weighted sum over all of its inputs is greater than or equal to t_i, the ith neuron turns on and its state becomes $+1$. If the sum is less than t_i, the neuron turns off and its state becomes -1. The action of each neuron simulates a general threshold function (see the Appendix) of $N-1$ variables (the states of all the other neurons):

$$u_i = \operatorname{sgn}\left(\sum_{j=1}^{N} w_{ij} u_j - t_i \right).$$

Let W be an $N \times N$ real-valued, zero-diagonal symmetric matrix. The entries of W are the w_{ij} defined above; w_{ij}

0018-9448/85/0700-0461$01.00

is the strength of the synaptic connection from neuron j to neuron i. Let the threshold vector $\boldsymbol{t}$ be a real-valued vector whose ith component is the threshold voltage of the ith neuron. Each choice of W and $\boldsymbol{t}$ defines a specific neural network of N neurons with specific values for the strengths of the synaptic connections and the threshold voltages of the neurons. The network starts in an initial state and runs with each neuron randomly and independently reevaluating itself. Often, the network enters a stable point in the state space in which all neurons remain in their current state after evaluating their inputs. This stable vector of states constitutes a stored word in the memory, and the basic operation of the network is to converge to a stable state if we initialize it with a nearby state vector (in the Hamming sense).

Hopfield [2] proposed a specific scheme of constructing the matrix W that makes a given set of vectors $\boldsymbol{u}^1, \cdots, \boldsymbol{u}^K$ stable states of the neural network. The scheme is based on the sum of the outer products of these vectors. We shall make no assumptions here about how the matrix W is constructed in terms of the vectors $\boldsymbol{u}^1, \cdots, \boldsymbol{u}^K$, and all the results are valid even if Hopfield's particular construction scheme is not followed.

III. Information Capacity

A Hopfield network represents a memory that stores information, and it is appropriate to ask how much information we can store in a network of N neurons. To define the information capacity C, we start with a familiar example and try to extend it.

If we have a random access memory with M address lines and one data line (an $M \times 1$ RAM, consisting of 2^M memory locations, where each location is accessed by an M-bit address and contains one bit of stored data), it is clear that we can store 2^M b of information. This is because given an *arbitrary* string of 2^M b, we can load the $M \times 1$ RAM with the string and be able to retrieve the whole string from the memory later on.

There is also another way to look at it, if we consider the string as a single object. We can store and retrieve any string (of length 2^M b) in the $M \times 1$ RAM, and there are 2^{2^M} such strings. Thus the memory can distinguish between 2^{2^M} cases. *We define the information capacity of a memory to be the logarithm of the number of cases it can distinguish between*, in this case $C = \log 2^{2^M} = 2^M$ b.

How does this definition apply to the Hopfield model? Consider a neural network with N neurons. The w_{ij} and the t_i are what distinguish one network from the other. If we had access to these values and were able to read them, the information capacity of the memory would be infinite, since a real number constitutes an infinite amount of information. However, we can only sense these values through the state transitions of the neurons. The question now becomes, how many different sets of values for w_{ij} and t_i can we distinguish between merely by observing the state transition scheme of the neurons? This corresponds to the number of distinguishable networks of N neurons. If this number is c, the capacity of the network will be $C = \log c$ b.

The key factor in estimating the number of distinguishable networks is the known estimate for the number of threshold functions (see the Appendix). The action of each neuron simulates a general threshold function of $N-1$ variables (the states of all the other neurons). There are at most $2^{(N-1)^2}$ such functions [3]. Since there are N neurons, there will be at most $(2^{(N-1)^2})^N$ distinguishable networks. The logarithm of this number is an upper bound for the information capacity C. Hence

$$C \le \log(2^{(N-1)^2})^N = O(N^3) \text{ b}.$$

Let us consider the lower bound now. There are at least $2^{\alpha n^2}$ threshold functions of n variables, where $\alpha \approx 0.33$ [6]. The symmetry of the matrix W makes the N threshold functions dependent, but we can take the submatrix of W consisting of the first $\lfloor N/2 \rfloor$ rows and the last $\lfloor N/2 \rfloor$ columns, and consider the partial threshold functions defined by this submatrix. Since the entries of this submatrix are independent, we have at least $\lfloor N/2 \rfloor$ functions each of $n = \lfloor N/2 \rfloor$ variables. Therefore, the number of distinguishable networks is a least $(2^{\alpha \lfloor N/2 \rfloor^2})^{\lfloor N/2 \rfloor}$. The logarithm of this number is a lower bound for the information capacity C. Hence

$$C \ge \log\left(2^{\alpha \lfloor N/2 \rfloor^2}\right)^{\lfloor N/2 \rfloor} = \Omega(N^3) \text{ b}.$$

The conclusion is that the information capacity C of a Hopfield neural network with N neurons is exactly of the order N^3 b. This definition of information capacity is quite general, and it is interesting to investigate how it is affected by imposing certain restrictions on the format of information storage. This aspect is addressed in the next section, where the storage format is restricted to stable states.

IV. Stable States

Information in the Hopfield model is stored as stable states. A stable state $\boldsymbol{u}^s$ is a state that is a fixed point of the neural network. Each of the N neurons randomly and repeatedly looks at the weighted sum of all its inputs and then decides not to change from its previous state. To see how information is stored in the model, look at the example of pattern recognition and error correction.

A person sees a face X and wants to decide if the face is that of person A or that of person B. The visual picture of the face is processed and the description is encoded into a binary vector $\boldsymbol{u}^X$, which contains the information describing the face. $\boldsymbol{u}^X$ is then fed into the particular neural network that remembers the faces of persons A and B. That is, $\boldsymbol{u}^A$ and $\boldsymbol{u}^B$, which contain the information describing faces A and B, respectively, are stable states of this particular network. The vector $\boldsymbol{u}^X$ is fed into the network by setting the initial state of the ith neuron to the same value as the ith component of the binary vector $\boldsymbol{u}^X$.

After a period of time, the state of the network is evaluated. If $\boldsymbol{u}^X$ is close to $\boldsymbol{u}^A$, then $\boldsymbol{u}^A$ will be the

network's final state. The face is then recognized as belonging to person A and similarly if $\boldsymbol{u}^X$ is close to $\boldsymbol{u}^B$. If $\boldsymbol{u}^X$ is in between $\boldsymbol{u}^A$ and $\boldsymbol{u}^B$, the system will randomly converge to one or the other of the two states. Therefore, we have a model that makes decisions and has some error-correcting capability.

It is of interest to know the number of memories that can be stored in a Hopfield network of N neurons. What is the maximum number K such that any K vectors of N binary entries can be made stable in a network of N neurons by the proper choice of W and $\boldsymbol{t}$? Since we have to come up with a network for every choice of the K vectors, and since there are $\binom{2^N}{K}$ such choices, but less than 2^{N^3} such networks, it follows that

$$\binom{2^N}{K} \leq 2^{N^3}.$$

Restricting K to be at most 2^{N-1} because of the symmetry of the choice function, we get $K = O(N^2)$. To be able to store and retrieve the order of N^2 arbitrary stable states in a Hopfield network with N neurons seems quite ambitious. Hopfield predicted experimentally that $K \approx 0.15N$ [2], and McEliece showed a statistical bound of $K \leq N/2 \log N$ [4]. However, these estimates restrict the construction of W to the sum-of-outer-products scheme [2]. We now improve on the $O(N^2)$ bound and show that the number of stable states K can be at most N, for every N, no matter how the matrix W is constructed.

Theorem: Let W denote a real-valued zero-diagonal $N \times N$ matrix, and let $\boldsymbol{t}$ denote a real-valued N vector. Suppose that $K \leq 2^{N-1}$ is an integer satisfying the following condition.

For any K-set of binary N-vectors $\boldsymbol{u}^1, \cdots, \boldsymbol{u}^K$, there is a matrix W and a vector $\boldsymbol{t}$ such that

$$\operatorname{sgn}\left(\sum_{j=1}^{N} w_{ij} u_j^k - t_i \right) = u_i^k, \qquad \text{for } k = 1, \cdots, K \text{ and } i = 1, \cdots, N,$$

then $K \leq N$.

Proof: Suppose that K satisfies this property. We construct K vector $\boldsymbol{u}^1, \boldsymbol{u}^2, \cdots, \boldsymbol{u}^K$ as follows. The first entries in these vectors, namely $u_1^1, u_1^2, \cdots, u_1^K$, are binary variables $x^1, x^2, \cdots, x^K$ to be fixed later. The remaining $N-1$ entries in each vector are fixed ± 1's such that no two vectors have exactly the same entries (always possible since $K \leq 2^{N-1}$). We apply the condition of the theorem for $i = 1$. For any choice of $x^1, \cdots, x^K$, there must be real numbers $w_{12}, w_{13}, \cdots, w_{1N}, t_1$ such that

$$\operatorname{sgn}\left(\sum_{j=2}^{N} w_{1j} u_j^k - t_1 \right) = x^k$$

for $k = 1, \cdots, K$, since $w_{11} = 0$ (zero-diagonal). Therefore, for each of the 2^K choices for the values of $x^1, \cdots, x^K$, we must find a different threshold function of $N-1$ variables with K points in the domain. Let B_{N-1}^K be the number of the threshold functions of $N-1$ variables with K points in the domain. We must have

$$B_{N-1}^K \geq 2^K. \tag{1}$$

Cameron [1] and Winder [9] (see the Appendix), give the following upper bound to B_{N-1}^K:

$$B_{N-1}^K \leq 2 \sum_{i=0}^{N-1} \binom{K-1}{i}.$$

If $K > N$, then

$$B_{N-1}^K \leq 2 \sum_{i=0}^{N-1} \binom{K-1}{i} < 2 \sum_{i=0}^{K-1} \binom{K-1}{i} = 2 \times 2^{K-1} = 2^K.$$

So if $K > N$, then $B_{N-1}^K < 2^k$, which contradicts condition (1). Therefore K must be at most N and the proof is complete.

The theorem is a formalization of the fact that a Hopfield neural network cannot have more than N arbitrary stable states. Notice that the matrix W was not required to be symmetric, and this covers the generalization of the Hopfield model where the synaptic connections become directed (allowing $w_{ij} \neq w_{ji}$). Also, there is no restriction on the method of constructing W and $\boldsymbol{t}$ in terms of $\boldsymbol{u}^1, \cdots, \boldsymbol{u}^K$. McEliece and Posner [5] predicted that a zero-diagonal symmetric matrix has an exponential number of stable states on the average. The above theorem predicts at most a linear number of arbitrary stable states for a zero-diagonal matrix. The two results imply that the average number of parasitic stable states is exponential in N.

V. Conclusion

The information capacity of general forms of memory was formalized and applied to the Hopfield model of associative memory. Exact asymptotic estimates for the number of bits that can be stored in a neural network of N neurons were derived. A linear upper bound for the number of arbitrary stable states that can be stored in a neural network of N neurons was proved. This bound is reasonably close to the experimentally achievable capacity and to the statistically predicted capacity.

Appendix
Enumeration of Threshold Functions

A switching function $f(x_1, \cdots, x_n)$ of n binary variables $x_1, \cdots, x_n$ is defined by assigning either 0 or 1 to each of the 2^n points $(x_1, \cdots, x_n)$ in $\{0,1\}^N$. We are using a binary $(-1, +1)$ convention, which is strictly equivalent to the $(0,1)$ convention. A switching function $f(x_1, \cdots, x_n)$ of n variables is linearly separable if there exists a hyperplane π in the n-dimensional space, which strictly separates the "on" set $f^{-1}(1)$ from the "off" set $f^{-1}(-1)$. In other words, $f^{-1}(1)$ lies on one side of π, and $f^{-1}(-1)$ lies on the other, and $\pi \cap \{-1, +1\}^N$ is empty. Linearly separable switching functions are also called threshold functions [3]. A threshold function simulates a neuron examining its inputs and making its decision as to its next state. Cameron [1] and Winder [9] give the following upper bound on the number of

threshold functions of n variables defined on m points B_n^m:

$$B_n^m \le 2\sum_{i=0}^{n}\binom{m-1}{i}.$$

They arrive at their upper bound in the following manner. Define an $(n+1)$-dimensional space in which the coordinate axes correspond to the weights and to the threshold voltage. Consider a particular state $\boldsymbol{u}$. Plot $\boldsymbol{u}$ as a hyperplane in $n+1$ space, the set of all values of w_j and t such that

$$\sum_{j=1}^{n} w_j u_j - t = 0.$$

Note that the hyperplane passes through the origin and that it divides the space into two regions. Weights and threshold voltages from one of the regions make $\sum_{j=1}^{n} w_j u_j - t > 0$ and correspond to the threshold function on $\boldsymbol{u}$ being equal to 1. Weights and voltages from the other region make $\sum_{j=1}^{n} w_j u_j - t < 0$ and correspond to the threshold function on $\boldsymbol{u}$ being equal to -1. Each of the m points gives a similar hyperplane.

Thus we have m hyperplanes passing through the origin in $n+1$ space and partitioning the space into a number of regions. Each region corresponds to a threshold function. All points in any one of these regions correspond to values of w_j and t that produce the same threshold function. Two points in different regions correspond to two different functions as at least one $\boldsymbol{u}$ out of the m $\boldsymbol{u}$'s is mapped to $+1$ by one function and mapped to -1 by the other. Therefore B_n^m is less than or equal to the maximum number of regions (call the number C_{n+1}^m) made by m hyperplanes passing through the origin in $n+1$ space. Assume $m-1$ hyperplanes have made C_{n+1}^{m-1} regions in $n+1$ space. We add the mth hyperplane to make as many more regions as possible. The mth plane can intersect the other $m-1$ hyperplanes in at most $m-1$ hyperlines. The $m-1$ hyperlines can at most partition the mth plane into C_n^{m-1} hyperplane regions, since this is the same problem in n space. Since each region in the mth plane has been divided into a boundary between two regions in $n+1$ space, we have added C_n^{m-1} regions to the other C_{n+1}^{m-1} regions given by $m-1$ planes.

Therefore

$$C_{n+1}^m = C_n^{m-1} + C_{n+1}^{m-1}.$$

The solution of this recurrence relation is $C_{n+1}^m = 2\sum_{i=0}^{n}\binom{m-1}{i}$, which is an upper bound for B_n^m. If $m = 2^n$, i.e., the threshold function is defined for every binary n-vector, then we have an upper bound for the number of fully defined threshold functions of n variables (for $n \ge 4$):

$$B_n^{2^n} \le 2\sum_{i=0}^{n}\binom{2^n-1}{i} \le 2(n+1)\times\binom{2^n-1}{n}$$

$$\le 2(n+1)\frac{2^{n^2}}{n!} \le 2^{n^2}.$$

References

[1] S. H. Cameron, "An estimate of the complexity requisite in a universal decision network," Bionics Symposium, Wright Airforce Dev. Div. (WADD) Rep. 60-600, pp. 197–212, 1960.

[2] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," in *Proc. Nat. Academy Sci., USA*, vol. 79, 1982, pp. 2554–2558.

[3] P. M. Lewis and C. L. Coates, *Threshold Logic*. New York: Wiley, 1967.

[4] R. J. McEliece, private correspondence, 1984.

[5] R. J. McEliece and E. C. Posner, "The number of stable points of an infinite-range spin glass," unpublished manuscript, 1984.

[6] S. Muroga, "Generation of self-dual threshold functions and lower bounds of the number of threshold functions and a maximum weight," in *Proc. AIEE Symp. Switching Circuit Theory and Logical Design*, 1962, pp. 170–184.

[7] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[8] F. Tanaka and S. E. Edwards, "Analytic theory of the ground state properties of a spin glass: I. Ising spin glass," *J. Phys. F: Metal Phys.*, vol. 10, pp. 2769–2778, 1980.

[9] R. O. Winder, "Threshold logic," Ph.D. dissertation, Princeton Univ., Princeton, NJ, 1962.

[10] ——, "Bounds on threshold gate realizability," *IRE Trans. Electron. Comput.*, vol. EC-12, pp. 561–564, 1963.