http://dx.doi.org/10.1109/IJCNN.1992.227183

deSilva, C.J.S. and Attikiouzel, Y. (1992) Hopfield networks as discrete dynamical systems. In: International Joint Conference on Neural Networks, 1992. IJCNN, 7 - 11 June, Baltimore, Maryland, USA, pp. 115-120.

Hopfield Networks as Discrete Dynamical Systems

Christopher J. S. deSilva and Yianni Attikiouzel
Centre for Intelligent Information Processing Systems
Department of Electrical and Electronic Engineering
The University of Western Australia
Nedlands W.A. 6009, AUSTRALIA

Abstract

The analysis of Hopfield networks has been carried out principally by the methods of statistical mechanics and information theory. In this paper, a more geometrical approach based on dynamical systems theory is presented. Basic definitions from dynamical systems theory are presented and illustrated with a simple example. The eigenspaces of the weight matrix, the geometry of the energy manifold and the diffeomorphisms induced on the sphere are presented as tools for the study of network dynamics. Topics for further research are proposed.

1 Introduction

Since their introduction in [1], the analysis of Hopfield networks has been carried out principally by the methods of statistical mechanics (e.g. [2], [3], [4], [5]) and information theory (e.g. [6], [7], [8]). In this paper, a more geometrical approach based on dynamical systems theory is presented. This is in keeping with Hopfield's original approach, which used dynamical systems concepts in the description of the behaviour of the network. The approach is similar to that of Cottrell [9].

In this paper the eigenspaces of the weight matrix, the geometry of the energy manifold and the diffeomorphisms induced by the weight matrix on the sphere are presented as tools for the study of the network dynamics. Some simple results are given. A number of topics for further research are proposed.

2 Definitions

The following definition is slightly non-standard. Definitions of other mathematical terms used in the paper and background material on linear algebra and dynamical systems may be found in [10].

Definition 2.1 *Let S be a set. A* discrete dynamical system *on S is a function $U : S \to S$. S is the* state space *of the dynamical system.*

Generally, the state space is taken to be a topological space or smooth manifold, and the function U is assumed to be continous or smooth. For Hopfield networks, the state space will be a finite set and conditions of continuity or smoothness will not apply.

Definition 2.2 *The* orbit *of $x \in S$ under U is the set $\{x, Ux, U^2(x), \ldots, U^n(x), \ldots\}$.*

The points of S whose orbits exhibit persistent steady-state behaviour are the basis for the analysis of dynamical systems.

Definition 2.3 *$x \in S$ is a* fixed point *of U if $U(x) = x$. $x \in S$ is a* periodic point *of U if there exists a positive integer p such that $U^p(x) = x$. The least such p is the* period *of X.*

Fixed points are considered to be special cases of periodic points (with period 1).

In what follows, $\mathcal{R}$ will denote the real numbers and $\mathcal{R}^n$ the n-dimensional vector space consisting of n-tuples of real numbers.

3 A Simple Example

The definitions of the previous section will be illustrated by a simple example.

Consider a Hopfield network consisting of five units. We will assume that the states of the units are either $+1$ or -1, and that the thresholds of the units are all 0. If we assign some ordering to the units, the state space of the network becomes $\{1,-1\}^5$. This is a subset of $\mathcal{R}^5$, so the states of the network can be represented as vectors in that space.

We will find it convenient to represent the state of the network also as a string of plus and minus signs: $[++--+]$ will denote the network state in which the first two units are in the $+1$ state, the next two are in the -1 state and the last unit is in the $+1$ state.

Suppose that we wish the network to store two patterns: $[++-++]$ and $[+-+-+]$. The weight matrix is then

$$W = \begin{pmatrix} 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & -2 & 2 & 0 \\ 0 & -2 & 0 & -2 & 0 \\ 0 & 2 & -2 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The state space $\mathcal{S}$ of the network is $\{1,-1\}^5$, which we consider as a subset of $\mathcal{R}^5$. The weight matrix W defines a linear transformation on $\mathcal{R}^5$, which will map points of S to points of $\mathcal{R}^5$ that will generally not belong to $\mathcal{S}$. This linear transformation will also be denoted by W.

We will define a projection mapping $\pi : \mathcal{R}^5 \rightarrow \mathcal{S}$ by requiring that π will map $\mathbf{x} = (x^1, x^2, x^3, x^4, x^5)^T \in \mathcal{R}^5$ to $(\sigma(x^1), \sigma(x^2), \sigma(x^3), \sigma(x^4), \sigma(x^5))^T \in \mathcal{S}$, where σ denotes the function that maps $x \in \mathcal{R}$ to 1 if $x > 0$ and to -1 if $x < 0$.

The dynamical system U on $\mathcal{S}$ will be the composition $\pi \circ W$. If $\mathbf{x}$ denotes the state of the network at some time, $U(\mathbf{x})$ will be the state of the network after all the units have been updated synchronously, provided that none of the components of $W(\mathbf{x})$ is 0. This will happen in a number of cases in this particular example; we will assume that the definition of the projection π is modified in such a way that if some component of $W(\mathbf{x})$ is 0, then the value of the corresponding component of $U(\mathbf{x})$ is the same as that of the same component of $\mathbf{x}$. The dynamics of the network can now be studied by considering the orbits of the states under U.

The state space consists of 32 states. Of these, there are four fixed points:

$$[++-++], [+-+-+], [--+--], \text{ and } [-+-+-].$$

The first two are the stored patterns from which the weight matrix was computed; the other two are obtained from the first two by reversing the state of each unit. We will call these states F_1, F_2, R_1, R_2 respectively.

There are also four states of period 2, making up two cycles:

$$[++-+-] \mapsto [--+-+] \mapsto [++-+-]$$

and

$$[+-+--] \mapsto [-+-++] \mapsto [+-+--].$$

We will call these cycles C_1 and C_2 respectively.

There are three other states which map to F_1 under U:

$$[+++++], [++--+], [+--++].$$

Similarly, there are three states which map to F_2 under U:

$$[+++-+], [+-+++], [+---+].$$

It will be seen that these are the only states from which the network can reach one of the stored patterns.

Three states map to each of the other fixed points:

$$[-+++-], [-+---] \text{ and } [---+-] \text{ map to } R_1,$$

while

$$[-++--],[--++-] \text{ and } [-----] \text{ map to } R_2.$$

The remaining states all map to one of the cycles:

$$[++++-],[++---],[+--+-],[-++++],[-+--+],[---++]$$

map to C_1, while

$$[+++--],[+-++-],[+----],[-++-+],[--+++],[----+]$$

map to C_2.

These types of behaviour were all described by Hopfield [1].

4 Eigenspaces of the Weight Matrix

Most of the properties of the Hopfield network that we shall derive are consequences of two simple facts: first, the states of the units are either $+1$ or -1, and, second, the connections in the network are symmetric and no unit is connected to itself. The second fact determines the properties of the weight matrix and consequently of the linear transformation it defines.

We will consider a Hopfield network with N units, so that the state space is $\{1,-1\}^N$. This is a subset of $\mathcal{R}^N$, which we will call the *embedding space*. We will denote the state space by $\mathcal{S}$.

We will suppose that P patterns are to be stored in the network. We will denote these patterns by $\mathbf{p}_1,\ldots,\mathbf{p}_P$, where $\mathbf{p}_i = (p_i^1,\ldots,p_i^N)^T, i = 1,\ldots,P$, and each $p_i^n \in \{1,-1\}$. The $\mathbf{p}_i$ are points in $\mathcal{R}^N$. They span a linear subspace of $\mathcal{R}^N$ whose dimension is no more than P. We will denote the set of patterns to be stored by $\mathcal{P}$. The subspace spanned by $\mathcal{P}$ will be denoted by $\mathcal{E}$.

The components of the weight matrix W are given by $w_{ij} = \sum_{k=1}^P p_k^i p_k^j$ for $i \neq j$ and $w_{ii} = 0, i,j = 1,\ldots,N$.

W is an $N \times N$ matrix, and it defines the linear transformation $W : \mathcal{R}^N \rightarrow \mathcal{R}^N$. Since $w_{ij} = w_{ji}$ for all i and j and $w_{ii} = 0$ for all i, W is a symmetric real matrix with zero trace. As a consequence, all its eigenvalues are real and their sum is zero. This means that, if W is not the zero matrix, then it must have both positive and negative eigenvalues. We can consider W to be a (linear) discrete dynamical system on $\mathcal{R}^N$.

Proposition 4.1 *For all* $\mathbf{x} \in \mathcal{R}^N, W(\mathbf{x}) = \sum_{k=1}^P < \mathbf{p}_k,\mathbf{x} > \mathbf{p}_k - P\mathbf{x}$.

Proof. Let $\mathbf{x} = (x^1,\ldots,x^N)^T$, and let $\mathbf{y} = (y^1,\ldots,y^N)^T = W(\mathbf{x})$. Then for $i = 1,\ldots,N$,

$$y^i = \sum_{j=1}^N w_{ij}x^j = \sum_{j=1,j\neq i}^N \sum_{k=1}^P p_k^i p_k^j x^j,$$

since $w_{ii} = 0$. Then

$$y^i = \sum_{k=1}^P \sum_{j=1,j\neq i}^N p_k^i p_k^j x^j = \sum_{k=1}^P \sum_{j=1}^N p_k^i p_k^j x^j - \sum_{k=1}^P (p_k^i)^2 x^i = \sum_{k=1}^P p_k^i \sum_{j=1}^N p_k^j x^j - Px^i$$

since $(p_k^i)^2 = 1$ for all i and k. Thus

$$y^i = \sum_{k=1}^P p_k^i < \mathbf{p}_k,\mathbf{x} > -Px^i.$$

Hence $W(\mathbf{x}) = \sum_{k=1}^P < \mathbf{p}_k,\mathbf{x} > \mathbf{p}_k - P\mathbf{x}$. □

$\mathcal{R}^N$ is endowed with the standard inner product, denoted by $<,>$. If S is any subset of $\mathcal{R}^N$, its *orthogonal complement*, denoted $S^\perp$, is the set $\{x \in \mathcal{R}^N :< x,s >= 0 \text{ for all } s \in S\}$.

The orthogonal complement of S is a linear subspace of $\mathcal{R}^N$. $\mathcal{P}^\perp = \mathcal{E}^\perp$ because $\mathcal{E}$ is spanned by $\mathcal{P}$.

If $\mathbf{x} \in \mathcal{R}^N$ is a non-zero vector with the property that $W(\mathbf{x}) = \lambda\mathbf{x}$ for some $\lambda \in \mathcal{R}$, then $\mathbf{x}$ is an *eigenvector* of W with *eigenvalue* λ.

The following results are immediate consequences of the proposition above.

Proposition 4.2 *If $\mathbf{x} \in \mathcal{E}^{\perp}$, then $\mathbf{x}$ is an eigenvector of W with eigenvalue $-P$.*

Proof. By the previous proposition, $W(\mathbf{x}) = \sum_{k=1}^{P} < \mathbf{p}_k, \mathbf{x} > \mathbf{p}_k - P\mathbf{x}$. Since $\mathbf{p}_k \in \mathcal{E}$ for all k, and $x \in \mathcal{E}^{\perp}$, $< \mathbf{p}_k, \mathbf{x} >= 0$ for all k, so $W(\mathbf{x}) = -P\mathbf{x}$, and hence $\mathbf{x}$ is an eigenvector of W with eigenvalue $-P$. □

Corollary 4.2.1 *If $\mathbf{x} \in \mathcal{P} \cap \mathcal{E}^{\perp}$, then $\mathbf{x}$ is a periodic point with period 2.*

Proof. Since $\mathbf{x} \in \mathcal{E}^{\perp}, W(\mathbf{x}) = -P\mathbf{x}$. Since $\mathbf{x} \in \mathcal{P}$, all the components of $\mathbf{x}$ are non-zero, so $\pi(-P\mathbf{x}) = -\mathbf{x}$. So $U(\mathbf{x}) = -\mathbf{x}$. $-\mathbf{x} \in \mathcal{P} \cap \mathcal{E}^{\perp}$, so $U(-\mathbf{x}) = \mathbf{x}$. □

The foregoing results are true regardless of the particular set of stored patterns, provided $\mathcal{P}$ does not span the entire state space. The following results require additional conditions to be placed on the set of stored patterns.

Proposition 4.3 *In the case where the pattern vectors to be stored are pairwise orthogonal, that is, $< \mathbf{p}_i, \mathbf{p}_j >= 0$ for $i \neq j$, and $\mathbf{x} \in \mathcal{E}$, then $\mathbf{x}$ is an eigenvector of W with eigenvalue $(N - P)$.*

Proof. As before, $W(\mathbf{x}) = \sum_{k=1}^{P} < \mathbf{p}_k, \mathbf{x} > \mathbf{p}_k - P\mathbf{x}$. The $\mathbf{p}_k$ form an orthogonal basis for $\mathcal{E}$, so

$$\mathbf{x} = \sum_{k=1}^{P} \frac{< \mathbf{p}_k, \mathbf{x} >}{< \mathbf{p}_k, \mathbf{p}_k >} \mathbf{p}_k.$$

$< \mathbf{p}_k, \mathbf{p}_k >= N$ for all k, so

$$N\mathbf{x} = \sum_{k=1}^{P} < \mathbf{p}_k, \mathbf{x} > \mathbf{p}_k.$$

So $W(\mathbf{x}) = (N - P)\mathbf{x}$, and $\mathbf{x}$ is an eigenvector of W with eigenvalue $(N - P)$. □

Corollary 4.3.1 *If the pattern vectors to be stored are pairwise orthogonal, they are fixed points of U. The pattern vectors obtained by reversing the states of the units in the stored patterns are also fixed points of U.*

Proof. If the pattern vectors to be stored are pairwise orthogonal, $N - P > 0$. It follows that $U(\mathbf{p}_k) = \mathbf{p}_k$ for all k. Reversing the states of $\mathbf{p}_k$ gives $-\mathbf{p}_k$, which is also an eigenvector with eigenvalue $N - P$. □

The following question is a topic for further research: *How are the stored patterns related to the eigenvalues and eigenvectors of W ?*

5 The Geometry of the Energy Manifold

Hopfield [1] based his analysis of the network on the energy function defined by

$$E(\mathbf{x}) = -\mathbf{x}^T W \mathbf{x}, \mathbf{x} \in \mathcal{R}^N.$$

The fact that W is a symmetric, zero-trace matrix means that the energy function has a simple form. It should be noted that the fact that E is not a positive definite quadratic form means that is it is not an energy function in the physical sense; in particular, states of negative energy are possible. In fact, as we shall see, the stored patterns should be states of negative energy.

Let us consider the graph of the energy function, which is an N-dimensional submanifold of $\mathcal{R}^{N+1}$. This manifold is often referred to as the *energy surface* in the literature, even though its dimension is greater than 2. We shall adopt the more correct term *energy manifold.*

The fact that W is symmetric with zero trace means that the energy manifold is a hyperboloid, and that it has a single critical point at the origin, which is a generalized saddle.

Further information about the energy manifold can be obtained from the derivative of E, given by $DE(\mathbf{x})(\mathbf{y}) = -\mathbf{x}^T W \mathbf{y} - \mathbf{y}^T W \mathbf{x} for \mathbf{x}, \mathbf{y} \in \mathcal{R}^N$.

Suppose that $\mathbf{x} \in \mathcal{E}^{\perp}$, so $W(\mathbf{x}) = -P\mathbf{x}$, then $DE(\mathbf{x})(\mathbf{y}) = P(\mathbf{x}^T\mathbf{y} + \mathbf{y}^T\mathbf{x})$, and in particular, $DE(\mathbf{x})(\mathbf{x}) = 2P\mathbf{x}^T\mathbf{x}$. This means that the variation in the energy function along any ray from the origin in $\mathcal{E}^{\perp}$ is a increasing quadratic function of the distance from the origin. The same is true for a ray in the direction of any other eigenvector with a negative eigenvalue.

If $\mathbf{x}$ is an eigenvector with eigenvalue $\lambda > 0$, then $DE(\mathbf{x})(\mathbf{x}) = -\lambda\mathbf{x}^T\mathbf{x}$, so that the energy function decreases as a quadratic function of distance from the origin along a ray from the origin in the direction $\mathbf{x}$.

To study the behaviour of the energy function on the pattern space $\mathcal{S}$, it is convenient to start by considering its behaviour on the sphere of radius $\sqrt{N}$, since all the points of $\mathcal{S}$ lie on this sphere. This sphere will be called the *embedding sphere*; as it has dimension $N-1$, it will be denoted by S^{N-1}.

Extreme values of the energy function restricted to S^{N-1} can be found by using the method of Lagrange Multipliers. To find the extreme values of $E(\mathbf{x}) = -\mathbf{x}^T W\mathbf{x}$ subject to the constraint $C(\mathbf{x}) = \mathbf{x}^T\mathbf{x} = N$, we introduce the Lagrange Multiplier μ, and solve the equations $DE(\mathbf{x})(\mathbf{y}) + \mu DC(\mathbf{x})(\mathbf{y}) = 0$ and $C(\mathbf{x}) = N$ for $\mathbf{x}$ and μ. (The first equation must hold for all values of $\mathbf{y}$.)

We have

$$DE(\mathbf{x})(\mathbf{y}) + \mu DC(\mathbf{x})(y) = -\mathbf{x}^T W\mathbf{y} - \mathbf{y}^T W\mathbf{x} + \mu\mathbf{x}^T\mathbf{y} + \mu\mathbf{y}^T\mathbf{x}.$$

By symmetry of W,

$$DE(\mathbf{x})(\mathbf{y}) - \mu DC(\mathbf{x})(y) = -2\mathbf{y}^T W\mathbf{x} + 2\mu\mathbf{y}^T\mathbf{x}.$$

So we must have $\mathbf{y}^T W\mathbf{x} - \mu\mathbf{y}^T\mathbf{x} = 0$ for all $\mathbf{y}$. This is only possible if $W\mathbf{x} = \mu\mathbf{x}$; that is, if $\mathbf{x}$ is an eigenvector of W with eigenvalue μ.

The following question is a topic for further research: *How do the values of E on S^{N-1} relate to values of E on $\mathcal{S}$?* In particular, when are the points of $\mathcal{S}$ maxima or minima of E on S^{N-1} ?

6 Induced Diffeomorphisms of the Embedding Sphere

So far, we have dealt with two discrete dynamical systems: the linear system $W : \mathcal{R}^N \to \mathcal{R}^N$ and the composition $U = \pi \circ W$ on $\mathcal{S}$. There is also a dynamical system on the embedding sphere which can be used to study U. This map, which we will denote by $V : S^{N-1} \to S^{N-1}$, is obtained by considering points of the sphere to be points of R^N, finding their images under W and projecting them back to S^{N-1}. We have the following diagram:

$$\begin{tikzcd}
\mathcal{R}^N \arrow[r, "W"] & \mathcal{R}^N \arrow[d] \\
S^{N-1} \arrow[u] \arrow[r, "V"] & S^{N-1} \arrow[d] \\
\mathcal{S} \arrow[u] \arrow[r, "U"] & \mathcal{S}
\end{tikzcd}$$

In the diagram, the maps on the left side are inclusions and the maps on the right side are projections. V is the composition $\pi_S \circ W \circ \iota_S$, where ι_S is the inclusion of the embedding sphere in $\mathcal{R}^N$ and $\pi_S : \mathcal{R}^N \to \mathcal{S}^{N-1}$ is the projection given by $\pi_S(\mathbf{x}) = \sqrt{N}\mathbf{x}/||\mathbf{x}||$, where $||\mathbf{x}||^2 =< \mathbf{x}, \mathbf{x} >$.

If $\mathbf{x} \in \mathcal{S}$ is a fixed point of U, the components of $V(\mathbf{x})$ must have the same signs as the components of $\mathbf{x}$. If $\mathbf{x} \in \mathcal{S}$ is a fixed point of V, then it is also a fixed point of U. Further, if $\mathbf{x} \in \mathcal{S}$, and each component of $V(\mathbf{x})$ differs from the corresponding component of $\mathbf{x}$ by less than 1, then $\mathbf{x}$ is a fixed point of U. It is more convenient to work with V than with W, because the projection π_S factors out the rescaling effects of W.

The following questions are topics for further research: *Given W, what are the fixed points and periodic points of V ? Which points of S^{N-1} are attracted to the various fixed and periodic points ? What is the relationship between fixed and periodic points of U and fixed and periodic points of V ?*

7 Directions for Further Research

This paper has attempted to set out the beginnings of a geometric approach to the dynamics of Hopfield networks based on dynamical systems theory. The approach appears to have considerable potential as a tool for the study of these networks, as well as other networks with similar characteristics.

Some questions that appear to merit further research have been set out in the preceeding sections. There are also a number of broader research issues which may be considered.

First, this approach relates to the synchronous method for updating the network. Can it be adapted and applied to the asynchronous updating method ?

Second, the proofs presented above depend on the units of the network having two states and zero threshold. It is likely that modifying the proofs for the case of non-zero unit thresholds will be a simple matter. However, it would appear that extending this approach to the case of units which can take any of a range of values would require different techniques.

Third, for the mathematician, the induced diffeomorphisms on the embedding sphere appear to form an interesting class of dynamical systems in their own right. A classification of such systems and their behaviour should be a fairly straightforward matter.

8 Acknowledgement

The author gratefully acknowledges the financial support provided by the Engineering Foundation of the University of W.A.

9 References

[1] Hopfield, J.J. "Neural networks and physical systems with emergent collective computational abilities", *Proceedings of the National Academy of Sciences,* **79**, pp. 2554–2558, 1982.

[2] Amit, J., Gutfruend, H., and Sompolinsky, H. "Spin-glass models of neural networks", *Physical Review A,* **32**, pp. 1008–1018, 1985.

[3] Peretto, P., and Niez, J. "Stochastic dynamics of neural networks", *IEEE Transactions on Systems, Man, and Cybernetics,* **SMC-16**, pp. 73–83, 1986.

[4] Amit, J., Gutfruend, H., and Sompolinsky, H. "Statistical mechanics of neural networks near saturation", *Annals of Physics,* **173**, pp. 30–67, 1987.

[5] Amari, S.-I., and Maginu, K. "Statistical neurodynamics of associative memory", *Neural Networks,* **1**, pp. 63–74, 1988.

[6] McEliece, R., Posner, E., Rodemich, E., and Venkatesh, S. "The capacity of the Hopfield associative memory", *IEEE Transactions on Information Theory,* **IT-33**, pp. 461–482, 1987.

[7] Sussman, H.J. "On the Number of Memories that can be Perfectly Stored in a Neural Net with Hebb Weights", *IEEE Transactions on Information Theory,* **IT-35**, pp. 174–178, 1989.

[8] Piret, P. "Analysis of a Modified Hebbian Rule", *IEEE Transactions on Information Theory,* **IT-36**, pp. 1391–1397, 1990.

[9] Cottrell, M. "Stability and attractivity in associative memory neural networks", *Biological Cybernetics,* **58**, pp. 239–250, 1988.

[10] Hirsch, M.W., and Smale, S. *Differential Equations, Dynamical Systems, and Linear Algebra,* Academic Press, New York, 1974.