Pergamon

Neural Networks, Vol. 9, No. 5, pp. 747–763, 1996
Copyright © 1996 Elsevier Science Ltd. All rights reserved
Printed in Great Britain
0893-6080/96 $15.00 + .00

0893-6080(95)00077-1

CONTRIBUTED ARTICLE

Self-similar Neural Networks Based on a Kohonen Learning Rule

SIMON CLIPPINGDALE AND ROLAND WILSON

University of Warwick

(*Received* 24 *October* 1994; *Accepted* 2 *May* 1995)

Abstract—*One of the most striking features about the perceptual machinery of mammals is its regularity of structure. This is particularly evident in the mammalian visual system, as the work pioneered by Hubel and Wiesel has demonstrated. The likely source of this regularity is the visual stimulus, which does not change randomly from instant to instant, but is affected primarily by motions of both the animal and objects in the environment. These motions induce structured changes in the visual stimulus, which might well be expected to have a significant effect in shaping the structure of the visual machinery, whether through individual plasticity or longer-term genetic changes.*

The work reported in this paper is an investigation of the structures that may evolve in a simple artificial neural network driven not by random changes of input pattern, but directly by transformations which are themselves related to transformations of the input signal through an analysis of motion-prediction error. Results are presented which demonstrate that such networks can evolve a remarkable degree of regularity which reflects the underlying symmetry group of the transformations, both in one and two dimensions. An appropriate and visually plausible choice of transformation group can lead to the development of foveal structures in two-dimensional networks. We also present some preliminary results on parametrised function spaces which support the general conclusion that global structure bearing a considerable resemblance to that found in the mammalian visual system can evolve as the result of a simple learning rule in networks driven by transformations similar to those typically encountered in vision.

Copyright © 1996 Elsevier Science Ltd

Keywords—Transformations in vision, Self-similarity, Visual representation, Retinal and foveal development, Signal prediction.

1. INTRODUCTION: LEARNING SYMMETRIES

The recognition that symmetry is an important issue in the modelling of perception is not a new one—it dates back at least to the work of Pitts and McCulloch (1947) on the cat. The study of geometric invariants also played a significant part in Minsky and Papert's demonstration of the weaknesses of the single-layer perceptron (Minsky & Papert, 1969). Its importance in perceptual psychology has also been argued by many authors, notably Shepard (1981). Perhaps its most obvious biological expression is the regularity of the neural machinery found at all levels of the visual system which have been systematically investigated hitherto, from the array of photoreceptors in the retina to the *hypercolumnar* organisation of the striate cortex, first uncovered by Hubel and his co-workers (Hubel & Wiesel, 1962; Hubel, 1988). There have been attempts to link these physiological findings with the requirements of invariant pattern recognition, but these can hardly be described as conclusive (Cavanagh, 1978; Schwartz, 1980). Work on neural networks for invariant PR such as the neocognitron (Fukushima et al., 1983), or those based on moments or similar techniques, may well provide practicable solutions to the problem, but they are generally based on computational methods which directly exploit the underlying symmetry (e.g., translation invariance) rather than acquiring it through learning. Previous work with direct relevance to the modelling of biological vision has tended to concentrate on the *microstructure*: the development of artificial neurons having receptive field

Acknowledgements: This work was supported by the UK SERC/EPSRC. The authors would also like to thank Ian Stewart, Peter Mason, Takayuki Ito, Tetsuro Kuge and Toshio Nakagawa for valuable discussions, and Wilf Kendall for help with the convergence proof.

Requests for reprints should be sent to Roland Wilson, Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK; e-mail: simon@dcs.warwick.ac.uk, rgw@dcs.warwick.ac.uk

profiles (rfp's) similar to those found in biological systems (Barrow, 1987; Kammen & Yuille, 1988; Linsker, 1988). Important though such studies are, they do not address the question of *macrostructure*: why is the visual system so regular in its structure? There are at least two plausible answers to the question:

1. It may be that the only efficient way to organise a massive parallel computation, such as that confronting the visual system (particularly at low levels) is with a regular structure.
2. It may be that the structure has evolved in direct response to the structure of the visual stimulus.

Of course, the two are not mutually exclusive—if it were too "expensive" for the system to follow the dictates of the stimulus, then no doubt it would adopt a cheaper compromise solution. It seems more effective to start by testing the assumption that the second assertion is true, by exploring artificial neural systems driven by appropriately structured inputs. This is the approach adopted in the present work. The structuring of the input is entirely due to the use of transformations corresponding to motions, such as rotation, dilation and translation. The simplest example of a network driven by transformations is a form of Kohonen net (Kohonen, 1984), in which each neuron corresponds to a point $\mathbf{x}_i \in \mathscr{R}^m$, and the input stimulus is derived by applying a randomly chosen co-ordinate transformation to the network itself, giving a new set of points $\{\mathbf{y}_i\} = \{\mathbf{T}\mathbf{x}_i\}$. The learning rule is designed to make the randomly transformed network $\{\mathbf{y}_i\}$ as similar as possible on average to the original network $\{\mathbf{x}_i\}$; hence the title *self-similar neural network*. We have found that these simple networks are capable of acquiring a highly symmetric structure, which reflects the group of symmetries from which the transformations are selected. We have also applied the idea to networks whose units correspond to parametrised Gaussian functions of randomly selected initial positions and scales; a simple modification of the learning rule leads to a *self-similar function network*, which also self-organises in response to randomly chosen co-ordinate transformations. Some preliminary results are presented on these networks. The striking feature about these networks is that it is possible under some plausible assumptions about the nature of the transformations acting in vision to evolve synthetic arrays which bear a considerable resemblance to the retinal structure found in animals. While not conclusive, this demonstrates that it may indeed be possible for global visual structures to evolve primarily in response to the "demands of the stimulus", rather than being organised solely on computational criteria.

After a description of the generic network and learning rule, we present results and analysis (including a proof of convergence in the Appendix) for 1-D periodic networks. We then describe a set of simulations illustrating the behaviour of networks in 2-D under increasingly large subgroups of the affine group. We examine the issue of self-similarity in relation to a minimum error criterion in motion prediction, and lastly we show some results of experiments using Gaussian functions, to show that the basic principle of the network is readily extended to spaces of functions. This has implications in terms of the rfp's of cells at various levels in the visual system. The paper is concluded with a discussion of our findings in relation to the problem of symmetry in vision.

2. NETWORK STRUCTURE AND LEARNING RULE

The networks used in this work consist of a single "layer" of N units, each of which is characterised by an m-dimensional *network vector* $\mathbf{x}_i \in \mathscr{R}^m$, $0 \leqslant i \leqslant N-1$. In this respect, they are similar to the units of the Kohonen self-organising feature map (SOFM) or the learning vector quantiser (LVQ) (Kohonen, 1984). Time dependence is indicated by a parenthesised argument: at time n (i.e., following the nth iteration of the training procedure), the ith network vector is denoted $\mathbf{x}_i(n)$.

At the nth iteration, a set of N *input vectors* $\mathbf{y}_i(n) \in \mathscr{R}^m$, $0 \leqslant i \leqslant N-1$, is presented to the network. These input vectors induce adjustments to the network vectors $\mathbf{x}_i(n-1)$, $0 \leqslant i \leqslant N-1$, according to the following learning rule:

$$\mathbf{x}_j(n) = \mathbf{x}_j(n-1) + \alpha(n) \sum_{i \in \Lambda_j(n)} \left(\mathbf{y}_i(n) - \mathbf{x}_j(n-1)\right) \qquad (1)$$

where

$$\Lambda_j(n) = \{i : \|\mathbf{y}_i(n) - \mathbf{x}_j(n-1)\| < \|\mathbf{y}_i(n) - \mathbf{x}_k(n-1)\|,\ k \neq j\} \qquad (2)$$

indexes all those input vectors $\mathbf{y}_i(n)$ to which $\mathbf{x}_j(n-1)$ is the closest network vector; $\alpha(n)$, $0 < \alpha(n) < 1$ is an adjustable learning parameter. Note that on any given iteration, a particular network vector $\mathbf{x}_j(n-1)$ may be the closest network vector to zero or more input vectors, i.e., the set $\Lambda_j(n)$ may have zero or more members. Hence that network vector $\mathbf{x}_j(n-1)$ may undergo no update, or it may be moved toward the mean of a number of input vectors. The various cases are illustrated in Figure 1, which shows an input set of four 2-D vectors and its effect on a four-unit network for a value of the learning parameter $\alpha(n)$ of around 0.5.

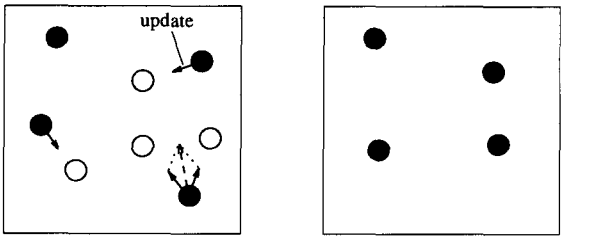


FIGURE 1. Network vectors $x_j(n-1)$ (solid circles) and updates to them induced by input vectors $y_i(n)$ (open circles). The network vector near the upper left corner is not affected by any input vector, whereas the network vector near the lower right corner is affected additively by two input vectors.

The learning rule resembles that of the Kohonen SOFM, but our networks do not possess its local connective topology: units are updated individually, without affecting other units in the network. This is equivalent to a neighbourhood or "bubble" size in the SOFM of zero, or to the LVQ with only attractive, and no repulsive, interactions. A further difference is that we use N input vectors per iteration rather than one, because we wish to consider inputs which are themselves transformed versions of the network. Indeed, a number of different possible sources of the input vector set $\{\mathbf{y}_i(n)\}$ come to mind. It might, for instance, be:

(A) a different sample at each iteration, drawn from some density (cf. Kohonen, 1984);
(B) a single sample (prototype array) of randomly chosen vectors, transformed differently at each iteration;
(C) the current network vector set, suitably transformed.

These methods generate input vector sets which differ in the degree of structure which they possess and in their relationship to the network; B and C give rise to input sets which are primarily specified in terms of transformations, whereas A does not; C gives rise to recursion—the input set is directly related to the existing structure of the network—whereas A and B do not. This work deals principally with one- and two-dimensional networks of type C.

3. ONE-DIMENSIONAL PERIODIC NETWORKS

In networks of type C, the input vector set at each iteration is a randomly transformed version of the current network vector set,

$$\mathbf{y}_i(n) = \mathbf{T}(n)\mathbf{x}_i(n-1), \tag{3}$$

with $\mathbf{T}(n)$ drawn from a density on some symmetry group of transformations. The simplest such network

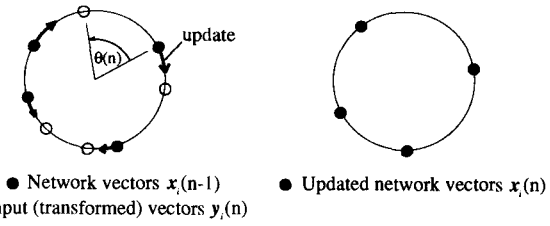


FIGURE 2. 1-D example of a network driven directly by random rotations. The left figure shows an input vector set $\{y_i(n)\}$ derived from the network vector set $\{x_i(n-1)\}$ by applying the randomly-chosen rotation $\theta(n)$. We find the nearest network vector to each input vector and move the former toward the latter, resulting in the updates shown as arrows, in this case for a value of learning parameter α of about 0.5. The updated network is shown on the right. For this particular value of the random rotation $\theta(n)$, each network vector is affected by exactly one input vector.

is a one-dimensional periodic network (the units correspond to points on a circle) where the input set is a randomly rotated version of the current network set. We examine the behaviour of these simple networks in some detail, since they show interesting and provable convergence properties which shed light on the behaviour of their less analytically tractable 2-D counterparts.

The steps involved in the nth iteration are as follows, and are depicted in Figure 2 for a four-unit system:

(i) Select a random rotation $\theta(n)$ uniform on $[0, 2\pi)$.
(ii) Apply this rotation to each network vector, now characterised by a single angle $x_i(n-1)$, to give an input vector $y_i(n)$:

$$y_i(n) = x_i(n-1) + \theta(n). \tag{4}$$

(iii) Update the network vector set $\{x_i(n-1)\}$ according to the learning rule of (1) and (2) to give the new network $\{x_i(n)\}$.

Figure 3 shows a real network of this type, with 32 units. Both the rotations $\theta(n)$ and the initial

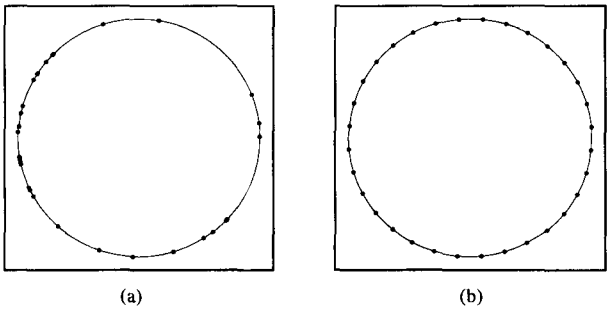


FIGURE 3. 32-unit network on the circle, driven by random rotations. (a) initial configuration; (b) after 5000 iterations with learning parameter $\alpha(n) = \alpha = 0.1$. The network converges to exactly regular spacing within a few thousand iterations.

configuration shown in Figure 3a are drawn from a uniform density on $[0, 2\pi)$. The configuration after 5000 iterations, with the learning parameter $\alpha(n)$ set to a constant value of $\alpha(n) = \alpha = 0.1$, is shown in Figure 3b and is exactly equally spaced, up to numerical precision.

3.1. Existence of a Fixed Point

It is easy to see that once regular spacing is established, it will persist. If the network point set $\{x_i(n-1)\}$ is regularly spaced, then the input set $\{y_i(n)\}$ is also regularly spaced. Exactly one network vector $x_j(n-1)$ is the closest to each input vector $y_i(n)$, and all such x, y pairs are separated by the same angle. The updates to the network vectors are all identical, and the regularity persists. Such regular configurations therefore constitute a fixed point of the system up to an arbitrary rotation, or in other words a stable *orbit* or *attractor*. The following theorem restates this result.

THEOREM 1. *Let* $\{x_j, 0 \leqslant j \leqslant N-1\}$ *be a set of points on the circle, with input set given by* (4) *and updates performed according to* (1) *and* (2) *above (restated here in scalar form), i.e., at iteration n,*

$$x_j(n) = x_j(n-1) + \alpha(n) \sum_{i \in \Lambda_j(n)} \big(y_i(n) - x_j(n-1)\big), \quad (5)$$

where

$$\Lambda_j(n) = \{i : |y_i(n) - x_j(n-1)| < |y_i(n) - x_k(n-1)|,\ k \neq j\} \quad (6)$$

and

$$y_i(n) = x_i(n-1) + \theta(n). \quad (7)$$

Then the point sets $\{x_j = 2\pi j/N + \psi, 0 \leqslant j \leqslant N-1\}$, *where* ψ *is arbitrary, form a stable orbit for the system* (5)–(7).

Proof. Let the point set at iteration $n-1$ be uniformly spaced,

$$x_j(n-1) = 2\pi j/N + \psi,\ 0 \leqslant j \leqslant N-1 \quad (8)$$

and substitute for $x_j(n-1)$ from (8) into (7) and (6), with $\theta(n)$ also arbitrary. It follows immediately that for each j, there is exactly one point $y_i(n)$, $i \in \Lambda_j(n)$, and hence that

$$x_j(n) = 2\pi j/N + \psi + \alpha(n)\left(\frac{2\pi(i-j)}{N} + \theta(n)\right) = 2\pi j/N + \psi', \quad (9)$$

which, because the difference $i-j$ is independent of j, describes another uniformly-spaced configuration, as was to be proved. □

3.2. Convergence Properties

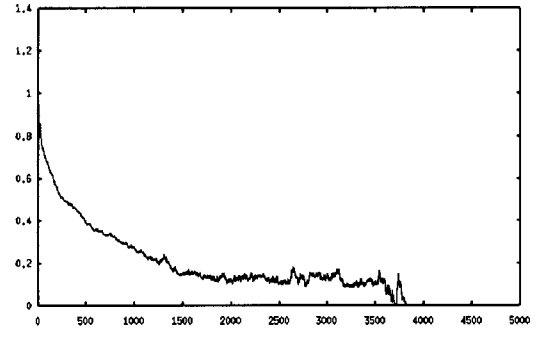
While the persistence of regular configurations is easily seen, the convergence properties of the system are not straightforward. Its evolution may be analysed in terms of the intervals γ_i between the points (with indices modulo N):

$$\gamma_i = x_i - x_{i-1},\ 0 \leqslant i \leqslant N-1 \quad (10)$$

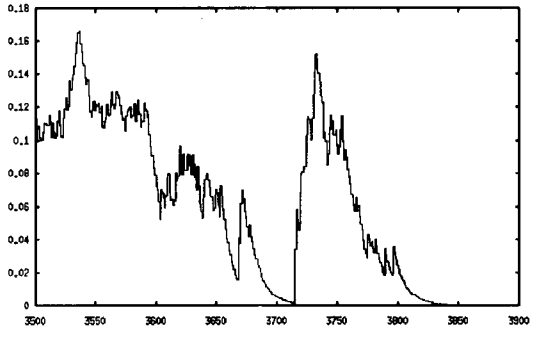
and in terms of a state vector $\phi(n)$, defined as

$$\phi = \left(\gamma_0 - \frac{2\pi}{N}, \gamma_1 - \frac{2\pi}{N}, \ldots, \gamma_i - \frac{2\pi}{N}, \ldots, \gamma_{N-1} - \frac{2\pi}{N}\right)^{\mathrm{T}}. \quad (11)$$

The norm $\|\phi\|$ of the state vector constitutes one possible metric of irregularity for the system, with $\|\phi\| = 0$ if and only if the system is exactly equally



(a)



(b)

FIGURE 4. Convergence of 32-unit circular network of Figure 3 driven by random rotations. Irregularity, expressed as the norm $\|\phi(n)\|$ of the state vector (see text) against iteration number n: (a) coarse resolution; (b) fine resolution.

spaced; $\|\phi(n)\|$ is plotted against iteration number n for a typical run of the system, at low resolution in Figure 4a and at higher resolution in Figure 4b close to convergence. In the Appendix, we prove that the system will converge with probability 1 to exactly equal spacing from an arbitrary initial configuration. Convergence is obviously erratic, but certain features are evident, particularly in Figure 4b, and are confirmed by the analysis in the Appendix. First, the irregularity increases, often dramatically, at certain iterations and decreases, less dramatically, at others. The appearance of Figure 4b at low ordinate values is one of steady, roughly exponential reduction of irregularity, punctuated by relatively large positive-going jumps. Second, the graph of Figure 4b is generally rougher at moderate or high ordinate values and smoother at low ordinate values, implying that there are fewer of these positive-going jumps as the system approaches regularity.

The key factor which determines the behaviour of the network at a given iteration is the mapping (2) between input (rotated) and network points. For certain configurations—particularly those which are close to regular—and certain values of the random rotation, the input points map to network points *bijectively*, or one-to-one, as in the artificial example of Figure 2. It is shown in the Appendix [see eqn (A.24)] that under these circumstances, the irregularity of the system cannot increase, and indeed usually decreases. Thus the positive-going jumps in irregularity visible in Figure 4b correspond necessarily to mappings from input to network which are *injective*, or many-to-one. As was noted above in Theorem 1, when the network is exactly regular, this can never happen since all possible mappings are bijective. If the system is perturbed slightly from regularity, a small set of rotations will now result in an injection while most will still yield a bijection. As the network approaches regularity, therefore, the probability that a randomly chosen rotation will result in an injection—which may dramatically disrupt the system—is reduced. Hence the greater smoothness of the graph of Figure 4b at low ordinate values and the rougher appearance at higher ordinate values. Equation (A.27) confirms a linear dependence on $\|\phi\|$ of an upper bound on the probability that an arbitrary rotation will yield an injective mapping.

The time required for convergence varies widely for a given number of units N, depending on the initial network configuration and the particular sequence of random transformations applied. Indeed, the variance of this quantity is so large as to render its mean rather uninformative about the likely behaviour of a network of N units. It does seem clear, however, that the time required for convergence increases rapidly with the number N of units in the network. Table 1 shows the sample mean and sample standard deviation of the number of iterations required for convergence, for networks containing different numbers N of units, for a constant value of the learning parameter $\alpha(n) = \alpha = 0.1$. The sensitivity to N of the average time to convergence may be at least partially explained by the quadratic dependence on N of the injection probability bound in (A.27); the average time to convergence clearly depends directly on the probability of potentially disruptive injections.

TABLE 1
Convergence Statistics for Circular Network with Constant Learning Parameter $\alpha(n) = \alpha = 0.1$, as a Function of the Number of Units N; 64 Runs Were Made for Each Value of N, and "Convergence" Implies that the Norm $\|\phi(n)\|$ of the State Vector Falls below a Threshold of 10^{-6} and Remains there for at Least a Further 1000 Iterations. Entries Shown Are the Sample Mean and Sample Standard Deviation σ_{64} of the Number of Iterations Required for Convergence

No of Points N	Sample Mean	Sample Standard Deviation
4	138	42
8	278	105
16	1105	642
32	6329	4409

3.3. Discussion

The salient point about the regular arrangement to which the network of Figure 3 converges is that *the exact regularity is an emergent global organisation, induced by random transformations of the network with a local learning rule.* It is also noteworthy that the network attains exact regularity even with a constant value $\alpha(n) = \alpha > 0$ of the learning parameter, without the need for annealing [the gradual reduction of $\alpha(n)$ to zero] which is commonly employed to force networks to converge (cf. Kohonen, 1984). We do, however, consider in the sequel annealed networks with larger numbers N of units.

3.3.1. *A Measure of Self-similarity.* The behaviour of the network on the circle has no obvious connection with self-similarity: in what sense is such a network, driven by transformations, self-similar? Starting from The Euclidean metric in $\mathscr{R}^m$, we are led to the mean square error associated with the point set $X=\{\mathbf{x}_i, 0 \leqslant i \leqslant N-1\}$ under transformations $\mathbf{T} \in \mathscr{T}$
in some group and governed by a probability measure $P(T)$,

$$\varepsilon(X) = \int dP(T) \sum_{i=0}^{N-1} \min_j \|\mathbf{T}\mathbf{x}_i - \mathbf{x}_j\|^2 \qquad (12)$$

which can be simplified by noting that the measure $P(T)$ on the group induces a corresponding density

$p_i(\mathbf{x})$ on each point $\mathbf{x}_i \in X$, representing the probability with which it is displaced by $\mathbf{x}$, giving

$$\varepsilon(X) = \sum_{i=0}^{N-1} \int d\mathbf{y}\, p_i(\mathbf{y} - \mathbf{x}_i) \min_j \|\mathbf{y} - \mathbf{x}_j\|^2. \quad (13)$$

Now each point $\mathbf{x}_i$ has associated with it a Voronoi cell, $R_i \in \mathscr{R}^m$, for which

$$\|\mathbf{x} - \mathbf{x}_i\| = \min_j \|\mathbf{x} - \mathbf{x}_j\|, \ \mathbf{x} \in R_i. \quad (14)$$

It follows that the m.s.e. of (13) can be written in the form

$$\varepsilon(X) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \int_{R_j} d\mathbf{y}\, p_i(\mathbf{y} - \mathbf{x}_i) \|\mathbf{y} - \mathbf{x}_j\|^2. \quad (15)$$

For a given density of displacements, this is just the squared error between the displaced version of a point and the nearest point in the set, averaged over all points in the set and all displacements. This seems a reasonable definition for a measure of self-similarity under the specified transformation probability distribution. For example, if the point set X were the set of vertices of a regular polygon and the group were the (finite) symmetry group of the figure, then the error would be zero. Because the groups of interest are continuous, the error will always be non-zero, but it will be smaller, the better matched is the spatial distribution of X to the transformation statistics.

3.3.2. *Network Dynamics as Gradient Descent in the Mean.* For rotations on the circle, because the displacement density is independent of the origin, (15) takes the particularly simple form

$$\varepsilon(X) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \int_{\frac{x_{j-1}+x_j}{2}}^{\frac{x_j+x_{j+1}}{2}} dy\, p(y - x_i)(y - x_j)^2. \quad (16)$$

Differentiating (16) with respect to the point x_i and simplifying, we get

$$\frac{\partial \varepsilon}{\partial x_i} = -2 \sum_{j=0}^{N-1} \int_{\frac{x_{i-1}+x_i}{2}}^{\frac{x_i+x_{i+1}}{2}} dy\, p(y - x_j)(y - x_i) + \sum_{j=0}^{N-1} \int_{\frac{x_{j-1}+x_j}{2}}^{\frac{x_j+x_{j+1}}{2}} dy \frac{\partial p(y - x_i)}{\partial x_i} (y - x_j)^2 \quad (17)$$

which takes account of the fact that varying the limits of the integral has no effect on the m.s.e. Now from (5) and (6), the expected change in x_i for a constant value of the learning coefficient $\alpha(n) = \alpha$ is given by

$$E[\Delta x_i] = \alpha \sum_{j=0}^{N-1} \int_{\frac{x_{j-1}+x_j}{2}}^{\frac{x_j+x_{j+1}}{2}} dy\, p(y - x_j)(y - x_i) \quad (18)$$

and so if the second term in (17) is zero, we have

$$E[\Delta x_i] = -\frac{\alpha}{2} \frac{\partial \varepsilon}{\partial x_i}. \quad (19)$$

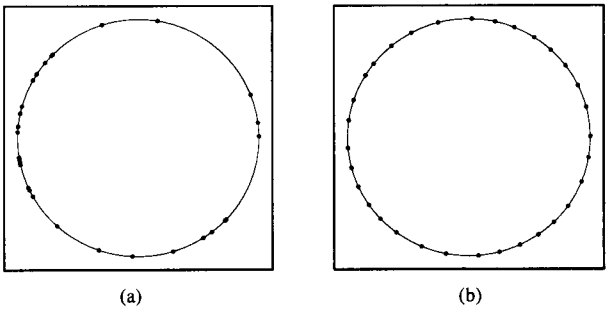
Thus in the case of rotations on the circle under a uniform density, the network update rule performs gradient descent *in the mean* on the m.s.e. defined in (16). It is not hard to see that this carries over to networks on the m-torus driven by m-D translations. In general, however, (17) must be replaced by

$$\frac{\partial \varepsilon}{\partial \mathbf{x}_i} = -2 \sum_{j=0}^{N-1} \int_{R_i} d\mathbf{y}\, p_j(\mathbf{y} - \mathbf{x}_j)(\mathbf{y} - \mathbf{x}_i) + \sum_{j=0}^{N-1} \int_{R_j} d\mathbf{y} \frac{\partial p_i(\mathbf{y} - \mathbf{x}_i)}{\partial \mathbf{x}_i} \|\mathbf{y} - \mathbf{x}_j\|^2. \quad (20)$$

If $p_i(\mathbf{x})$ is smooth and the density of points is high, the second term, which is quadratic in the radius of the Voronoi cells, will make a negligible contribution to the gradient. It follows that in this case too, the network will tend on average to perform gradient descent on the m.s.e.

3.3.3. *Comparison with Network Driven by Input Transformations.* For comparison with the directly transformation-driven network (type C), we consider briefly a similar network driven by input transformations (type B). Such networks lack the recursion inherent in the type C networks, and are effectively just Kohonen SOFM networks without the usual annealing of the learning parameter.

The learning rule remains unchanged, but the input vector set at each iteration is now a randomly rotated version of a fixed prototype $\{z_i\}$. Thus the nth iteration involves the following steps:



(a) (b)

FIGURE 5. 32-unit network on the circle, driven by random rotations of a fixed random prototype pattern. (a) Initial network configuration and random prototype; (b) network after 5000 iterations with learning parameter $\alpha(n) = \alpha = 0.1$. The network attains only approximate regularity, and is perturbed by each new input presentation in the absence of annealing.

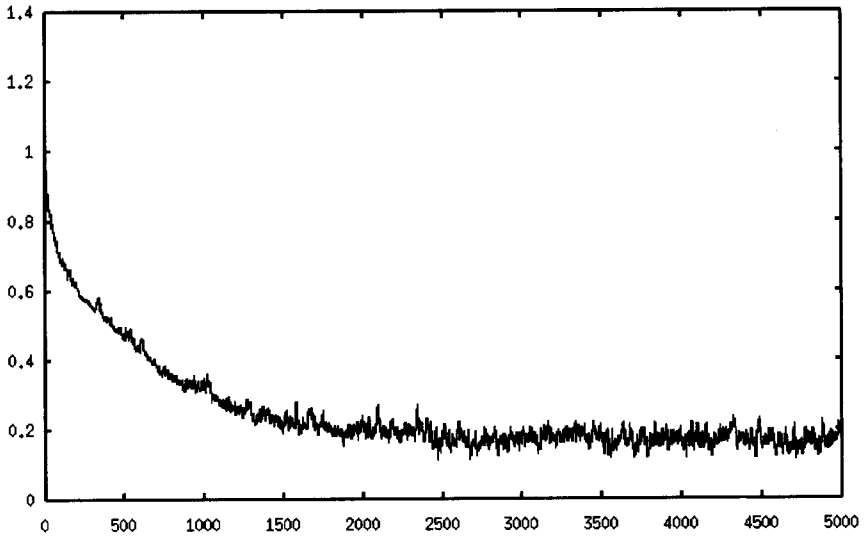


FIGURE 6. "Convergence" of 32-unit circular network of Figure 5, driven by random rotations of a fixed random prototype pattern. Irregularity of the network, expressed as the norm $\|\phi(n)\|$ of the state vector. Unlike the network driven directly by transformations (Figures 3 and 4), this network does not in fact converge.

(i) Select a random rotation $\theta(n)$ uniform on $[0, 2\pi)$.
(ii) Apply this rotation to each vector z_i in the random prototype, to give an input vector $y_i(n)$:

$$y_i(n) = z_i + \theta(n). \tag{21}$$

(iii) Update the network vector set $\{x_i(n-1)\}$ according to the learning rule of (1) and (2) to give the new network $\{x_i(n)\}$.

Figure 5a shows the initial configuration of such a network on the circle (it is identical to Figure 3a), which is also used as the random prototype. Figure 5b shows the configuration after 5000 iterations with $\alpha(n) = \alpha = 0.1$. For comparison with Figure 4a, Figure 6 shows the same measure of irregularity for the network of Figure 5. The present network clearly does not attain exact regularity; it attains an approximate regularity in response to the gross density of input points, but in the absence of annealing, continues to respond to each new input presentation and hence the irregularity fluctuates randomly about a nonzero mean level visible in the figure.

4. TWO-DIMENSIONAL NETWORKS

We have also experimented with networks in which the units are characterised by vectors in $\mathcal{R}^2$, using transformations drawn from a variety of symmetry groups and under various boundary conditions.

4.1. Networks on the Torus Driven by 2-D Translations

The simplest such system is the 2-D analogue of the circular network described in Section 3 above. It has periodic (toroidal) boundary conditions, and is driven by 2-D translations drawn from a uniform density on $[0, 2\pi) \times [0, 2\pi)$. The initial configuration is similarly distributed.

4.1.1. *A 32-unit Toroidal Network.* Figure 7 shows the behaviour of a 32-unit toroidal network with the learning parameter $\alpha(n)$ set to a constant value of 0.1. The initial configuration is shown in Figure 7a, and the result after 10,000 iterations in Figure 7b. This network converges to regularity within several thousand iterations, although as in the 1-D case, the time taken varies widely for different runs with different initial conditions and random transformations applied.

Convergence is to a lattice or doubly periodic structure, just as the 1-D version converged to a singly periodic (regular) distribution on the circle. This lattice structure is a fixed point up to an

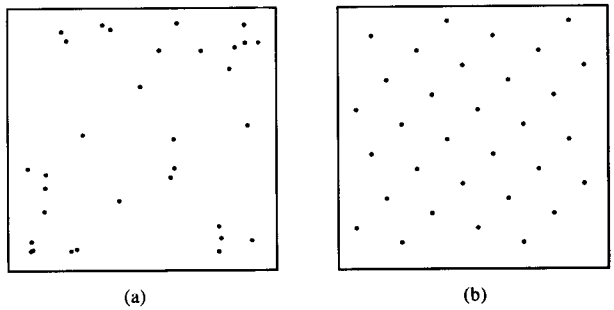
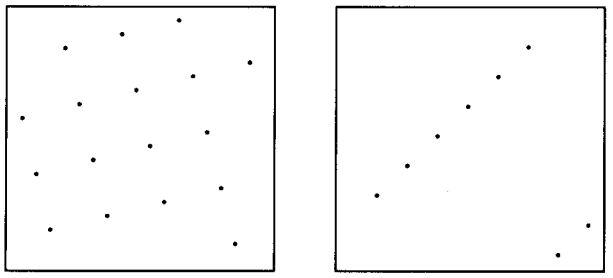
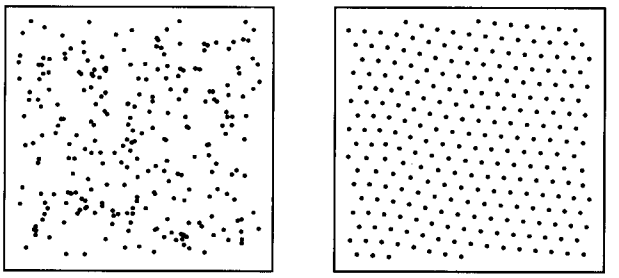


FIGURE 7. 32-unit network on the torus, driven by random 2-D translations. (a) Initial configuration; (b) after 10,000 iterations with learning parameter $\alpha(n) = \alpha = 0.1$. The network again converges to exactly regular spacing after several thousand iterations. Note that there is a narrow margin between the boundary of the displayed torus and the frame.



(a) (b)

FIGURE 8. (a) Typical fixed point of a 17-unit network on the torus; (b) a helical lattice is a stable fixed point, as is any regular lattice on the torus. This 8-unit network was initialised close to the helical configuration shown.



(a) (b)

FIGURE 9. 256-unit network on the torus, driven by random 2-D translations. (a) Initial configuration; (b) after one million iterations. In this and subsequent results for 256-unit networks, the learning parameter $\alpha(n)$ was annealed to force approximate convergence in reasonable time (see text). The structure obtained, however, still shows considerable regularity over much of the network.

arbitrary 2-D translation, by a trivial extension of Theorem 1. Convergence to a regular lattice structure has been obtained for all numbers N of points up to $N = 32$, with the time taken again increasing rapidly with N on the whole. Particular numbers of points, e.g., $N = 17$, seem to require many iterations on average to converge; it seems that the network has some difficulty in finding the basins of attraction of stable fixed points in such cases. A configuration to which a 17-unit network does eventually converge is shown in Figure 8a.

While an approximately hexagonal lattice is preferred in general, any regular lattice is a fixed point of the learning rule under 2-D translations on the torus. Such configurations appear also to be stable, in that if the network is initialised sufficiently close to such a configuration, it will converge there rather than to another, say near-hexagonal, configuration. Figure 8b shows a stable helical lattice configuration in an eight-unit network.

4.1.2. *Annealed 256-unit Toroidal Network.* One solution to the dramatic increase in convergence time with network size is to force convergence, albeit to some perhaps less regular state, by reducing the learning parameter $\alpha(n)$ with the iteration number n according to some *annealing schedule*. This approach has been widely used in stochastic optimisation schemes (e.g., Geman & Geman, 1984) and was used by Kohonen (1984) to force networks to approach some stable configuration rather than continuing to oscillate about that configuration with each new input presentation like the unannealed network of Section 3.3.3. In all of the work reported here on two-dimensional networks of $N = 256$ units, $\alpha(n)$ was maintained at a constant value of 0.01 for the first 99% of the iterations performed, and thereafter was reduced linearly to zero (cf. Kohonen, 1984).

We first consider a larger toroidal network of the type described above: the network consists of 256 units, each characterised by a 2-D vector on $[0, 2\pi) \times [0, 2\pi)$ (the *arena*), driven by uniform 2-D translations for one million iterations. Figure 9 shows the initial and final configurations. While the network has not attained complete regularity, substantial areas of regularity are visible and a hexagonal lattice pattern predominates, punctuated by dislocations. As will become apparent, the hexagonal lattice pattern is characteristically induced in large networks driven by random 2-D translations.

4.2. Annealed 256-unit Networks on a Bounded Disk

We consider in this section 256-unit networks on a bounded, nonperiodic arena. The learning parameter $\alpha(n)$ is annealed according to the same schedule as employed for the 256-unit toroidal network discussed above. The learning rule on a bounded arena is slightly modified:

(i) units which leave the arena under transformations are ignored, and do not induce updates to any network unit;
(ii) despite (i), network units can still migrate beyond the arena due to experiencing multiple additive updates at some iteration. Such units are replaced on the arena boundary.

4.2.1. *Network on the Disk Driven by Elementary Transformations.* Figure 10a shows a uniform initial scattering of points on a bounded disk. Figure 10b shows the effect of rotations about the centre, drawn from a density uniform on $[0, 2\pi)$. Only 10,000 iterations were performed in this case; since updates are along chords rather than arcs, a system driven only by rotations will collapse toward the centre. Figure 10c shows the effect of dilations about the centre, with the log of the dilation factor uniform on $[\log(1/16), \log(16)]$ for 100,000 iterations. Figure 10d shows the effect of 2-D translations, uniform on a disk of the same size as the arena with the null

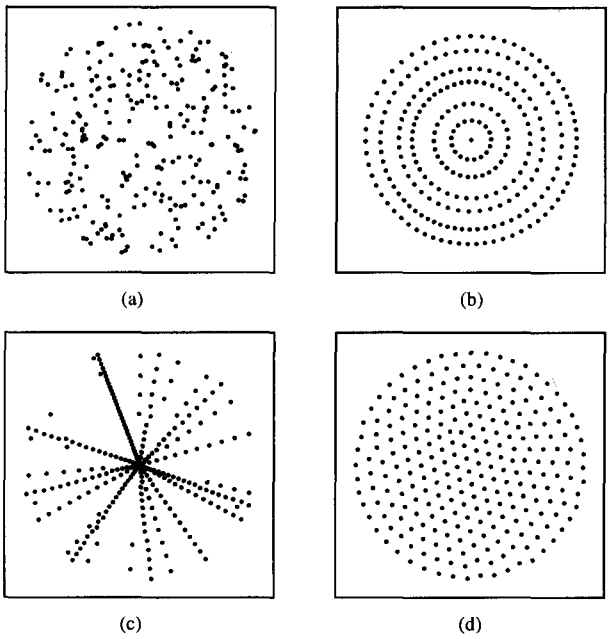


FIGURE 10. 256-unit network on a disk, driven by elementary transformations. (a) Typical initial configuration; (b) driven by rotations about the centre, 10,000 iterations; (c) driven by dilations about the centre, 100,000 iterations; (d) driven by 2-D translations uniform on a disk of the same size as the arena, 100,000 iterations.

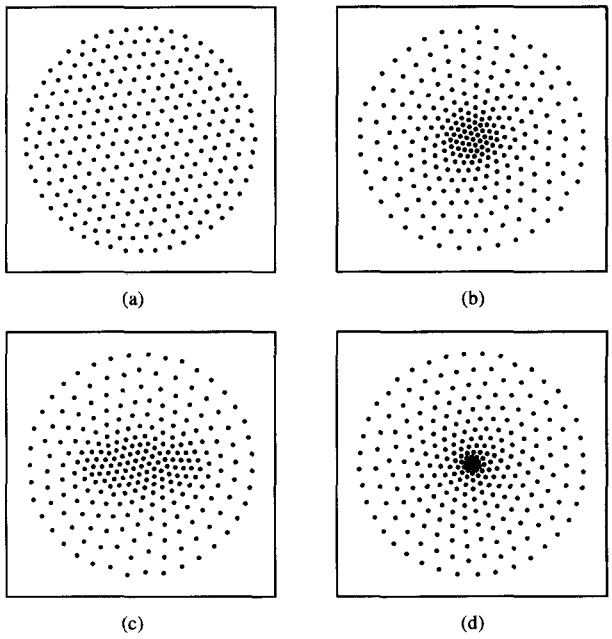


FIGURE 11. 256-unit network on a disk, driven by composite transformations (rotation + dilation + translation) with varying translational component. (a) Translations uniform on a disk the same size as the arena; (b) translations uniform on a disk 0.2 times the linear size of the arena; (c) translations uniform on an ellipse 0.5 times the size of the arena horizontally and 0.2 times the size of the arena vertically; (d) no translations (i.e., point translation density). One million iterations were performed in each case.

translation $[0,0]^T$ at the centre, again for 100,000 iterations.

It is clear that such systems tend to become regularly spaced with respect to the symmetry groups from which the transformations are drawn; 2-D translations, whether on the torus or the bounded disk, tend to induce a periodic lattice arrangement with a hexagonal lattice being preferred in general. Rotations on the disk induce first an aggregation into concentric rings, and then a more even angular distribution of units within each ring (plus a tendency for the whole figure to contract in the process, as noted above). Dilations induce an aggregation into radial "spokes", and then a more even distribution with respect to log r of units within each spoke.

4.2.2. *Network on the Disk Driven by Composite Transformations.* Figure 11a shows the effect on the network of composite transformations which are a combination of the elementary transformations used in Figures 10b–10d. Each transformation consists of a random rotation about the centre uniform on $[0, 2\pi)$, a random dilation about the centre with log of the dilation factor uniform on $[\log(1/16), \log(16)]$, and a random 2-D translation uniform on a disk of the same size as the arena. One million iterations were performed. Despite the use of composite transformations, the result is dominated by the effects of the translation component. An approximately hexagonal lattice covers much of the figure, with significant distortion only at, and induced by, the boundary. The circular and radial organisations characteristic of rotations and dilations (Figures 10b–10c) are not in evidence.

Of considerable interest is the effect of reducing the maximum magnitude of the translation component from the relatively large value used in the example of Figure 11a. In vision, effective object translations may be restricted by the use of eye movements which track translating objects (but do not, of course, compensate for object rotation or dilation), and the tendency of mammalian vision to attempt to fixate a point in the visual environment will lead to small translations as the gaze drifts from the nominal fixation point and is brought back by active correction. Therefore it seems appropriate to consider the effects of rotations, dilations, and restricted translations as arguably a more representative model of the transformations "seen" by visual systems under certain conditions. Figure 11b shows the effect of such transformations. The rotation and dilation components were as in Figure 11a, but the translations are now uniform on a disk only 0.2 times the linear size of the arena. A *foveal* region of high unit density, possessing the hexagonal lattice structure characteristically induced by translations, is present at the centre of the network. Its spatial extent corresponds approximately to the density from

which the translations are drawn, at least for the parameter set used in this case. (The foveal size has also been observed to depend rather strongly on the range of dilations used, for example: a wider range of dilations leads to a smaller and denser fovea for a given density of translations.)

To test the relationship between the translation density and the structure of the resulting fovea, we varied the shape of the density. Figure 11c shows the result for translations uniform on an ellipse of 0.5 times the linear size of the arena horizontally and 0.2 vertically. Figure 11d shows the result for no translations (i.e., a "point" translation density). From these figures, it seems that the spatial extent of the induced fovea indeed closely reflects the density from which the translation component of the composite transformations is drawn. The periphery tends to possess the spiral organisation characteristically induced by the rotational and dilational components and visible in isolation in Figure 11d, but also shows a local hexagonal structure.

4.3. Discussion

4.3.1. *Self-similarity of Foveal Networks.* While it is difficult to show analytically that such a foveal arrangement is optimal under a criterion of self-similarity, there are heuristic and empirical reasons for supposing it to be so (see also Sections 3.3.1 and 3.3.2). Consider the sample-average squared error $\varepsilon(\mathbf{T})$ between the transformed and network vector sets, defined as

$$\varepsilon(\mathbf{T}) = 1/N_a \sum_i \min_j \|\mathbf{T}\mathbf{x}_i - \mathbf{x}_j\|^2 \qquad (22)$$

where the sum is over the $N_a \leqslant N$ transformed vectors $\mathbf{T}\mathbf{x}_i$ which fall within the arena. We estimated the average error $E[\varepsilon(\mathbf{T})]$ [cf. the self-similarity measure defined in (12)] as a sample average over 10,000 transformations, for each of the four networks shown in Figure 11, under each of the four transformation densities used to derive those networks. The results are shown in Table 2 and should be compared within each column, corresponding to the performance of the various networks on a given density of transformations. It is clear that each network is more self-similar in terms of this error measure than any of the others, with respect to the density of transformations under which it evolved.

TABLE 2
Measured Self-similarity of the Various Foveal Networks Shown in Figure 11 with Respect to Each of the Transformation Densities Used to Derive those Networks. The Measure Tabulated is the Normalised (per Column) Sample Average Squared Error Attained by Each Network under Each Transformation Density (see Text). The Averages were Performed over 10,000 Randomly Selected Composite Transformations

	Tx (a)	*Tx* (b)	*Tx* (c)	*Tx* (d)
Nwk (a)	1	1.62	1.44	1.87
Nwk (b)	1.29	1	1.07	1.06
Nwk (c)	1.27	1.08	1	1.18
Nwk (d)	1.25	1.02	1.09	1

4.3.2. *Perturbation Behaviour.* While the results of the previous section are encouraging, they do not show that the networks have found *maxima* of self-similarity (i.e., minima of the average error $E[\varepsilon(\mathbf{T})]$) even locally, let alone globally. Accordingly, we repeated the experiment with perturbed versions of the networks of Figure 11. The networks were perturbed by applying a random 2-D translation to each unit, uniform on a small disk centred on the unit. As the radius of that disk (the *perturbation radius*) increases, the network becomes more perturbed. Figure 12 shows, for each network, the dependence of the average error on the normalised perturbation radius (= perturbation radius/arena radius). Each perturbed network was driven with the same density of transformations under which its unperturbed counterpart evolved.

In each case, the relative squared error increases roughly parabolically for values of the perturbation radius significantly smaller than the inter-unit spacing, whence the perturbation does not affect the input-network mapping for most points and most

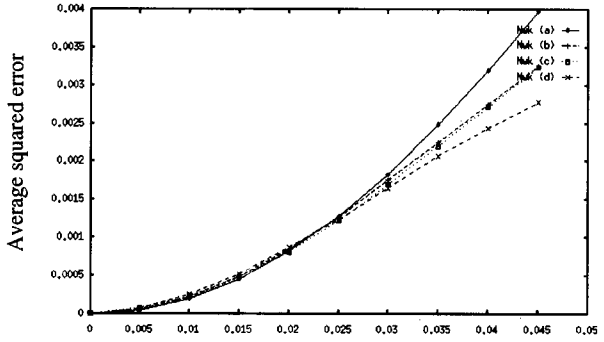


FIGURE 12. Self-similarity of perturbed versions of the various foveal networks of Figure 11, with respect in each case to the transformation density used to derive the unperturbed network. The graph for each network shows the sample average squared error between transformed and untransformed versions of the perturbed network, as a function of normalised perturbation radius (see text). The error was averaged over 1000 transformations for each data point, and the error for zero perturbation has been subtracted from all points so that the graphs pass through (0, 0) (i.e., the error shown is relative to the unperturbed case). In absolute terms, the squared error at the largest value of perturbation shown here is about 20–30% larger than in the unperturbed case.

transformations. At larger values of perturbation radius, the error falls below parabolic; the network which exhibits this effect first is network (d) since it has the smallest inter-unit spacing of any of the four networks at its centre. While this experiment does not prove the networks to be maximally self-similar under the respective densities, it strongly suggests that they have found at least a local minimum of the error, and the clear relationship between the form of the transformation density and the structure of the network suggests that the minimum found in each case is close to the global minimum.

4.3.3. *Prediction and Self-similarity.* There are two qualitatively different levels at which these networks organise under the influence of transformations. There is a *local* level of organisation, at which, for example, 2-D translations induce a regular, approximately hexagonal lattice structure in the network. Then there is a *global* level of organisation, at which, for example, the foveal region occupied by such a lattice structure is of a spatial extent determined by the spatial extent of the density from which the translations are drawn.

These results suggest that foveal structures, as well as providing a higher central resolution, might also be close to optimal under the apparently unrelated criterion of self-similarity, under 2-D transformations which include a restricted translational component, such as might arise in vision as a result of tracking eye movements. Furthermore, where the translation density is anisotropic, it seems that a comparably anisotropic foveal structure maximises self-similarity under that density. While not conclusive, it appears from visual physiology (Hughes, 1977) that possession of a pronounced retinal *visual streak*—an elongated horizontal region of high receptor and ganglion cell density—is correlated with species which dwell in open country, and which might be expected to scour their visual environment with a bias toward horizontal eye movements. The visual streak seems to be largely absent in forest- and undergrowth-dwelling species, for whom eye movement is less likely to have a horizontal bias. From Hughes (1977, p. 709):

> . . . the visual streak is common to terrestrial species whose field of view is not completely obscured by nearby vegetation [. . .]; the terrain surface is overt in all but the most overgrown woods, above the stream bed to a fish, or above the sea to a flying or floating bird.

There is an interesting connection between self-similar point sets and signal prediction, which may help to shed some light on why such a mechanism might be involved in the development of retinal structures. Suppose that a system's only information about a continuous spatiotemporal signal $s(\mathbf{x}, t)$ is a set of samples taken at some instant of time $t = 1$, say, at the points $\mathbf{x}_i \in X$, giving sample values $s_1(\mathbf{x}_i), 0 \leqslant i \leqslant N-1$. From these samples, it is required to predict the signal at the next sampling time $t = 2$, $s_2(\mathbf{x}) = s_1(\mathbf{T}\mathbf{x})$, which is obtained from the signal at time $t = 1$ by a motion, represented by the transformation $\mathbf{T}$. The transformations are drawn at random from some group, with probability measure $P(T)$. If we take as a prediction for each sample $s_2(\mathbf{x}_i)$ the sample $s_1(\mathbf{x}_j)$ which minimises, over the signal statistics, the expected error

$$e_i = E\left[\left|s_2(\mathbf{x}_i) - s_1(\mathbf{x}_j)\right|^2\right] = \min_k E\left[\left|s_2(\mathbf{x}_i) - s_1(\mathbf{x}_k)\right|^2\right] \quad (23)$$

then which set of sampling points $X = \{\mathbf{x}_i\}$ would minimise over the transformation statistics the total mean squared prediction error (m.s.p.e.)?

For a given transformation $\mathbf{T}$, the average error associated with the ith point is

$$\begin{aligned} e_i(\mathbf{T}) &= \min_k E\left[\left|s_1(\mathbf{T}\mathbf{x}_i) - s_1(\mathbf{x}_k)\right|^2\right] \\ &= 2\min_k (R_s(\mathbf{0}) - R_s(\mathbf{T}\mathbf{x}_i - \mathbf{x}_k)), \end{aligned} \quad (24)$$

where it is assumed that the signal is spatially wide sense stationary, with isotropic autocorrelation function $R_s(\mathbf{x}) = R_s(\|\mathbf{x}\|)$. Since $R_s(\mathbf{x})$ is an even function of $\|\mathbf{x}\|$ and has a maximum at $\mathbf{0}$, its Taylor expansion can be written as

$$R_s(\mathbf{x}) = R_s(\mathbf{0}) + \frac{1}{2!} R_s''(\mathbf{0})\|\mathbf{x}\|^2 + O(\|\mathbf{x}\|^4). \quad (25)$$

Thus, provided the sampling density is high enough, the average error $e_i(\mathbf{T})$ is given by

$$e_i(\mathbf{T}) = -R_s''(\mathbf{0}) \min_k \|\mathbf{T}\mathbf{x}_i - \mathbf{x}_k\|^2. \quad (26)$$

Note that the second derivative of the autocorrelation is negative at $\mathbf{0}$. Summing over all points $\mathbf{x}_i \in X$ and taking expectations with respect to the transformations gives the total m.s.p.e. as

$$\begin{aligned} e(X) &= -R_s''(\mathbf{0}) \sum_i \int dP(T) \min_k \|\mathbf{T}\mathbf{x}_i - \mathbf{x}_k\|^2 \\ &= -R_s''(\mathbf{0}) \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \int_{R_j} d\mathbf{y}\, p_i(\mathbf{y} - \mathbf{x}_i) \|\mathbf{y} - \mathbf{x}_j\|^2 \end{aligned} \quad (27)$$

which is identical to the original equation defining self-similarity (15), apart from the multiplicative constant $-R_s''(\mathbf{0})$. In other words, the optimal sampling point set X is just the one which is self-similar under the transformations. Although the above demonstration is based on prediction using

single points, it seems likely that a similar result will obtain if small *neighbourhoods* within X are used in the prediction, rather than single points.

The above discussion shows why we have concentrated on self-similar networks. If the input signal transforms according to some density on a group, then the network which provides optimal sampling for motion prediction is maximally self-similar under that density, and may be generated by driving a network directly with the same transformations. Furthermore, if some adaptive mechanism acts (whether over the lifetime of an individual or over many generations) to reduce the average motion prediction error, then this mechanism will give rise in the same way to a maximally self-similar array.

5. FUNCTION NETWORKS

The self-organisational principles under which these networks operate may be extended to networks in which the units are characterised by parametrised functions corresponding, for example, to the receptive field profiles of visual cortical neurons. Indeed, such receptive field profiles are most commonly described in terms of parameters such as orientation, position and spatial frequency, often with bandwidths attached which express the degree of selectivity in each parameter. The local mapping of such parameters in primary visual cortex has received much attention (Schwartz, 1980; Swindale, 1985; Hubel, 1988; Durbin & Mitchison, 1990) but rather less has been paid to the global regularity of the machinery—which we suggest may approach self-similarity under typical visual transformations—and to the view of transformations as fundamental in shaping the structure of perception, which has been largely the preserve of psychologists and group theorists (Gibson, 1950; Hoffman, 1966; Shepard, 1981; Caelli & Dodwell, 1982; Dodwell, 1983).

Whereas the point networks of the previous sections can represent an impulsive sampling grid, a function network can represent a basis function set into which an input might be decomposed, much as visual input is represented by the activities of neurons with particular forms of receptive field profile. A *maximally self-similar basis function set* under a given set of transformations is then a set of functions which most closely resembles itself on average under the coordinate transformations. This implies that the effect of motions on the input can be approximated by permutations of the output (cf. Section 4.3.3 and Simoncelli et al., 1992), thus decoupling the representation of the *form* of the input (the perceptual "what") from the representation of the transformation acting upon it (the perceptual "where").

In the point networks of the previous sections, the metric used to rank the similarity of transformed and untransformed points was simple Euclidean distance. In function networks, we have to replace that metric by an inner-product-based measure, and also ensure that all basis functions are normalised in order to render such inner products meaningful in terms of similarity. However, there remains a problem of computational burden, and for this reason we have chosen to consider functions represented by points in a parameter space rather than to directly manipulate the functions, which would involve performing a numerical integration for each inner product and normalisation. In a suitable parameter space, while the class of functions is perhaps unrealistically restricted, an inner product between any two functions can be defined in closed form in terms of their respective parameters. This approach has also been used by Durbin and Mitchison (1990) in their work on cortical maps.

5.1. Parametrised Gaussian Networks

Each unit in these networks is associated with a three-dimensional parameter vector. For the ith unit, the parameters u_i, v_i encode (centre) position, just as in the 2-D point networks. Now there is an extra parameter σ_i^2 which encodes the variance (spatial scale) of a circularly symmetric unit-energy 2-D Gaussian centred at (u_i, v_i) and given by

$$G_i(u,v) = \frac{1}{\sigma_i\sqrt{\pi}} \exp - \frac{(u-u_i)^2 + (v-v_i)^2}{2\sigma_i^2}. \qquad (28)$$

A co-ordinate transformation applied to this Gaussian will give another Gaussian with parameters which are simply related through the transformation to those of the original.

At each iteration, we apply a random composite transformation to the network. Having obtained a set of parameter vectors describing the transformed network, we then find, for each transformed unit, the "nearest" untransformed unit. As discussed above, this now involves an inner product similarity measure rather than Euclidean distance.

The inner product of two Gaussians with different positions and scales has a simple closed-form expression in terms of the parameters,

$$\langle G_i, G_j \rangle = \iint G_i(u,v) G_j(u,v) \, du \, dv$$
$$= \frac{2\sigma_i\sigma_j}{\sigma_i^2 + \sigma_j^2} \exp - \frac{(u_i-u_j)^2 + (v_i-v_j)^2}{2(\sigma_i^2+\sigma_j^2)}. \qquad (29)$$

We must also reformulate the learning rule to specify how the various parameters will be updated. We update the centre position linearly within limits

imposed by the arena boundary, with learning parameter $\alpha(n)$, exactly as for the point networks of previous sections. We also update the variance σ_i^2 linearly within hard limits, with the same value of learning parameter. Thus if each unit is characterised by a vector $\mathbf{x}_i = [u_i, v_i, \sigma_i^2]^T$, the form of update rule given by eqn (1) still holds within the hard limits. The set defined by (2), however, is now expressed in terms of inner products:

$$\Lambda_j(n) = \{i : \langle \mathbf{y}_i(n), \mathbf{x}_j(n-1) \rangle > \langle \mathbf{y}_i(n), \mathbf{x}_k(n-1) \rangle,\ k \neq j\}, \tag{30}$$

where the inner product of the parameter vectors $\mathbf{a} = [u_a, v_a, \alpha_a^2]^T$ and $\mathbf{b} = [u_b, v_b, \sigma_b^2]^T$ is defined in terms of their components as

$$\langle \mathbf{a}, \mathbf{b} \rangle = \frac{2\sigma_a\sigma_b}{\sigma_a^2 + \sigma_b^2} \exp - \frac{(u_a - u_b)^2 + (v_a - v_b)^2}{2(\sigma_a^2 + \sigma_b^2)}. \tag{31}$$

Figure 13 shows the results of 128- and 256-unit Gaussian networks on the disk, directly driven by composite transformations for 100,000 iterations. The annealing schedule was somewhat "hotter" over the early stages than for the point networks, with $\alpha(n)$ maintained at a constant value of 0.05 for the first 95% of the iterations performed, and thereafter reduced linearly to zero. The figures plot each Gaussian as a circle drawn at one standard deviation from the centre, with thickness increasing with radius because this makes it easier to see the various scales present in any given region (a slight defocusing of the reader's vision may also help).

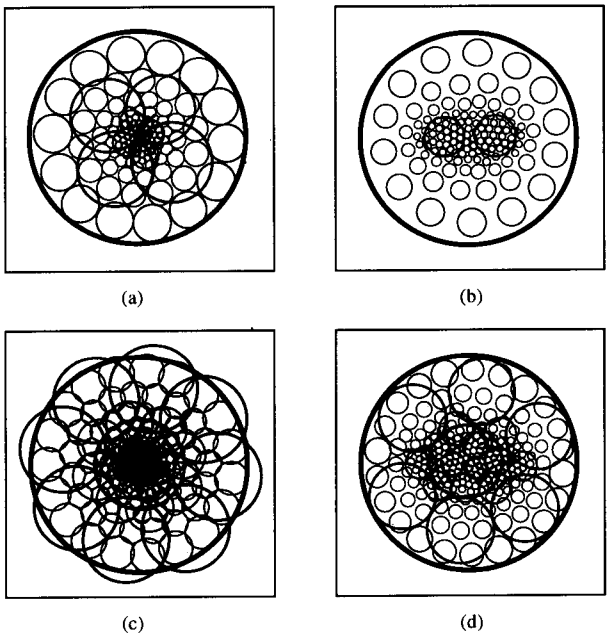


FIGURE 13. 128- and 256-unit Gaussian networks on the disk, driven by composite transformations (rotation + dilation + translation) with varying translational component. (a) 128-unit network: translational component uniform on a disk 0.2 times the linear size of the arena; (b) 128-unit network: translational component uniform on an ellipse 0.5 times the size of the arena horizontally and 0.2 times the size of the arena vertically; (c) 256-unit variant of (a); (d) 256-unit variant of (b). 100,000 iterations were performed in each case.

Figure 13a shows a Gaussian network of 128 units, driven by transformations drawn from the same density as induced the circular fovea in the network of Figure 11b. Figure 13b shows a similar 128-unit Gaussian network, but driven by transformations drawn from the density which induced the elliptic fovea in the network of Figure 11c. Figures 13c and 13d are 256-unit variants of (a) and (b) respectively.

5.2. Discussion

Again, the application of transformations with restricted translational component induces a foveal region in the networks. The fovea contains the smallest Gaussian units, packed in a configuration which is close to a hexagonal lattice, and is of a spatial extent which corresponds to the density of the translational component of the applied transformations. Figure 13a shows a roughly circular fovea and Figure 13b an elliptic fovea, of comparable dimensions to the foveal regions which emerged in Figures 11b and 11c. In addition, we now observe that multiple scales are present at any spatial position. (Note that the largest circle in each of these examples is a Gaussian of large scale, not a picture frame.) Figures 13a and 13b hint that the foveal structure may be similar over a range of scales. For example, the pair of larger-scale units near the centre of figure 13b is oriented horizontally like the major axis of the smaller-scale fovea. However, 128 units is too few to express such a structure clearly, and hence we show corresponding examples of 256-unit networks in Figures 13c and 13d.

While more crowded, the 256-unit examples, and particularly the elliptic-fovea example of Figure 13d, illustrate that the foveal structure is indeed similar at different scales. In that figure, there are two clearly visible foveal groupings of different average scales which occupy the same elliptically-shaped region at the centre of the figure, resembling overlying layers. There is also a larger-scale structure at the centre at a further two scales, including the single unit which comprises the largest Gaussian in the network.

In all these examples, it is clear that a more or less orderly arrangement of multiple scales emerges, with a spacing out in the scale dimension of units located at similar centre positions, together with a spacing out of spatial position within each scale

layer. The smallest scales are present only within the foveal region, the spatial extent of which depends on the density of the translational component of the applied transformations, and which possesses a local organisation which approximates the hexagonal lattice pattern characteristically induced by 2-D translations. Within each layer, the unit scale varies little within the fovea, but tends to increase steadily with eccentricity in the periphery. In the example of Figure 13c, for example, a wide range of scales is represented at the centre and in the fovea as a whole, but in the far periphery, only units of larger scale are found. This organisation exhibits certain features of both artificial and biological vision systems. The fact that the structure is layered in scale, rather than more continuously or randomly distributed, is a feature shared with all artificial multiple resolution systems to date (Burt & Adelson, 1983; Rosenfeld, 1984), while the increase in receptive field size in the periphery is characteristic of mammalian vision (Hughes, 1977).

6. CONCLUSIONS: TRANSFORMATIONS IN VISION

We have shown in this work that a simple adaptation of the Kohonen SOFM can, with appropriate choice of "stimulus", lead to self-similar 1-D and 2-D point and function sets. We have examined the convergence of the networks and proved a strong convergence result in one important case. We have also shown how self-similarity leads to efficient motion prediction.

Clearly, any system concerned with perception of an environment dominated by motion will find some advantage in optimising its prediction of that motion: only by such means can it reduce the flow of data impinging on it to manageable proportions. Even the rudimentary motion estimation methods used in the current generation of image sequence coding systems give significant reductions in data rate (LeGall, 1991). It therefore seems plausible that the nonuniform sampling exhibited by the retinae of so many animals may have been adapted over time for the purposes of motion prediction and representation. That this should lead to self-similar sampling arrays is perhaps not obvious at first glance, but turns out to be the case. Whether the crude mechanism that we have reported, based on a Kohonen update rule, has any biological plausibility is another question. The only conclusion we can reach from the work reported here is that through comparatively simple adaptation mechanisms, it is possible to develop visual representations which are well adapted to the representation of visual motion.

There is still much to do, both in extending the theoretical results and in exploring function networks. We feel that the results we have achieved thus far are encouraging enough to make such developments worthwhile.

REFERENCES

Barrow, H. G. (1987). Learning receptive fields. *Proceedings of the IEEE First Annual Conference on Neural Networks.*

Burt, P. J., & Adelson, E. H. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, **C-31**, 532–540.

Caelli, T., & Dodwell, P. C. (1982). The discrimination of structure in vectorgraphs: local and global effects. *Perception and Psychophysics*, **32**(4), 314–326.

Cavanagh, P. (1978). Size and position invariance in the visual system. *Perception*, **7**, 167–177.

Dodwell, P. C. (1983). The Lie transformation group model of visual perception. *Perception and Psychophysics*, **34**(1), 1–16.

Doob, J. L. (1990). *Stochastic processes.* orig. 1953, Wiley Classics Library Edition, 1990.

Durbin, R., & Mitchison, G. (1990). A dimension reduction framework for understanding cortical maps. *Nature*, **343**, 644–647.

Fukushima, K., Miyake, S., & Ito, T. (1983). Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Transactions on Systems, Man and Cybernetics*, **SMC-13**, 826–834.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**, 721–741.

Gibson, J. J. (1950). *The perception of the visual world.* Boston, MA: Houghton Mifflin.

Hoffman, W. C. (1966). The Lie algebra of visual perception. *Journal of Mathematical Psychology*, **3**, 65–98.

Hubel, D. H. (1988). *Eye, brain and vision.* San Francisco: Freeman.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology*, **160**, 106–154.

Hughes, A. (1977). The topography of vision in mammals of contrasting life style: comparative optics and retinal organization. In F. Crescitelli (Ed.), *Handbook of sensory physiology, VII/5, The visual system in vertebrates.* New York: Springer-Verlag.

Kammen, D. M., & Yuille, A. L. (1988). Spontaneous symmetry-breaking energy functions and the emergence of orientation selective cortical cells. *Biological Cybernetics*, **59**, 23–31.

Kohonen, T. (1984). *Self-organization and associative memory.* Berlin: Springer-Verlag.

LeGall, D. (1991). MPEG: a video compression standard for multimedia applications. *Communications of the ACM*, **34**, 46–58.

Linsker, R. (1988). Self-organization in a perceptual network. *IEEE Computer*, **21**(3), March 1988, 105–117.

Minsky, M., & Papert, S. (1969). *Perceptrons.* Cambridge, MA: MIT Press.

Pitts, W. H., & McCulloch, W. S. (1947). How we know universals: the perception of auditory and visual forms. *Bulletin of Mathematical Biophysics*, **9**, 127–147.

Rosenfeld, A. (Ed.) (1984). *Multiresolution image processing and analysis.* New York: Springer-Verlag.

Schwartz, E. L. (1980). Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding. *Vision Research*, **20**, 645–669.

Shepard, R. N. (1981). Psychophysical complementarity. In M.

Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization*. Hillsdale, NJ: Erlbaum.

Simoncelli, E. P., Freeman, W. T., Adelson, E. H., & Heeger, D. J. (1992). Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, **IT-38**, 587–607.

Swindale, N. V. (1985). Iso-orientation domains and their relationship with cytochrome oxidase patches. In D. Rose & V. G. Dobson (Eds.), *Models of the visual cortex*. New York: John Wiley.

APPENDIX: PROOF OF CONVERGENCE FOR N POINTS ON THE CIRCLE UNDER UNIFORM ROTATIONS WITH CONSTANT LEARNING COEFFICIENT

After some preliminaries, it is shown in Lemma 1 that from any initial state, the system will approach the origin in the state space—corresponding to regular spacing of the points—as closely as desired provided that the learning coefficient is chosen sufficiently small. Then Lemma 2 shows that if the system starts sufficiently close to the origin, there is a non-zero probability that it will reach the origin in the limit. These results are combined in a theorem which proves that the system will reach the origin in the limit with probability 1, from any initial state, given a suitable choice of learning coefficient.

Preliminaries

Since the rotations applied are uniform on $[0, 2\pi)$, expectations with respect to that density are given simply by

$$E[\nu] = \frac{1}{2\pi}\int_0^{2\pi} \nu(\omega)d\omega. \tag{A.1}$$

Using the conventional notation for inner products and norms, the conditional mean squared norm of the updated state vector $\phi(n+1)$ is given by

$$E[\|\phi(n+1)\|^2|\phi(n)] = \|\phi(n)\|^2 + 2\langle E[\Delta\phi(n+1)|\phi(n)], \phi(n)\rangle + E[\|\Delta\phi(n+1)\|^2|\phi(n)]. \tag{A.2}$$

Bounds on the second (inner product) and third (incremental variance) terms on the right of (A.1) may be established.

Bound on the inner product term

A simple integration gives the inner product term as a function of the intervals γ_i as

$$\langle E[\Delta\phi(n+1)|\phi(n)], \phi(n)\rangle = \frac{\alpha N}{16\pi}\sum_{i=0}^{N-1}\gamma_i(\gamma_{i-1}^2 - 2\gamma_i^2 + \gamma_{i+1}^2) \tag{A.3}$$

which can be rewritten as

$$\langle E[\Delta\phi(n+1)|\phi(n)], \phi(n)\rangle = -\frac{\alpha N}{16\pi}\sum_{i=0}^{N-1}(\gamma_i + \gamma_{i+1})(\gamma_i - \gamma_{i+1})^2 \leqslant 0. \tag{A.4}$$

To proceed, we momentarily replace the N variables γ_i by the related β_i, defined by

$$\gamma_i = \gamma_{\min} + \beta_i(\gamma_{\max} - \gamma_{\min}) \tag{A.5}$$

where

$$\gamma_{\max} = \max_i(\gamma_i), \gamma_{\min} = \min_i(\gamma_i). \tag{A.6}$$

Clearly, $0 \leqslant \beta_i \leqslant 1$, $0 \leqslant i \leqslant N-1$, and at least one β_i is zero and at least one other is unity. Then the inner product term is bounded by

$$\begin{aligned}\langle E[\Delta\phi(n+1)|\phi(n)], \phi(n)\rangle &= -\frac{\alpha N(\gamma_{\max} - \gamma_{\min})^3}{16\pi} \times \sum_{i=0}^{N-1}(\beta_i + \beta_{i+1})(\beta_i - \beta_{i+1})^2 \\ &\quad - \frac{\alpha N\gamma_{\min}(\gamma_{\max} - \gamma_{\min})^2}{8\pi}\sum_{i=0}^{N-1}(\beta_i - \beta_{i+1})^2 \\ &\leqslant -\frac{\alpha N\gamma_{\min}(\gamma_{\max} - \gamma_{\min})^2}{8\pi}\sum_{i=0}^{N-1}\frac{1}{N^2} \\ &= -\frac{\alpha\gamma_{\min}}{8\pi}(\gamma_{\max} - \gamma_{\min})^2.\end{aligned} \tag{A.7}$$

Bound on the incremental variance

The effect on the ith interval γ_i of a rotated point x may be expressed by the *update function* $\mu_i(x)$ *associated with that interval,*

$$\mu_i(x) = \begin{cases} x_{i-1} - x, & \frac{x_{i-2} + x_{i-1}}{2} \leqslant x \leqslant \frac{x_{i-1} + x_i}{2} \\ x - x_i, & \frac{x_{i-1} + x_i}{2} \leqslant x \leqslant \frac{x_i + x_{i+1}}{2} \\ 0, & \text{else.} \end{cases} \tag{A.8}$$

The ith component of the squared increment may then be expressed in terms of the $\mu_i(\cdot)$ and the rotation θ as

$$(\Delta\phi_i)^2 = \alpha^2\sum_{j=0}^{N-1}\sum_{k=0}^{N-1}\mu_i(x_j + \theta)\mu_i(x_k + \theta) \tag{A.9}$$

and the incremental variance is then

$$E[\|\Delta\phi\|^2|\phi] = \frac{\alpha^2}{2\pi}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\sum_{k=0}^{N-1}R_i(x_j - x_k), \tag{A.10}$$

where $R_i(\cdot)$ is a correlation function associated with the ith interval,

$$R_i(x) = \int_0^{2\pi}\mu_i(\nu)\mu_i(\nu + x)d\nu. \tag{A.11}$$

The incremental variance is bounded by noting that

$$R_i(x) \leqslant R_i(0) = \frac{\gamma_{i-1}^3 + 2\gamma_i^3 + \gamma_{i+1}^3}{24} \tag{A.12}$$

and so for any ϕ,

$$E[\|\Delta\phi\|^2|\phi] \leqslant \frac{\alpha^2}{2\pi}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\sum_{k=0}^{N-1}R_i(0) = \frac{\alpha^2N^2}{12\pi}\sum_{i=0}^{N-1}\gamma_i^3 \leqslant \frac{\alpha^2N^3\gamma_{\max}^3}{12\pi}. \tag{A.13}$$

With these bounds established, the results follow:

LEMMA 1. *Let $\phi(0)$ be the arbitrary initial state of the network. Then for any $\varepsilon > 0$, there exists an $\alpha > 0$ and $n < +\infty$ such that with probability* 1, $\|\phi(n)\| \leqslant \varepsilon$.

Proof. For any ϕ, the incremental variance is bounded by (A.13). Now let $\|\phi\| > \varepsilon/2$. Since

$$\gamma_{\max} - \gamma_{\min} \geqslant \frac{2}{\sqrt{N}}\|\phi\| \tag{A.14}$$

it follows from (A.7) that the inner product term is bounded above by

$$\langle E[\Delta\phi(n+1)|\phi(n)], \phi(n)\rangle \leqslant -\frac{\alpha\gamma_{\min}\varepsilon^2}{8\pi N}. \tag{A.15}$$

It then follows from (A.2), (A.13) and (A.15) that choosing

$$\alpha < \frac{3\gamma_{\min}\varepsilon^2}{N^4\gamma_{\max}^3}$$

implies that for $\|\phi\| > \varepsilon/2$,

$$E[\|\phi(n+1)\|^2|\phi(n)] < \|\phi(n)\|^2. \tag{A.16}$$

Now let n_0 be the smallest integer n for which $\|\phi(n)\| \leqslant \varepsilon/2$ and $n_0 = +\infty$ if no such integer exists. Define the process $\{\check{\phi}(n)\}$ as $\{\phi(n)\}$ with stopping, i.e.,

$$\check{\phi}(n) = \begin{cases} \phi(n), & n \leqslant n_0 \\ \phi(n_0), & n > n_0. \end{cases} \tag{A.17}$$

Then $\{\|\check{\phi}(n)\|\}$ is a supermartingale, and by virtue of (A.16),

$$E\left[\|\check{\phi}(n+1)\|^2|\check{\phi}(n)\right] < \|\check{\phi}(n)\|^2,\ n \leqslant n_0. \tag{A.18}$$

Then by the supermartingale convergence theorem (Doob, 1990, Theorem 4.1s, p. 324, $\{\|\check{\phi}(n)\|\}$ has a limit d_∞ almost everywhere, and as a consequence of (A.16),

$$d_\infty \leqslant \varepsilon/2. \tag{A.19}$$

But then there must exist some finite $m \leqslant n_0$ for which

$$\|\check{\phi}(m)\| - d_\infty < \varepsilon/2 \tag{A.20}$$

i.e., such that

$$\|\check{\phi}(m)\| \leqslant \varepsilon. \tag{A.21}$$

In other words, with probability 1, there exists an $m < +\infty$ for which

$$\|\phi(m)\| < \epsilon \tag{A.22}$$

as was to be proved. □

LEMMA 2. *If*

$$\|\phi(0)\| < \frac{\pi\rho}{N^2\sqrt{2}}$$

with ρ given by

$$\rho^2 = \alpha^2 + (1-\alpha)^2 + 2\alpha(1-\alpha)\cos\frac{2\pi}{N}$$

then it follows that Prob $\{\lim_{n\to\infty}\phi(n) = \mathbf{0}\} > 0$.

Proof. Let θ be the rotation applied at iteration $n+1$, and suppose that there exists a j such that for every point x_i,

$$\frac{x_{i-1}+x_i}{2} \leqslant x_{i-j} + \theta < \frac{x_i + x_{i+1}}{2}. \tag{A.23}$$

In other words, each point x_{i-j} lands in a different "bin" under the rotation θ, and affects only the corresponding point x_i. The mapping between rotated and unrotated points is *bijective*, or one-to-one. It is easy to show that for bijections, $\|\phi\|$ is unchanged or reduced:

$$\|\phi(n+1)\|^2 = \sum_{i=0}^{N-1}\left[(1-\alpha)\phi_i(n) + \alpha\phi_{i-j}(n)\right]^2 \leqslant \|\phi(n)\|^2. \tag{A.24}$$

Hence $\|\phi\|$ can increase only if the mapping between rotated and unrotated points is *injective*, or many-to-one. (We will assume for mathematical convenience that the converse is also true—*all* injections increase $\|\phi\|$, and *only* bijections can leave $\|\phi\|$ unchanged or reduce it—and we will prove that the system converges even in this worst case.) Let $P_+(n)$ be the probability that $\|\phi\|$ increases,

$$P_+(n) = \text{Prob}\{\|\phi(n+1)\| > \|\phi(n)\|\|\phi(n)\}. \tag{A.25}$$

$P_+(n)$ is bounded by observing that it cannot exceed the probability of an injection, or equivalently the probability that at least one point receives a zero update,

$$P_+(n) \leqslant \frac{1}{2\pi}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\max\left(0, \gamma_j - \frac{\gamma_i + \gamma_{i+1}}{2}\right) \leqslant \frac{N^2}{2\pi}(\gamma_{\max} - \gamma_{\min}). \tag{A.26}$$

Now since $\gamma_{\max} - \gamma_{\min} \leqslant \sqrt{2}\|\phi\|$, it follows that the probability that $\|\phi\|$ increases is bounded by

$$P_+(n) < \frac{N^2}{\pi\sqrt{2}}\|\phi(n)\|. \tag{A.27}$$

It follows from (A.24) that $\|\phi\|$ is unchanged if and only if $\phi_i(n) = \phi_{i-j}(n)$, $0 \leqslant i \leqslant N-1$. Since this implies that $j = 0$ or that $\phi(n)$ is periodic, the probability $P_0(n)$ with which $\|\phi\|$ is unchanged satisfies the relation

$$P_0(n) = \text{Prob}\{\|\phi(n+1)\| = \|\phi(n)\|\|\phi(n)\} < \frac{1}{2},\ \|\phi(n)\| > 0. \tag{A.28}$$

All other bijections yield a reduction in $\|\phi\|$, and a simple variational argument shows that in this event,

$$\begin{aligned}\|\phi(n+1)\|^2 &\leqslant \left[\alpha^2 + (1-\alpha)^2 + 2\alpha(1-\alpha)\cos\frac{2\pi}{N}\right]\|\phi(n)\|^2 \\ &= \rho^2\|\phi(n)\|^2,\ \rho < 1.\end{aligned} \tag{A.29}$$

Hence if $\|\phi\|$ is reduced at all, it is reduced by at least a factor of ρ, and the probability of a subsequent increase is itself reduced in accordance with (A.27). Now we introduce a new index set

$\{m_k, k = 0, 1, 2, \ldots; m_k > m_{k-1}, m_0 = 0\}$ which indexes only those iterations at which $\|\phi\|$ changes:

$$\{m_k\} = \{n : \|\phi(n)\| \neq \|\phi(n-1)\|,\ n > 0\}. \tag{A.30}$$

The probability that the change in $\|\phi\|$ at iteration m_{k+1} is a reduction is given by

$$\text{Prob}\{\|\phi(m_{k+1})\| < \|\phi(m_k)\|\} = (1 - P_+(m_k) - P_0(m_k)) \sum_{i=0}^{\infty} (P_0(m_k))^i$$

$$= 1 - \frac{P_+(m_k)}{1 - P_0(m_k)}. \tag{A.31}$$

Then the probability that *all* changes in $\|\phi\|$ are reductions is given by

$$P = \prod_{k=1}^{\infty} \left(1 - \frac{P_+(m_k)}{1 - P_0(m_k)}\right). \tag{A.32}$$

From (A.27), choosing sufficiently small

$$\|\phi(0)\| < \frac{\pi\rho}{N^2\sqrt{2}}$$

gives $P_+(0) < \rho/2$, and it then follows from (A.28) that the product on the right of (A.32) converges, giving

$$P \geqslant \prod_{k=1}^{\infty} (1 - \rho^k) \geqslant \prod_{k=1}^{\infty} \exp(-2\rho^k) = \exp\left(-\frac{2\rho}{1-\rho}\right) > 0. \tag{A.33}$$

There is therefore a probability $P > 0$ that the sequence $\|\phi(n)\|$ is monotonic non-increasing and therefore possesses a limit d_∞, say. Let $\phi(n)$ be the limit state. Then it follows from Lemma 1 that there exists some $\alpha > 0$ for which $E[\|\phi(n+1)\| | \phi(n)] < d_\infty$ unless $\gamma_{\max} = \gamma_{\min}$, i.e., unless $\phi = 0$. Hence

$$\phi(n) \xrightarrow{n \to \infty} \mathbf{0}$$

and

$$Prob\left\{\lim_{n \to \infty} \phi(n) = \mathbf{0}\right\} = P > 0 \tag{A.34}$$

as was to be proved. □

Combining Lemmas 1 and 2, the theorem follows:

THEOREM 2. *There is a learning coefficient α for which, for any initial state $\phi(0)$, the network will converge with probability* 1.

Proof. Choosing α such that by Lemma 1 there is an $n < +\infty$ for which

$$\|\phi(n)\| \leqslant \frac{\pi\rho}{N^2\sqrt{2}} = \lambda_N \tag{A.35}$$

guarantees by Lemma 2 that there is a probability $P > 0$ that the network will converge from that state. Now define the ith transition time n_i as follows: n_1 is the first n for which $\|\phi(n)\| < \lambda_N$, and n_i, $i > 1$, is the smallest n for which both (i) $\|\phi(n)\| < \lambda_N$, and (ii) $\|\phi(m)\| \geqslant \lambda_N$ for some $m : n_{i-1} < m \leqslant n_i$. Call C_i the event that the network converges from $\phi(n)$,

$$C_i = \{\|\phi(n)\| \leqslant \|\phi(n-1)\|,\ n > n_i\}. \tag{A.36}$$

Now C_i depends only on $\phi(n_i)$ and not on $\phi(n)$, $n < n_i$. From Lemma 2, $\text{Prob}\{C_i\} = P$, and it follows that the probability that the network does not converge is just

$$\text{Prob}\{\bar{C}\} = \lim_{n \to \infty} (1 - \text{Prob}\{C_i\})^n = 0. \tag{A.37}$$

Thus the network converges almost surely, regardless of $\phi(0)$, and by Lemma 2 it must converge to $\phi(\infty) = \mathbf{0}$. □

Note. The bounds on α given here should be taken with a pinch of salt: in practice, convergence is consistently observed for $0 < \alpha \leqslant 0.9$ for moderate numbers of points (of the order of $N = 10$). For larger N one still observes convergence within reasonable time even for quite large values of α. For example, for $N = 64$ points, convergence reliably occurred within one million iterations (and often considerably fewer) over all values of α tested within the range $0.01 \leqslant \alpha \leqslant 0.3$.