Papers

International Journal of Bifurcation and Chaos, Vol. 8, No. 7 (1998) 1463–1478

ROUTES TO CHAOS IN NEURAL NETWORKS WITH RANDOM WEIGHTS

D. J. ALBERS and J. C. SPROTT
Department of Physics, University of Wisconsin, Madison, 1150 University Avenue, Madison, WI 53706, USA

W. D. DECHERT
Department of Economics, University of Houston, Houston TX 77204-5882, USA

Received January 13, 1998; Revised May 9, 1998

Neural networks are dense in the space of dynamical systems. We present a Monte Carlo study of the dynamic properties along the route to chaos over random dynamical system function space by randomly sampling the neural network function space. Our results show that as the dimension of the system (the number of dynamical variables) is increased, the probability of chaos approaches unity. We present theoretical and numerical results which show that as the dimension is increased, the quasiperiodic route to chaos is the dominant route. We also qualitatively analyze the dynamics along the route.

1. Introduction

The study of complex systems over the past few decades has led to many interesting and diverse results. The dynamics encountered in many different systems occurs across diverse disciplines. This apparent similarity in dynamics has motivated our Monte Carlo study of dynamical systems using neural networks, since neural networks are dense in the set of continuous functions on a bounded interval. We investigate the general dynamic properties, including the probability of chaos, and the power spectrum. As the dimension is increased, the probability of chaos approaches unity. With this in mind, we study the route to chaos as the connection strengths are increased. Numerically we find that as the dimension is increased the probability of the first bifurcation being Hopf increases to near unity. Based on observing the bifurcation diagrams of hundreds of networks, we conject that the quasiperiodic route dominates at high dimension. Besides the evidence and conjectures put forth by Doyon *et al.*, [1993] and Brock [1997] regarding very general systems, there is evidence that some specialized systems such as coupled map lattices and coupled identical period-doubling systems also tend to, or do undergo the quasiperiodic route to chaos [Wang & Cerdeira, 1996; Reick & Mosekilde, 1995].

2. Bifurcation Theory and Chaotic Dynamics

Define X to be an open subset of a Banach space, E. Next set $I = [a, b] \in R$. Now define the map $f_\mu : X \mapsto E$ for $\mu \in I$, thus forming a parameterized family of maps for which μ is the bifurcation parameter. Define $\mu_0 \in I$ such that f_{μ_0} has a fixed point at $x_0 \in X$. Define Γ such that $\Gamma = \{(\mu, x) \in I \times X : f_\mu x = x\}$. Now, suppose that $(\mu_0, x_0) \in \Gamma$, where x_0 is not hyperbolic. A bifurcation occurs where $D_{x_0} f_{\mu_0}$ has a eigenvalue on the unit circle. There are three generic ways this can occur, (and hence three generic bifurcations): Flip, where a negative real eigenvalue lies on the unit circle ($D_{x_0} f_{\mu_0} = -1$), (this corresponds to a period-two oscillation); saddle-node, where

a positive real eigenvalue lies on the unit circle ($D_{x_0} f_{\mu_0} = 1$), (this corresponds to the appearance of two branches of stable equilibria); and Hopf,[1] where a complex conjugate pair of eigenvalues lie on the unit circle ($D_{x_0} f_{\mu_0} = |a \pm bi| = 1$, $b \neq 0$), (this corresponds to the appearance of a limit cycle or torus). The three types of bifurcations described above are generic when Γ is a nonsingular smooth curve. We have yet to impose any symmetry groups, but it should be noted that the presence of symmetry groups changes which bifurcations are generic [Ruelle, 1989].

A "route to chaos" is the path of bifurcations that a system undergoes from a steady state to a chaotic state as a control parameter is varied. Since we are only concerned with the generic bifurcations, all our routes to chaos must follow combinations of them. The main theme of this study is determining the most likely first bifurcation along this route for a random system. We also concern ourselves with the general dynamics along the route, in the chaotic regime, and then the transition out of chaos at very large values of the control parameter. For our purposes, we will define a system to be chaotic if its largest Lyapunov exponent is positive.

3. General Neural Networks

Single layer feed-forward neural networks of the form

$$f(y) = \sum_{i=1}^{n} \beta_i \phi \left(s\omega_{i0} + s \sum_{j=1}^{d} \omega_{ij} y_j \right) \tag{1}$$

where $f : R^d \to R$, with arbitrary squashing functions ϕ, can uniformly approximate any continuous function on any compact set, and any measurable function arbitrarily well, given a sufficient number of hidden units [Hornik *et al.*, 1989]. In Eq. (1), n represents the number of hidden units or neurons, d is the embedding dimension of the system which for our purposes is the number of time lags, and s is a scaling factor on the weights. The function ϕ represents a neuron or activation function. If $\phi \in S_p^m(R, \mu)$, (i.e. is made up of functions in $C^m(U)$ having derivatives up to order m and $L_p(U, \mu)$-integrable), and does not vanish everywhere, then Eq. (1) can approximate any function belonging to $C^\infty(R^r)$ and its derivatives up to order m arbitrarily well on compact sets [Hornik *et al.*, 1990]. In general the parameters are set in the following way:

$$\beta_i,\ w_{ij},\ y_j,\ s \in R \tag{2}$$

where the β_i's and w_{ij}'s are elements of weight matrices (which we hold fixed for each case), $(y_0, y_1, \ldots, y_d)$ represent initial conditions, and $(y_t, y_{t+1}, \ldots, y_{t+d})$ represent the current state of the system at time t. For our purposes we shall assume that the functions are sufficiently smooth such that the system dynamics are representable by an element of $C^2(M, M)$, the set of twice continuously differentiable functions from a compact manifold, M, of dimension m into itself. These dynamical systems are of the form

$$x_{t+1} = F(x_t) \tag{3}$$

where $F : M \to M$ and $x_t \in M$. For our purposes, we use a single neural network to generate a "time-series" of scalar data. A neural network forms a dynamical system on R^d by:

$$y_t = f(y_{t-d}, y_{t-d+1}, \ldots, y_{t-1}) \tag{4}$$

where $y_t \in R$. Systems of the form Eq. (4) are equivalent to systems of form of Eq. (3) since Eq. (3) can be stated via a mapping of R^d to itself:

$$(y_1, y_2, \ldots, y_d) \to (y_2, y_3, \ldots, f(y_1, y_2, \ldots, y_d)) . \tag{5}$$

Thus they form a subset of the d-dimensional dynamical systems. Takens [1980] has shown that systems of the form of Eq. (3) are diffeomorphisms that embed (generically) in R^d for some $d \leq 2m+1$. Thus, there is an open and dense set of dynamical systems, each element of which is topologically conjugate to a system of the form of Eq. (4). These latter systems can be uniformly approximated (on compacta) by neural networks.

The significance of uniform approximation to our work is considerable. In this study we are considering a function space and then attempting to gain insight into the physical world based on the dynamics of that function space. There are two main considerations; whether functions from

[1]Hopf proved the bifurcation theorem for vector fields; the "Hopf bifurcation for maps" was proved independently by Naimark [1959] and Sacker [1965]. To avoid confusion we will refer to the Naimark–Sacker bifurcation as Hopf.

our class mimic any dynamical system, and whether our method of sampling is representive of that class of dynamical systems. The answer to the former is the uniform approximation put forth by Hornik *et al.* [1989]. A discussion of the latter follows.

3.1. *Our networks*

For the purpose of our study we consider networks of the form:

$$y_t = \sum_{i=1}^{n} \beta_i \tanh\left(s\omega_{i0} + s\sum_{j=1}^{d} \omega_{ij} y_{t-j}\right) \qquad (6)$$

which is identical to Eq. (1) with the hyperbolic tangent taken as the squashing function, which, belonging to $S_p^m(R, \mu)$, is general. As previously stated, the β and w matrices are held fixed; s is held fixed for the probability of chaos study and also used as the bifurcation parameter. We pick the β's *iid* uniform over [0, 1], and then re-scale them to satisfy the following condition:

$$\sum_{i=1}^{n} \beta_i^2 = n \qquad (7)$$

The w_{ij}'s are picked *iid* normal with zero mean and unit variance. The s parameter is a real number, and it can be interpreted as the standard deviation of the w matrix of weights. The initial y_j's are chosen *iid* uniform on the interval $[-1, 1]$. All the weights and initial conditions are selected randomly using a pseudorandom number generator [l'Écuyer, 1988; Press *et al.*, 1992]. Iterating Eq. (6) gives a time-series whose dynamics we investigate.

The specific conditions used to pick the weights has an important effect on the dynamics. First, $\tanh(x)$, for $|x| \gg 1$ will tend to behave much like a binary function. Since binary functions have a finite number of states and must repeat, such systems cannot be chaotic. Therefore, if β or s become very large, the system will have a greatly reduced ability to be chaotic. There is a simple reason for the imposed condition on the β's as opposed to something like:

$$\sum_{i=1}^{n} |\beta_i| = k \qquad (8)$$

where k is a fixed constant. If the β_i are restricted to a sphere of radius k, as n is increased, $\langle {\beta_i}^2 \rangle$ goes to zero [Albers *et al.*, 1996]. Also, since the $\tanh(x)$ function is nearly linear when $|x| \ll 1$, choosing s to be small will force the dynamics to be mostly linear, again inhibiting chaos. As d is increased, the effect of any given w_{ij} decreases, making high-d networks with and without a bias term, w_{i0}, very similar.

3.2. *How we sample the space*

The generality of the results presented hinges on how we sample the space of neural networks and how that is related to the space of dynamical systems. First we must deal with how we are sampling the neural network function space in accordance with uniform approximation. Since we are dealing with a finite sample, the set we are studying is a set of measure zero of functions. What is more important is the method used to sample the space. Within the class of neural networks with the β's scaled as in our study with the w_{ij}'s such that $w_{ij} \in (-12, 12)$, (because of the approximation used to generate normally distributed weights), we produce a countable dense subset of the neural networks. A harder question is whether our sampling method gives a dense subset of a larger class of functions, (specifically dynamical systems). To answer this we would need to establish a specific norm on this larger class of functions. This is difficult. After establishing this norm, we would need to show that our neural networks are dense in that norm, and then that our sampling method is dense in our class of neural networks. The latter is not a problem; the previous two are. For this reason, the results at this point are suggestive of what dynamical system space is like. We demonstrate phenomena that are possible, and show just how complex things can get with a limited number of dimensions. None of our results can be generalized to the entire space of dynamical systems since we do not know what part of the dynamical system function space we are sampling. In this paper we use words like "typical" or "on average" with respect to the space of neural networks we are considering.

3.3. *Networks without bias terms*

Doyon *et al.* [1993] considered networks with dimensions starting near the upper range of our study ($d = 1024$). All the networks they considered were similar to the ones we studied but with $w_{i0} = 0$. Although their networks and our networks are different, they are in many ways dynamically equivalent. Excluding the bias term decreases the generality of the uniform approximation, but not the dynamics. Consider the following example; in

Eq. (6), $w_{i0} = 0\ \forall i$. By doing this, Eq. (6) can only uniformly approximate odd functions. If we map the compact set $I \subset R$ from $[a, b] \in I$ to $[0, \infty)$, the dynamics are preserved, even though the approximating is lost. Since networks of the form of Eq. (6) with and without bias terms are equivalent over $[0, \infty)$, the possible dynamics must also be equivalent. There is a subtle difference, however, in the genericity of bifurcations. Networks with no bias terms impose an odd symmetry group, making the pitchfork and not the saddle-node generic.

3.4. *Convergence to attractors*

Before presenting the numerical results, we discuss briefly numerical errors and how they might affect the results. Since chaotic attractors have a sensitive dependence on initial conditions, and we only keep d time lags, the exact original initial conditions are lost in d iterations of the map, and all that are left are points near the attractor within roundoff error. The shadowing lemma [Bowen, 1978; Newhouse, 1980] helps ensure that the orbit remains close to the attractor. We will give a version stated in [Guckenheimer & Holmes, 1983]:

Shadowing Lemma: Let Λ be a hyperbolic invariant set. Then for every $\beta > 0$, there is an $\alpha > 0$ such that every α-pseudo-orbit $x_j; b_{i=a}$ in Λ is β-shadowed by a point $y \in \Lambda$.

The existence of this lemma is encouraging, but in practice it is often very difficult to show a system to be hyperbolic; and many systems are not hyperbolic. In order to deal with this, Grebogi *et al.* [1990] suggest a method of containment guaranteeing a pseudo-orbit for nonhyperbolic sets. This method consists of constructing parallelograms that are oriented such that two sides are contracting and two are expanding. There are inherent problems with this method; namely, when an angle of the parallelogram is near zero the parallelogram effectively loses a dimension. These situations are rare, and this method works rather well. We do not use this method since it is numerically costly for the benefit gained, as shown below.

The first bifurcation of all the systems we consider is global. For brief substantiation, consider the following argument. Map our networks to the origin up to the first bifurcation. This can be done without affecting the dynamics. We are now considering the first bifurcation where the origin loses its stability. The time-dependent terms of the Jacobian of a system following the origin are:

$$a_i = \sum_{i=1}^{n} \beta_i s w_{ij}\ \mathrm{sech}^2 \left(s \sum_{j=1}^{d} \omega_{ij} y_j \right) \tag{9}$$

which simplifies to

$$a_i = \sum_{i=1}^{n} \beta_i s w_{ij} \tag{10}$$

when $y_j = 0$. Notice that the dependence on initial conditions is gone and the eigenvalues are simply a function of s. To verify that this works numerically, we tested the results of the first bifurcation both for cases where the bias terms were set to zero and for those which were not. The resulting data were the same within statistical error.

Beyond the first bifurcation this simple transformation cannot apply. However, beyond the first bifurcation we are only concerned with the probability of chaos and qualitative analysis of the dynamics. When the system is structurally stable, perturbations do not affect the dynamics, so the only place at which the qualitative dynamics are affected is the chaotic region. Since we are not trying to mimic any particular system, staying on the attractor is not as critical, albeit we have considerable experimental evidence that this is not a problem. We studied several specific cases over a variety of initial conditions, after various numbers of iterations. Correlation dimensions and largest Lyapunov exponent plots would overlay within experiment error, and the bifurcation diagrams, if they did not directly overlay, would dynamically overlay. (i.e. They were periodic and chaotic in the same places with the same features.) There are several cases where this matter is of concern; they will be considered in the following sections.

4. Numerical Results

4.1. *Probability of chaos*

The probability of chaos is the fraction of systems with positive largest Lyapunov exponent. We choose the parameters n, d, and s between 1 and 256. Weights and initial conditions are as previously described. We calculate the largest Lyapunov exponent by the following method [Wolf *et al.*, 1984]. First we randomly select an initial point $(y_0, \ldots, y_{d-1})$ and a nearby point

$(v_0, \ldots, v_{d-1})$ with a small separation ϵ at time t. Define

$$\Delta y_t = (y_t - v_t, \ldots, y_{t+d-1} - v_{t+d-1}) \tag{11}$$

and let $|\Delta y_0| = \epsilon$. Both points are advanced one time step, and the ratio

$$\frac{|\Delta y_1|}{|\Delta y_0|} \tag{12}$$

is recorded. The vector Δy_1 is then rescaled to a length ϵ, and the new neighbor,

$$(v_1, \ldots, v_d) = (y_1, \ldots, y_d) + \Delta y_1 \tag{13}$$

is advanced one time period along with $(y_1, \ldots, y_d)$. This process is repeated, and the largest Lyapunov exponent is estimated from the average of the logarithm of the scalings:

$$\lambda = t^{-1} \sum_{l=0}^{t-1} \ln \frac{|\Delta y_{l+1}|}{|\Delta y_l|} \tag{14}$$

where t is the number of iterations of the map over which the average is taken. We iterate the map a set number of times (usually 100 000) before beginning this calculation to help ensure that the orbit is near the attractor.

Figure 1 shows the percentage of chaos in networks for varying n and d with s fixed at 8. Notice that as d is increased, the probability of chaos approaches unity; the same is true of n. The d dependence does not change with the method used to pick the weight matrices; the n dependence does. The relatively straight contour lines are related to the fact that we scale the β's with n. In previous work [Albers *et al.*, 1996] we assigned the β_i's such that $k = 1$ in Eq. (8). The formulation in Eq. (8) gives "C"-shaped contour lines that depend on n. As stated in Sec. 3.1, as n is increased, the individual β_i's go to zero, forcing the system to be linear at high n. The important feature of Fig. 1 is that as the available complexity (d) increases the probability of chaos will always be high, unless the system is binary or linear. The dip in the first ten percent line in Fig. 1 is a numerical artifact related to how the plotting program positions the lines.

Figure 2 shows the percentage of chaotic networks for varying s and d with n fixed at 8. Notice the "C"-shaped contour lines; this would suggest that the s parameter can be optimized for

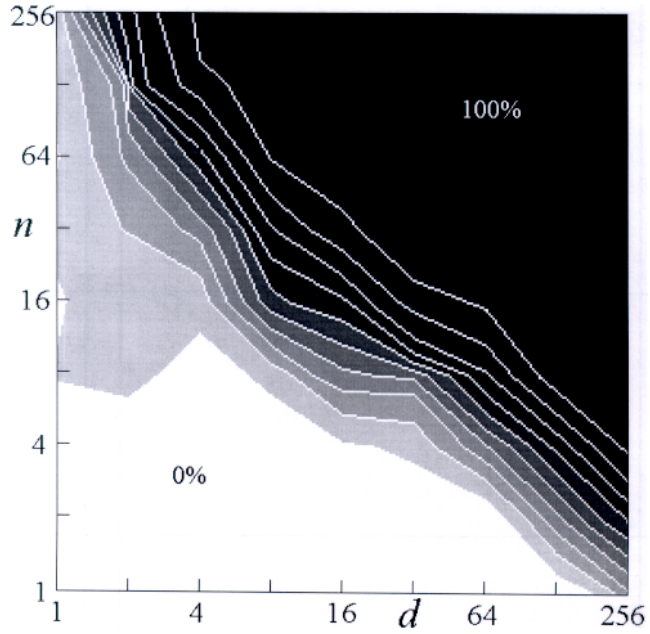

Fig. 1. Probability of chaos contour plot for $s = 8$ and various d and n.

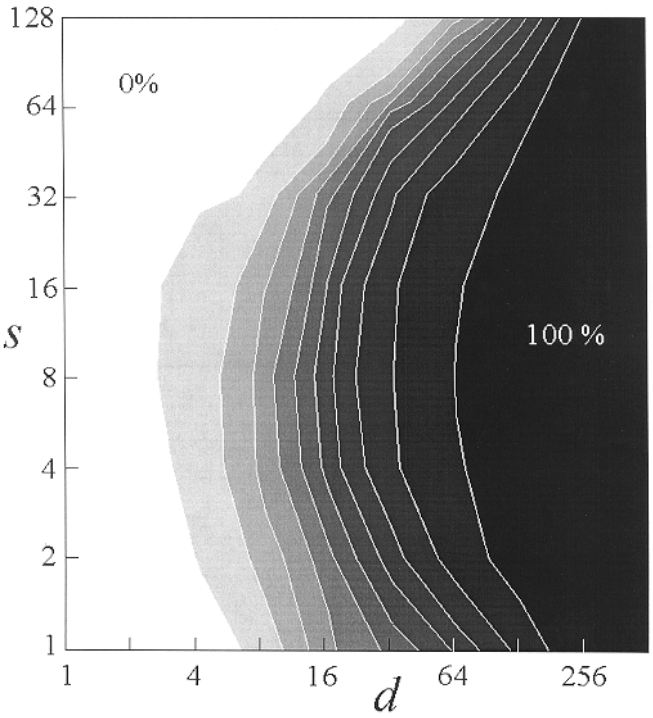

Fig. 2. Probability of chaos contour plot for $n = 8$ and various d and s.

maximum chaos. The s parameter works with the β_i's to place the argument of the squashing function in different regions of the domain. For the hyperbolic tangent squashing function, the value of x has a significant effect on the available dynamics. As previously stated, the hyperbolic tangent

function is linear when $|x| \ll 1$ and binary when $|x| \gg 1$. Thus if the β_i and s put the argument into either region, neither chaos nor limit cycles are possible.

4.2. *The power spectrum*

An interesting issue is the global topology of the strange attractor that results from the chaotic high-n, high-d systems and the power spectrum of the associated dynamics. For this purpose a time series of 32 000 points was calculated for over a dozen systems with $n = 64$, $d = 512$, and $s = 8$ after 5000 iterations to allow any initial transient to decay. The power spectra typically have one or more dominant incommensurate peaks and an approximately white (frequency-independent) background three or four orders of magnitude below that of the dominant frequency. Thus the global topology of the attractor resembles a limit cycle or higher dimensional torus, perturbed (sometimes strongly) by chaotic deviations. This behavior is confirmed in correlation dimension plots that show low, presumably integer, dimension on large scales, and an unmeasurably high dimension on smaller scales of the attractor.

4.3. *The basin of attraction and initial conditions*

The method used to pick initial conditions clearly influences how well the results generalize. We have determined through experiment that for one set of parameter values, using different initial conditions can usually produce several attractors. In preliminary experiments, as the dimension is increased, so are the number of attractors (we have seen as many as 9 at $d = 64$). When s is small, say 0.0125, there is only one attractor. As s is increased, the number of attractors increases, until the squashing function begins to saturate. As the squashing function approaches saturation the number of attractors decreases until there is again one attractor.

Different methods of picking initial conditions have different effects on the bifurcation diagrams, but the first point of instability (i.e. the point we call the first bifurcation) is a global bifurcation, which is independent of initial conditions. Before this bifurcation, there is only one attractor; a fixed point. After this bifurcation the dynamics are most often not global.

Figure 3 shows an interesting example of how the structure of the basin can affect the dynamics. As we increase the s parameter, not only does the system bifurcate, but it is "riding the fence" along a basin boundary between two attractors. As we increase s, since we are re-initializing with the same initial values, the system moves from one basin into another. There is a difference between "jumping" basins and bifurcation. A bifurcation is a qualitative change in dynamics when a control parameter is varied. What we call basin "jumping" is not a function of a control parameter. Rather this "jumping" occurs because of re-initializing the

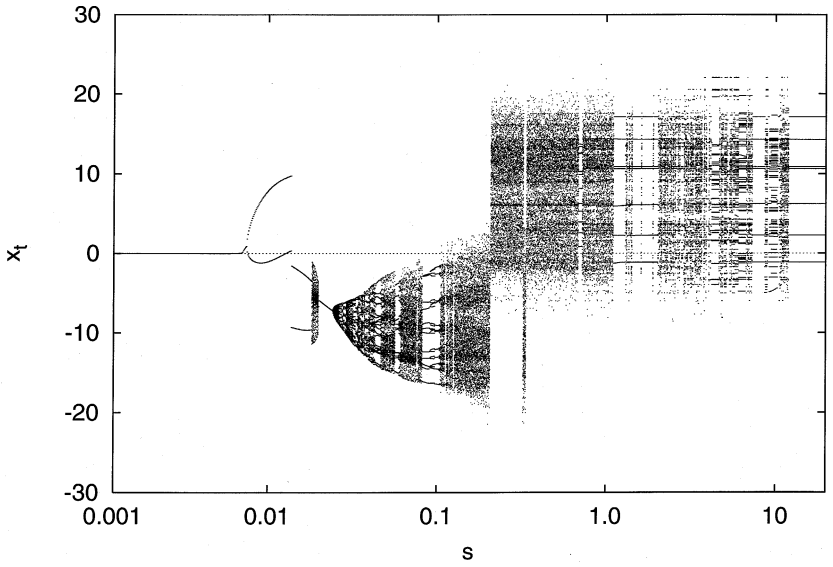

Fig. 3(a). Bifurcation diagram for $n = 64$ and $d = 4$.

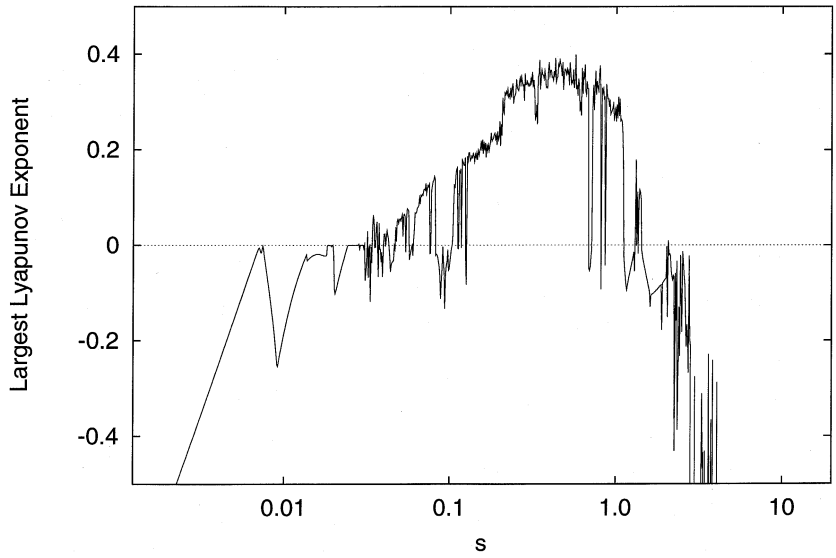

Fig. 3(b). The Lyapunov exponent for $n = 64$ and $d = 4$.

network in another basin, or structural instability. Figure 3(a) suggests three things: that most bifurcations are not global, that the positioning of the initial conditions in the basin is very important, and that understanding the underlying basin structure is crucial if one wants to know how a system is capable of behaving. For each value of s, we use the same initial conditions as for previous values of s. If we were to choose new initial conditions for each s, we would get a larger sampling of attractors. This larger sampling tends to make the diagrams look as though we had plotted several attractors at once.

4.4. *The first bifurcation*

The first bifurcation of a system occurs where its largest eigenvalue reaches the unit circle as some parameter is varied. We choose to vary the s parameter since it acts like a gain on the weights, taking the hyperbolic tangent from its linear range through the nonlinear range and into its binary range. For each case, we pick and fix the weights and initial conditions, run the case for an s value, calculate the eigenvalues and largest Lyapunov exponent, and then increase s by a constant multiple (usually close to one), re-initialize with the same weights and initial conditions, and repeat the process. When the modulus of the largest eigenvalue reaches the unit circle we decide what kind of bifurcation has occurred and move on to the next set of weights and initial conditions. This process was done over n and d values ranging from 1 to 256.

To calculate the eigenvalues for a given system, we create the Jacobian matrix:

$$\begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_{d-2} & a_{d-1} & a_d \\ 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & & & & \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \tag{15}$$

where:

$$a_i = \sum_{i=1}^{n} \beta_i s w_{i0} \operatorname{sech}^2 \left(s\omega_{i0} + s \sum_{j=1}^{d} \omega_{ij} y_j \right) \tag{16}$$

The eigenvalues of the above matrix are the eigenvalues for the system. We begin each system with an s value small enough such that the dynamics are that of a stable fixed point. Then, as we increase s, we push the system into the more nonlinear region of the squashing function, eventually saturating it into a binary function.

Figures 3(a)–6(a) are bifurcation diagrams for various n and d. In these figures the first 120 000 y_t values are discarded and the next 128 are plotted. Each is a typical bifurcation diagram for its given parameter values. Note that the meaning of typical or characteristic for one set of n and d values is different from that of another set of n and d values. There is more apparent variability in the dynamics at low d than at high d. At high d most of the diagrams look alike. At low d the diagrams are erratic but differ in detail.

Figures 3(b)–6(b) show the largest Lyapunov exponents corresponding to the Figs. 3(a)–6(a). Note that as n is increased the plots become smoother. From this plot alone it is not possible to tell saddle-node from flip bifurcations or second Hopf bifurcations from first Hopf. Chaos, however, can be differentiated from limit cycles and periodic orbits. The positive values indicate that the system is chaotic. By looking at the corresponding (a) and (b) figures, flips and saddle-nodes can be differentiated; limit-cycles/tori and chaos can be differentiated, and often second and third Hopfs can be differentiated.

Figure 4(a) is a bifurcation diagram for $n = 4$ and $d = 4$. In this figure the first bifurcation happens to be a Hopf. Figure 3(a) is a bifurcation diagram for $n = 64$, $d = 4$, and shows a typical saddle-node first bifurcation. Figure 5(a), $n = 4$, $d = 64$, shows a Hopf first bifurcation, as does Fig. 6(a), ($n = 64$, $d = 64$). We will return to the dynamics after the first bifurcation later.

Notice that as n is increased, the range of y_t values increases. This is due to the method used to choose β's. In general, as the dimension increases, the bifurcation diagrams are more centered and symmetric about the origin, almost as if there

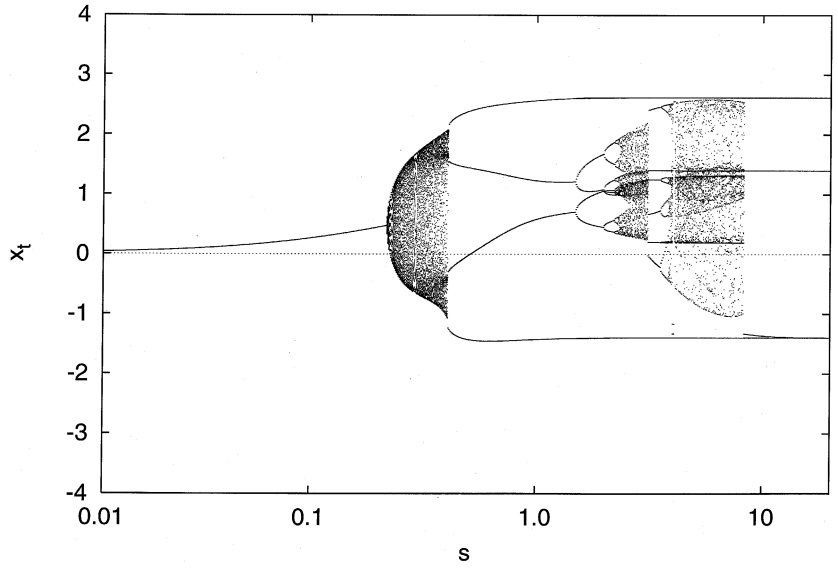

Fig. 4(a). Bifurcation diagram for $n = 4$ and $d = 4$.

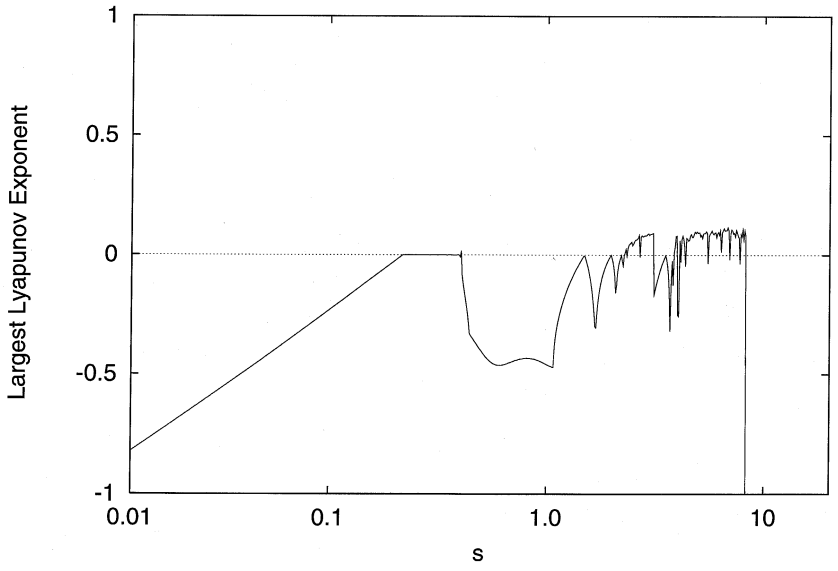

Fig. 4(b). The Lyapunov exponent for $n = 4$ and $d = 4$.

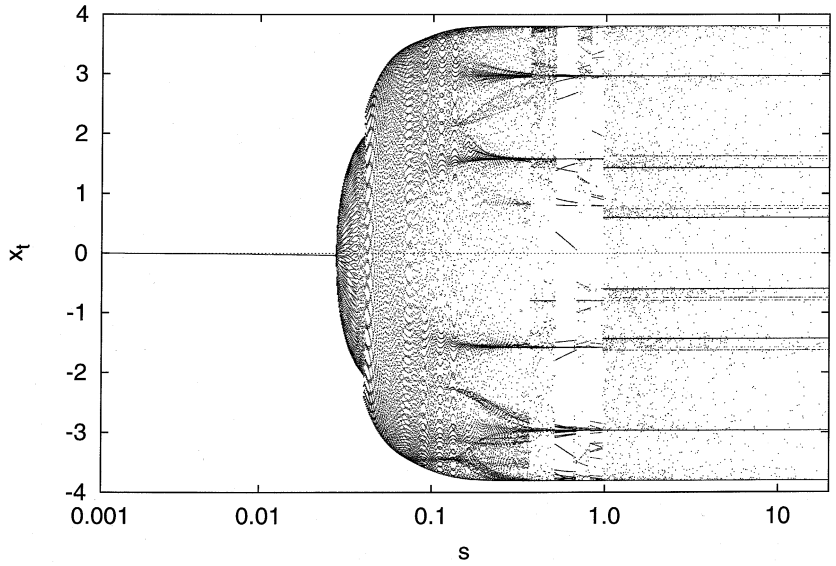

Fig. 5(a). Bifurcation diagram for $n = 4$ and $d = 64$.

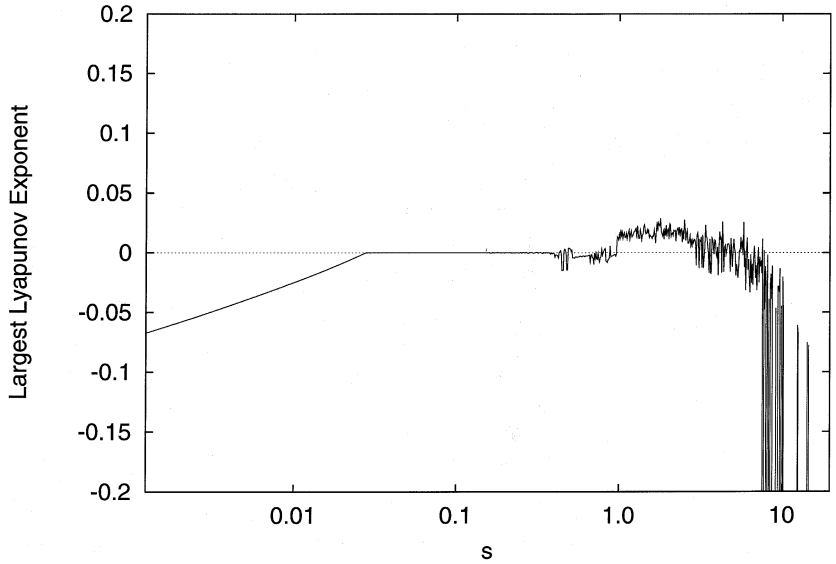

Fig. 5(b). Lyapunov exponent for $n = 4$ and $d = 64$.

were no bias term. This is because, as the number of w_{ij}'s is increased, the importance of the individual w_{ij} is decreased, thus the importance of each bias term is decreased. This does not affect the results since there is no dynamical difference between networks with and without bias terms (as stated in previous sections).

4.4.1. *A theoretical argument for the first bifurcation*

For random complex matrices, Girko [1983] proved a circular law which states that as the dimension of a matrix becomes large, the probability of encountering any real eigenvalues is zero. Given a random matrix, the eigenvalues will be uniformly distributed within the unit circle. Making the unwrought assumption that the set of Jacobians of our systems are the same as a random sample of Girko's random matrices, we will put forth the following argument. The set of eigenvalues lying on any particular axis is of measure zero, thus the probability of an eigenvalue being real is zero. Also, it would seem that there is no difference between negative and positive, thus no reason to favor a positive versus a negative largest eigenvalue. Therefore, we should see as many saddle-node bifurcations as flips. At low dimensions we should see a much higher percentage

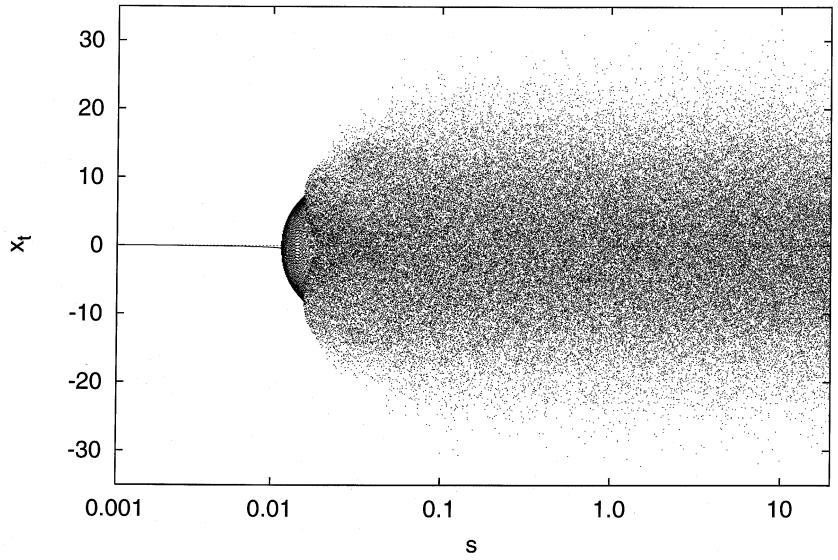

Fig. 6(a). Bifurcation diagram for $n = 64$ and $d = 64$.

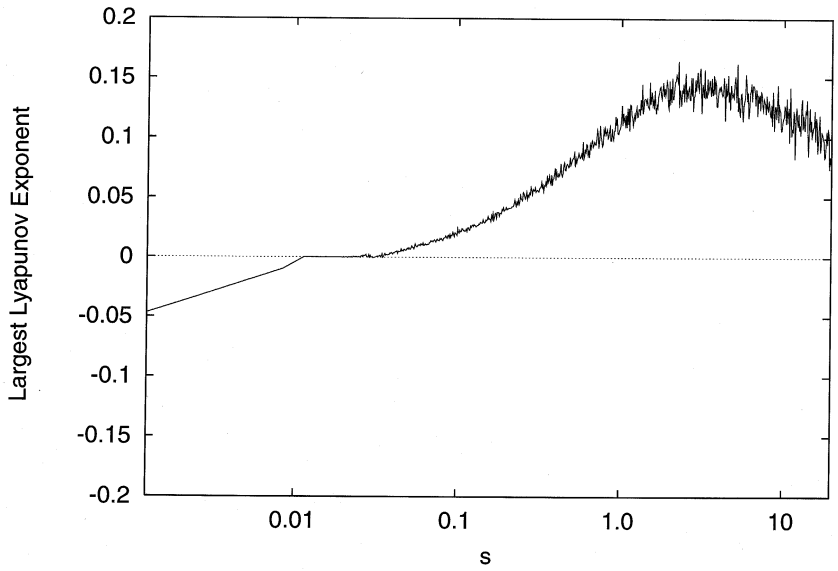

Fig. 6(b). Lyapunov exponent for $n = 64$ and $d = 64$.

of flips and saddle-node bifurcations because Hopf bifurcations need an even number of roots with no real eigenvalues since they occur in pairs. At low dimensions, arranging the roots such that they occur frequently in even numbers, which is necessary for complex solutions, is not as probable as at higher dimensions. We briefly examined this numerically for random Girko-like real matrices. As the dimension is increased the percentage of Hopf bifurcations goes from 0 to about 90 percent for a dimension of 64. Given the aforementioned assumption, it would be reasonable to assume that we would get approximately the same distribution. Note that since complex eigenvalues occur in complex conjugate pairs, given an odd dimension, at least one of the roots must be real.

4.4.2. *Numerical results*

Figure 7 shows the percentage of each bifurcation as the dimension is increased for an intermediate number of neurons. Much like the prediction above, the Hopf's start at about 40 percent of the first bifurcations at $d = 2$ and increase to almost unity at large d. Also, notice that the percentage of flips and saddle-nodes is, on average, equal throughout the range.

Unlike the random matrix case, we also have to deal with the n parameter. Figures 8 and 9 show

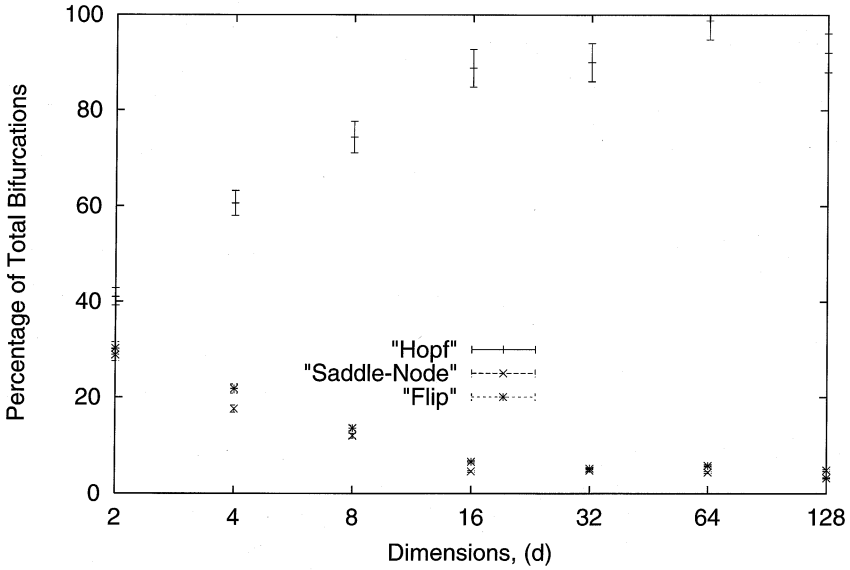

Fig. 7. Percent first bifurcation for $n = 16$, error bars represent the error in the probability.

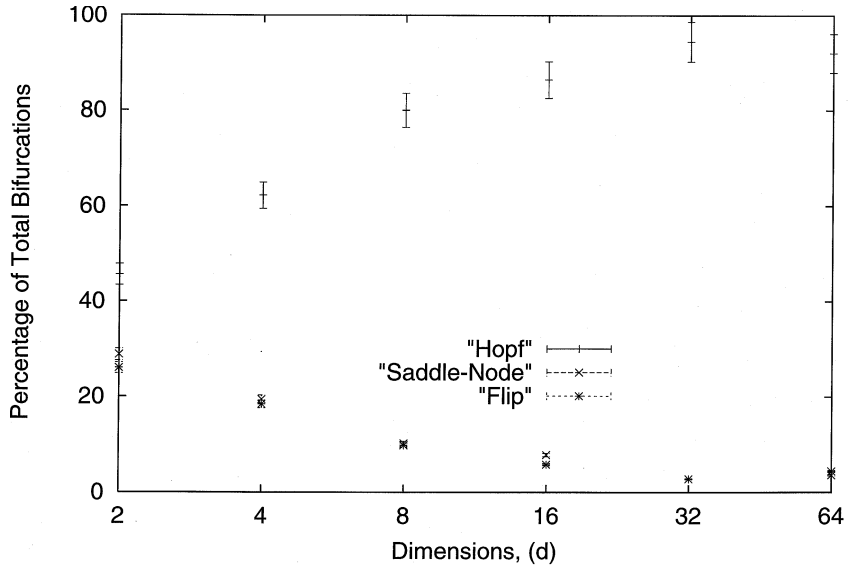

Fig. 8. Percent first bifurcation for $n = 4$, error bars represent the error in the probability.

the percentage of first bifurcation over an increasing range of d at low and high n ($n = 4$ and $n = 256$). Note at high-n and low-d, the percentage of each bifurcation is nearly equal. For the low-n, low-d cases, the percent of each bifurcation is not nearly as close as at high-n. As d is increased, the percentage of Hopf bifurcations rapidly increases so that, at $d = 8$, the percentages of each bifurcation is about equal to those of all n.

The n dependence is an artifact of how we choose the β matrix. Consider a two-dimensional system. For a Hopf bifurcation to occur, the discriminant must be negative. As we increase n, we are increasing the variance of the coefficients of the matrix, thus pushing the expected value of the discriminant positive. The result is a decrease in Hopf bifurcations as n is increased at low d. As d is increased this effect is washed out. Contrasting Figs. 7 and 8 you will notice that at $d = 2$ the percentages of first bifurcations are quite different, but for $d \geq 8$, Figs. 7 and 8 are almost identical.

In the high-d limit, we found that the Hopf bifurcation was overwhelmingly dominant. We looked at cases with d as high as 1024 and found the percent of first Hopf bifurcations approached unity. This confirms the result in [Doyon *et al.*, 1993], that in the limit of high d, the first bifurcation will be Hopf.

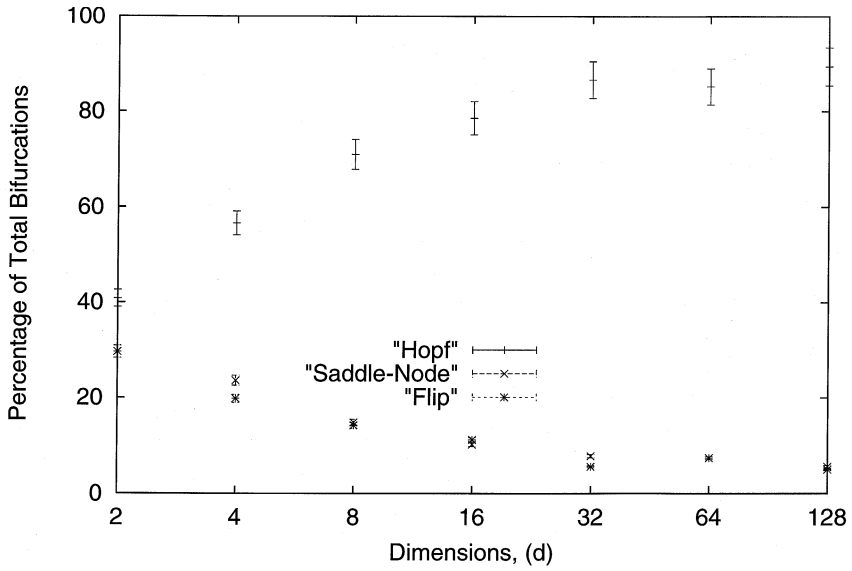

Fig. 9. Percent first bifurcation for $n = 256$, error bars represent the error in the probability.

4.5. *Dynamics after the first bifurcation*

Quantitative results for the probability of a given bifurcation after the first requires developing the Q.R. algorithm [Eckmann & Ruelle, 1985] to allow the tracking of quasiperiodic orbits. Tracking non-periodic orbits involves multiplying the Jacobians at each time step and renormalizing to calculate the eigenvalues. The numerical stability of specific eigenvalues is not good when the number of time steps is large. Although the stability of the modulus of the largest eigenvalue is acceptable and would probably allow us to know when the system bifurcated, which eigenvalue crossed the unit circle would not be certain. For this reason we will now present qualitative trends that occur in these systems after the first bifurcation as gauged by observing the largest Lyapunov exponent, period, correlation dimension, and bifurcation diagrams of hundreds of systems.

4.5.1. *A conjecture for the second bifurcation*

Doyon *et al.* [1993] proved a corollary of Girko's theorem showing that the quasiperiodic route would dominate when the dimension was high for their maps. For our purposes, we will first argue for flows and then adapt portions for maps. Consider a flow on U that has undergone a Hopf bifurcation and is now living on a limit cycle. Now take a local cross-section V which is everywhere transverse to the flow. Next induce a discrete-time map $P : U \to V$ thus creating the "first" return map. P is defined for a $q \in U$ such that $P(q) = h_t(q)$ where t is the time required for the orbit to return to V. This mapping has a fixed point which is the point q on the limit cycle. Considering the mapping P, increase the bifurcation parameter of the original flow and keep track of the eigenvalues of the Jacobian of the discrete map. This will, on average, describe which bifurcation will occur. Apply Girko's circular law to the Jacobian of this map as we did for the first bifurcation argument. For higher dimensions, consider a flow with dimension d. Instead of taking a 1-D "line transverse", cut the flow with a hyper-surface (which will have dimension $d - 1$) that is everywhere transverse to the flow. Map the hyper-surface to itself. As above, increase the bifurcation parameter and look at the eigenvalues of the Jacobian of the map. As d goes to infinity, apply the circular law; most of the eigenvalues lie on the unit disk. The set of real eigenvalues is a set of measure zero with respect to the limiting Girko distribution. Because the set of real eigenvalues has measure zero, the second bifurcation must be Hopf. This suggests that as the dimension is increased, the predominant route to chaos will be the quasiperiodic route.

The argument for maps is somewhat different because taking a Poincaré section for a map is more difficult. First consider the Jacobian at each time step as given by the aforementioned matrix A. Since the probability that $a_i = 0$, and thus the probability that a_d is zero, the A matrix will have full rank. If A_t has full rank $\forall t$, then none of the A

matrices have eigenvalues that are equal to zero. Since

$$\det(\prod A_t) = \det(A_1)\det(A_2)\cdots\det(A_t) \qquad (17)$$

and $\det(A_t) \neq 0\ \forall t$; $\det(\prod A_t) \neq 0$, and thus $\prod A_t$ is nonsingular. From here we only need to apply the circular law to the $\prod A_t$ to see that the quasiperiodic route will be dominant at high-d. Making the original assumption rigorous is beyond the scope of this paper [Brock, 1997].

4.5.2. *Between the first bifurcation and chaos*

As stated above, the dynamics change considerably with d but not with n. The effect of n, for all d considered, is just to "smooth" the dynamics. The Lyapunov exponents are much more steady and smooth, and the bifurcations are much more apparent. Figures 3(b)–6(b) show typical largest Lyapunov exponents over a range of s. Note that the dynamic transitions in Fig. 5(b) at low-n are much more "rough" than the changes in Fig. 6(b) at high-n. The largest Lyapunov exponent does not jump between positive and negative nearly as much. This behavior is typical over the range of n that we studied. However, n does not decrease the dynamic diversity (types of transitions and bifurcations) of a given network. Figures 3(a) and 4(a) show many of the same phenomena and diversity even though the number of neurons in Fig. 3(a) is much greater than in Fig. 4(a).

At $d = 4$ about 40 percent of the first bifurcations are Hopf. After an initial Hopf bifurcation, recognizing a second Hopf is quite difficult, and they do not seem very frequent. More often the second bifurcation appears to be a blue sky, out of a quasiperiodic orbit and into a periodic orbit. After this blue sky bifurcation the system either Hopfs again, or period doubles to chaos [Fig. 4(a)], or just becomes chaotic [Fig. 3(a)]. Occasionally the system will oscillate between a quasiperiodic and a periodic orbit several times before finally reaching the chaotic region, but this usually occurs only for low n. The other 60 percent of the bifurcations are either flips or saddle-nodes. The only difference between the dynamics after a saddle-node and after a flip is that the saddle-node sometimes will jump from one branch of the fork to the other, presumably because the initial conditions lie close to a basin boundary. Most often, after the flip or saddle-node, the system's next bifurcation is a Hopf. We did see period-doubling cascades, but they were infrequent prior to the onset of chaos except at low d. Increasing n increases the probability and strength (magnitude of Lyapunov exponent) of chaos in the system.

The effect of d on the system dynamics is quite significant. As d is increased, the dynamic diversity is greatly reduced. Networks with high d tend to be quite symmetric about the origin; this is just an artifact of our method of choosing the w matrix. The most notable effect of increasing d is the increase of chaos for a given system. Flip and saddle-node bifurcations are almost never seen in high-dimensional networks. Also, once the system has become chaotic, it very rarely shows periodic windows until the s parameter is high enough to saturate the squashing function. In high-d systems, increasing n increases the range of the function. Doyon *et al.* [1993] conclude that for high-dimensional networks the second bifurcation is Hopf. We also see this when the dimension is high, but discerning chaos from limit cycles and tori is often difficult from the bifurcation diagrams. The Lyapunov exponent gives no insight if the second bifurcation is Hopf, but at the onset of chaos, it becomes positive. Doyon *et al.* [1993] also state the quasiperiodic route to chaos dominates at high d. Our results agree, but speculating on how many bifurcations occur up to chaos is difficult.

4.6. *The chaotic region*

4.6.1. *Bifurcation into chaos*

The largest Lyapunov exponent for a nonchaotic map equals the logarithm of the modulus of the largest eigenvalue

$$\lambda_{\text{Lyap}} = \log|\lambda_{\text{eigen}}| \qquad (18)$$

For systems that follow the quasiperiodic route to chaos as previously stated, it is very difficult to track specific eigenvalues, making it difficult to discern which bifurcations occur into chaos. For the same reason that it is difficult to calculate specific eigenvalues for limit cycles, it is difficult to track the eigenvalues in the chaotic region. We do track the modulus of the largest eigenvalue numerically (the Lyapunov exponent), which is used to distinguish chaos from limit cycles. In systems following periodic routes, calculating the eigenvalues is difficult. What is normally seen when plotting the modulus of the largest eigenvalue before and after a flip or saddle-node bifurcation is quite expected.

At the bifurcation point, the modulus of the largest eigenvalue spikes up to one, and then just as quickly drops back down. The largest Lyapunov exponent's behavior coincides with this behavior. What happens at the bifurcation into chaos is not surprising either. The modulus of the largest eigenvalue spikes to one, the largest Lyapunov exponent spikes to zero, and it then proceeds to rise above zero. At this point we can no longer calculate the largest eigenvalue (or any eigenvalues) in a theoretic sense, but we continue to track the Lyapunov exponent numerically. The modulus of the largest eigenvalue is greater than one.

4.6.2. *Dynamics in the chaotic region*

The chaotic region shows the least apparent dynamic variability of any region. For the most part, the return maps of the chaotic regions all look very similar. (Much of this could of course be due to that fact that we are projecting many dimensions on a plane.) As previously discussed, when n is increased, the range of the y_t values increases, (due to the β scaling) but that is the extent of the apparent variability.

At low d the chaotic region is highly nonuniform. The chaotic region contains windows of: Period-doubling sequences, limit cycles, low-period orbits bifurcating to Hopf and then back into chaos; basically every type of dynamics imaginable. As previously stated, Figs. 3(a) and 4(a) are typical of the dynamic diversity at low d. Figures 5(a) and 6(a) are very different. Once the chaotic region is reached, the system remains chaotic until it is forced to be periodic by the saturation of the squashing function. Figure 5(a) is an example of what we call point-intermittent chaos. The chaotic region is strongly dominated by a period-8 cycle with chaos. The chaotic region in Fig. 5(a) is the region corresponding to the positive Lyapunov exponent region in Fig. 5(b). Figure 6(a) has a much stronger chaotic region than Fig. 5(a), suggesting that increasing the available complexity (increase in n), increases the chaos. Figure 6(a) is typical of high d dynamics, limit cycles leading into a highly chaotic region.

4.6.3. *Point-intermittent chaos*

The chaotic region in Fig. 5 is especially interesting. When the network corresponding to Fig. 5 is saturated, i.e.

$$\tanh\left(s\omega_{i0} + s\sum_{j=1}^{d}\omega_{ij}y_{t-j}\right) = \pm 1 \ \forall i,\ j \qquad (19)$$

then there are eight possible states per dimension.[2] (The upper bound on the period is then $d2^n$.) The system is chaotic when the squashing function is not quite saturated. For most of the trajectory the squashing function is saturated and the trajectory tends toward one of the eight attracting points. Certain combinations of inputs cause a neuron to become unsaturated, causing a point that misses the attracting set by a significant amount causing a positive largest Lyapunov exponent at that time step. This miss effectively resets the periodic orbit; it also adds a nonattracting point to the y array. A periodic orbit is interrupted before it has completed one period by an intermittent point that is not one of the attracting points. After this miss occurs, the time-series resumes and again consists of points in the attracting set. If the system misses the attracting set enough times, the largest Lyapunov exponent will become positive on average. The average time between misses is proportional to s. This behavior creates a sequence of periodic trajectories that are strung together by these points that miss the periodic orbit. The system is locally (over a short, periodic region) not chaotic, but, because of the averaging, globally chaotic.

The focus now is the cause of missing the periodic points. Computers have round-off errors which can play a significant role in the dynamics; we argue that the chaos in Fig. 5 not such an artifact. Consider a function

$$\psi = \begin{cases} \pm 1 & \text{for } |x| > a \\ \tanh(x) & \text{for } |x| \le a \end{cases} \qquad (20)$$

When all $|x| > a$ the system is finite state and periodic. When $|x|$ is not greater than a for all x, then there are an infinite number of states that occur along with the eight attracting points. Every missed point not only "resets" the period, but since it adds a new point into the y_t array, allows for the possibility of saturating ψ in a different way than the 8 attracting points could. The existence of these intermittent points giving rise to other,

[2]There happen to be eight states for this system; there is a possibility of 2^n states. We will discuss this in a later section.

different intermittent points, allows for an infinite number of possible intermittent points. As s is increased, the probability that ψ will be saturated increases, thus giving rise to less intermittent points.

4.7. *Dynamics after the chaotic region*

The dynamics after the chaotic region (large s) are quite interesting and surprising. At low d (Figs. 3 and 4), there is a rapid (i.e. one or two increments of s), transition from chaos to periodicity. At low d, increasing n increases the complexity, as can be seen by comparing Figs. 3 and 4; high n tends to have a greater chance of chaotic windows after the periodic behavior starts, but the dynamics are qualitatively the same.

As before, at high d [Figs. 5(a) and 6(a)] the situation is quite different. At low n the transition from the chaotic to periodic region is quite difficult to find. Often there are strong periodic regions, with very small windows of high-period, quasiperiodic, and chaotic orbits. If s is increased enough, the system becomes purely periodic. At high n the chaotic region occurs over a much larger range of s. It is not until s is made very large (1000), that the periodicity becomes prominent. In this region, the transition between chaos and periodicity is very similar to that of the lower-n case, only much more gradual.

4.7.1. *Theoretical argument for the highest final period*

As the s parameter is taken to infinity, for all practical purposes (i.e. using a machine with limited precision), the squashing function becomes a binary or step function. This implies that the system must repeat, and thus cannot be chaotic. The dynamics at high s consist of periodic orbits of varying period. There is often a dominant period, but even at s values as high as 1024 we see periodic windows. At high s, each neuron has two states, there are d sets of n neurons, and thus the highest period the networks can see is:

$$P_h = d2^n \tag{21}$$

We ran some experiments replacing the hyperbolic tangent squashing function with a step function. At very low d we did see Eq. (21) reached. At d greater than 4 we never observed periods as high as Eq. (21) (c.f. [Kauffman, 1993]).

4.8. *Interpretation of the qualitative results*

Since all our results after the first bifurcation are very qualitative, we would like to give some general interpretations of the results. Increasing n as could be deduced from reading [Hornik *et al.* 1989, 1990], smoothes the dynamics. Increasing n also increases the available dynamics, thus increasing the complexity, which increases both the probability and strength of chaos. Increasing n increases the ability to approximate, thus more dynamics can be accounted for, and seen. The d parameter increases the embedding dimension of the network and increases the probability of chaos. Another effect of increasing d is the slowing of the dynamics. This decreases the largest Lyapunov exponent and also smoothes the dynamics. At high d most of the systems appear very similar, whereas for low d the dynamics are quite diverse. Low-d networks are almost exclusively nonchaotic. Since we looked only at cases we knew were chaotic, we sampled little of the low-d network space. Many low-d networks stay at fixed points throughout the range of s considered and thus look quite similar. Thus the apparent diversity of the low-d networks might be due to the fact that we are only sampling 4 to 10 percent of the low-d space.

5. Conclusion

The major results of this study are: (1) as the number of degrees of freedom are increased, the probability of chaos approaches unity given a system that is sufficiently nonlinear, (2) as the dimension is increased, the most probable first bifurcation is Hopf; the probabilities of saddle-node and flip bifurcations are equal, and (3) qualitatively the quasiperiodic route to chaos is the most probable as the dimension is increased. The generality of the results hinges on the methods used to assign the weight matrices, but given the conditions above, the results are general. The source code and additional details can be found at http://sprott.physics.wisc.edu/neural/.

Acknowledgments

We would like to thank William Brock, Ian Dobson, and Cosma Shalizi for many helpful discussions. The authors would like to thank Derek

Wright of the U.W. Madison Computer Science Condor group for help collecting data. Much of the data was run via the Condor High-Throughput Computing System; without this resource our data collection would be far more sparse. Information concerning the use of Condor can be found at http://www.cs.wisc.edu/condor/

References

Albers, D. J., Sprott, J. C. & Dechert, W. D. [1996] "Dynamical behavior of artificial neural networks with random weights," in *Intelligent Engineering Systems Through Artificial Neural Networks*, eds. Dagli, C. H., Akay, M., Chen, C. L. P., Fernanndez, B. R. & Ghosh, J. (ASME, NY), pp. 17–22.

Bowen, R. [1978] *On Axiom A Diffeomorphisms, CBMS Regional Conference Series in Mathematics* (A.M.S. Publications, Providence).

Brock, W. [1997] Private communication.

Doyon, B., Cessac, B., Quoy, M. & Samuelides, M. [1993] "Control of the transition to chaos in neural networks with random connectivity," *Int. J. Bifurcation and Chaos* **3**, 279–291.

Eckmann, J. P. & Ruelle, D. [1985] "Ergodic theory of chaos and strange attractors," *Rev. Mod. Phys.* **57**(3), 617–656.

Girko, V. L. [1983] "Circular law," *Theor. Prob. Appl.* **29**, 694–706.

Grebogi, C., Hammel, S., Yorke, J. & Sauer, T. [1990] "Shadowing of physical trajectories in chaotic dynamics: Containment and refinement," *Phys. Rev. Lett.* **65**, 1527–1530.

Guckenheimer, J. & Holmes, P. [1983] *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (Springer-Verlag, NY).

Hornik, K., Stinchocombe, M. & White, H. [1989] "Multilayer feedforward networks are universal approximators," *Neural Networks* **2**, 359–366.

Hornik, K., Stinchocombe, M. & White, H. [1990] "Universal approximation of unknown mapping and its derivatives using multilayer feedforward networks," *Neural Networks* **3**, 535–549.

Kaufman, S. [1993] *The Origins of Order, Self-Organization and Selection in Evolution* (Oxford University Press, NY).

l'Écuyer, P. [1988] "Efficient and portable combined random number generators," *Commun. ACM* **31**, 742–749.

Naimark, J. [1959] "On some cases of periodic motions depending on parameter," *Dokl. Akad. Nauk*, 736–739.

Newhouse, S. E. [1980] "Lecture on dynamical systems," *Dyn. Syst.* **8**, 1–114.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. [1992] *Numerical Recipes in C* (Cambridge University Press, Cambridge).

Reick, C. & Mosekilde, E. [1995] "Emergence of quasiperiodicity in symmetrically coupled, identical period-doubling systems," *Phys. Rev.* **E52**, 1418–1435.

Ruelle, D. [1989] *Elements of Differentiable Dynamics and Vector Fields* (Academic Press).

Sacker, R. S. [1965] "On invariant surfaces and bifurcations of periodic solutions of ordinary differential equations," *Comm. Pure Appl. Math.*, 717–732.

Takens, F. [1980] "Detecting strange attractors in turbulence," in *Lecture Notes in Mathematics* **898**, eds. Rand, D. & Young, L. (Springer-Verlag, Berlin), pp. 366–381.

Wang, W. & Cerdeira, H. A. [1996] "Dynamical behavior of the multiplicative diffusion coupled map lattices," *Chaos* **6**, 200–208.

Wolf, A., Swift, J. B., Swinney, H. L. & Vastano, J. A. [1984] "Determining Lyapunov exponents from a time series," *Physica* **D16**, 285–317.