Time series generation by multilayer networks

Liat Ein-Dor and Ido Kanter
Minerva Center and Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel

(Received 12 January 1998)

The properties of time series, generated by continuous valued multilayer networks consisting of one or two hidden layers, are studied analytically. The time series is generated by using past output values to determine the next input vector. The main results for the generic asymptotic behavior are (a) The attractor dimension is only a function of the number of hidden units in the first hidden layer; (b) the analytical solution for the time series generated by the networks mirrors the structure of the network itself. [S1063-651X(98)13906-5]

PACS number(s): 05.20.−y, 87.10.+e

I. INTRODUCTION

The main goal of analytical research in the field of neural networks during the past decade has been to examine the ability of various architectures to store, to retrieve, and to learn from *random* examples [1,2]. Nevertheless, the content of natural or artificial data streams is, generally speaking, expressed in the correlations, spatial and temporal, among the data points. Hence, extending the neural network approach to deal with time series is of great interest [3,4].

There are two main lines of approach in the investigation of time series. In the first approach, the time series is given and the following two questions must be answered: (1) is a given network capable of learning a segment of the sequence; and (2) what is the quality of the prediction on the part of the sequence that has not been shown to the network. In practice, for a given time series, predictors based on ideas from the realm of neural networks can be built and their success can be compared to other linear or nonlinear predictors. However, as long as the statistical nature of the examined sequences, their origin, and the available space for the architecture of the trained network are not well restricted, a general theory cannot be established.

In the second approach, we recently studied the statistical nature of time series generated by a given network with a particular architecture and dynamical rules. The focus is then placed on what kinds of time series (their complexity, etc.) a given network can generate and hence can predict accurately. Of course, forecasting of a particular sequence cannot be answered. However, we would like to build a classification of the possible outcome sequences as a function of the architecture and dynamical rules. This classification is a prerequisite for any theoretical insight in the field of time series prediction. For instance, the classification can answer the underlying question of which architecture has to be chosen for the predictor. Of course we would not choose an architecture that is incapable of learning the sequence regardless of the particular set of weights, fixed by the learning algorithm.

A beginning of such classification was recently developed [5,6] and indicates that there is an interplay between the architecture of a multilayer network with one hidden layer and the attractor dimension (d_A) of the time series generated by the multilayer feedforward networks (MLN), the d_A being a function of the number of hidden units [7]. This feature quantitatively distinguishes between the computational ability of MLN with a different number of hidden units. Adding additional hidden units vastly expands the set of sequences generable with the network.

In this paper we first report in detail results for MLN with one hidden layer, and enlarge the investigation to a restricted network with two hidden layers.

In Sec. II, the particular architectures and their dynamical rules are defined. In Sec. III, previous findings are briefly summarized and questions raised. In Sec. IV, results for MLN with one hidden layer are presented, and in Sec. V results are extended to MLN with two hidden layers. Results for a general set of weights between the input and the first hidden units are briefly discussed in Sec. VI. Conclusions are presented in Sec. VII.

II. ARCHITECTURES AND DYNAMICAL RULES

The examined architectures are multilayer feedforward networks, with one or two hidden layers. The network with one hidden layer is denoted as $N{:}M{:}1$, N input units S_j, $j=1,\ldots,N$, M hidden units σ_i^1, $i=1,\ldots,M$ and 1 output unit *out* (see Fig. 1). The symbol W_{ij} signifies the weight between the jth input unit and the ith hidden unit and, for simplicity, the weights between the hidden units and the output unit are set equal to 1 (see Fig. 1).

The network with two hidden layers is defined as

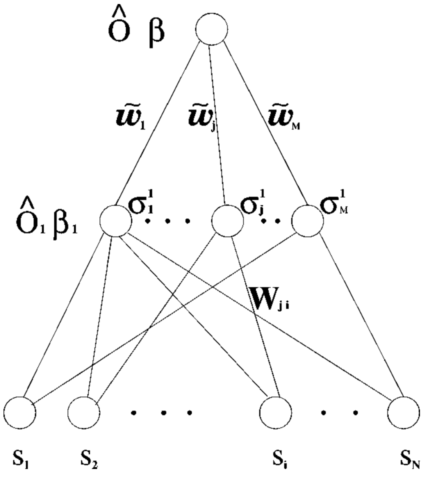

FIG. 1. The architecture $N{:}M{:}1$.

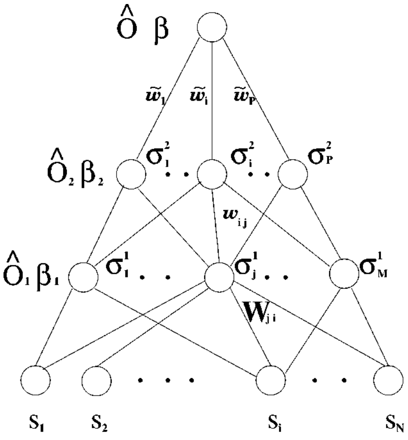

FIG. 2. The architecture $N{:}M{:}P{:}1$.

$N{:}M{:}P{:}1$, N input units S_i, $i=1,\dots,N$, M hidden units in the first hidden layer σ_i^1, $i=1,\dots,M$, P hidden units in the second hidden layer σ_i^2, $i=1,\dots,P$, and 1 output unit *out* (see Fig. 2). The symbol W_{ij} signifies the weight between the jth input unit and the ith hidden unit in the first layer. The symbol w_{ij} stands for the weight between the jth hidden unit in the first hidden layer and the ith hidden unit in the second hidden layer. Again, for simplicity, the weights between the second hidden layer and the output unit set equal to 1 (see Fig. 2).

Starting from an initial configuration for the N input units $\{S_1,S_2,\dots,S_N\}$ the dynamics is defined as follows. The ith hidden unit in the first hidden layer is fixed by

$$\sigma_i^1=\hat{o}_1\left[\beta_1\sum_{j=1}^{N} W_{ij}S_j\right], \tag{1}$$

where $\hat{o}_1$ is the activation function of the hidden units in the first layer, which, for simplicity, is taken to be the same for all hidden units, and β_1 is the gain factor. Similarly, the ith hidden unit in the second hidden layer is fixed by

$$\sigma_i^2=\hat{o}_2\left[\beta_2\sum_{j=1}^{M} w_{ij}\sigma_j^1\right], \tag{2}$$

where $\hat{o}_2$ is the activation function of hidden units in the second hidden layer with a gain factor β_2. The output of the network with one hidden layer is given by

$$\mathcal{O}=\hat{o}\left[\beta\sum_{j=1}^{M} \tilde{w}_j\sigma_j^1\right] \tag{3}$$

and for the network with two hidden layers is given by

$$\mathcal{O}=\hat{o}\left[\beta\sum_{j=1}^{P} \tilde{w}_j\sigma_j^2\right], \tag{4}$$

where in both cases $\tilde{w}_j$ denotes the weight between the jth hidden unit (in the last hidden layer) and the output. The input at each successive time step is chosen as follows: the inputs from the previous time step are shifted one unit to the right with the state of the leftmost input unit set equal to the state of the output unit in the previous time step. Symbolically,

$$S_j^{t+1}=S_{j-1}^{t} \quad j=2,\dots,N; \quad S_1^{t+1}=\mathcal{O}^t. \tag{5}$$

For time steps $t>N$ one can summarize the dynamical evolution of the network $N{:}M{:}1$ by the following equation:

$$S^t=\hat{o}\left\{\beta\sum_{i=1}^{M} \tilde{w}_j\hat{o}_1\left[\beta_1\sum_{j=1}^{N} W_{ij}S^{t-j}\right]\right\}, \tag{6}$$

where S^t is the output at time t, and of the network $N{:}M{:}P{:}1$ by

$$S^t=\hat{o}\left[\beta\sum_{m=1}^{P} \tilde{w}_m\hat{o}_2\left\{\beta_2\sum_{k=1}^{M} w_{mk}\hat{o}_1\left[\beta_1\sum_{j=1}^{N} W_{kj}S^{t-j}\right]\right\}\right]. \tag{7}$$

These equations indicate that the network generates an infinite sequence from an initial state of the input units in the following manner. The dynamical evolution of one degree of freedom, S^t, depends on its values in the previous N steps $S^t=f\{S^{t-1},S^{t-2},\dots,S^{t-N}\}$. The special form of the function f depends on the details of the architecture and the dynamical rules and is explicitly given by Eqs. (6) and (7).

To simplify the discussion, below we restrict the parameter space such that

$$\beta=\beta_1=\beta_2 \tag{8}$$

and the activation function in all levels is the same:

$$\hat{o}_i=\tanh \quad \text{or} \quad \hat{o}_i=\sin. \tag{9}$$

The choice of the tanh activation function seems to be natural, but the mathematical simplification of the sin activation function will be explained below.

III. QUESTIONS

In previous studies [7] we claim that a perceptron with the same dynamical rules exhibits the following characteristic features: (a) Flows can be periodic or quasiperiodic depending on the phase of the weights. A phase shift in the weights results in a frequency shift in the output. (b) The dimension of the attractor in the generic case is less than or equal to 1, regardless of the complexity of the weights. One can now conclude that a perceptron with these dynamical rules is capable of *generating* only time series that are characterized by the attractor dimension $d_A\leqslant 1$. Hence, under the same dynamical rules (known in other communities as one-time-lag dynamics or sliding windows [8,9]) one can possibly *learn* and *predict* with a perceptron only time series which are characterized by $d_A\leqslant 1$. We said ''possibly,'' since it is as yet unclear whether all possible time series with $d_A=1$ can be learned and predicted by a perceptron with freedom to choose the appropriate activation function.

The generalization of the perceptron to a MLN with one hidden layer consisting of M hidden units indicates that such a network is capable of generating time series with an integer $d_A\leqslant M$, where the d_A increases with the gain factor. The weights and the activation functions of the hidden units and the output unit only influence the shape of the attractor. The detailed calculations for a MLN with one hidden layer are presented below in Sec. IV.

However, a few questions remain to be answered.

(1) From the asymptotic behavior of the time series generated by a MLN with one hidden layer one can conclude that the d_A is a function of the number of hidden units, but has no interplay with the size of the input. At this stage a few scenarios are possible for more structured MLN with more than one hidden layer, $N{:}M_1{:}M_2{:}\dots{:}M_L{:}1$. It is plausible that the d_A is only a function of the size of the first hidden layer M_1, or that the d_A is only a function of the size of the last hidden layer, which feeds the output unit M_L or that the d_A is a function of the size and the order of all the hidden layers $\{M_1,\dots,M_L\}$.

(2) The translating solution of a MLN with one hidden layer mirrors the architecture of the network, regardless of details of the weights and the particular choice of the odd activation functions. Weights in the lowest level are acted upon by the activation function of the first hidden layer and then in turn acted upon by the activation function of the output unit. The question is whether this mathematical beauty is conserved also for more structured MLN. In an affirmative case, one can immediately find the form of the dynamical evolution of any MLN with these dynamical rules. Only the coefficients have to be determined explicitly via careful and tedious algebra.

(3) After the previous two questions are answered, and the interplay between the details of generated time series and the architecture and the dynamical rules of the MLN can be understood, one may ask the following question: when is it necessary to, or what is the advantage of, increasing the number of hidden layers? More precisely, what quantitatively distinguishes between the computational ability of MLN with a different number of hidden layers, and does adding additional layers vastly enlarge the set of sequences that can be generated with the network?

IV. A MLN WITH ONE HIDDEN LAYER

The dynamical evolution of the network $N{:}3{:}1$ (Fig. 1) and with tanh activation function is given by [see Eq. (6) and Eq. (8)]

$$S^t=\tanh\left\{\beta\sum_{i=1}^{3}\tilde{w}_i\tanh\left[\beta\sum_{j=1}^{N}W_{ij}S^{t-j}\right]\right\}. \qquad (10)$$

Let us consider first the case where the weight vector for each one of the hidden units consists of a single Fourier component, where more structured weights are examined in simulations. In particular, let

$$W_{ij}=R_i\cos[2\pi K_i j/N], \qquad (11)$$

where $K_i\neq 0$ denotes the wave number to the ith hidden unit and R_i is the amplitude. We assume in the following analytical treatment that the wave numbers $\{K_i\}$ are relatively prime, where in other cases similar solutions can be found. The dynamical solution of Eqs. (10) and (11) is given by

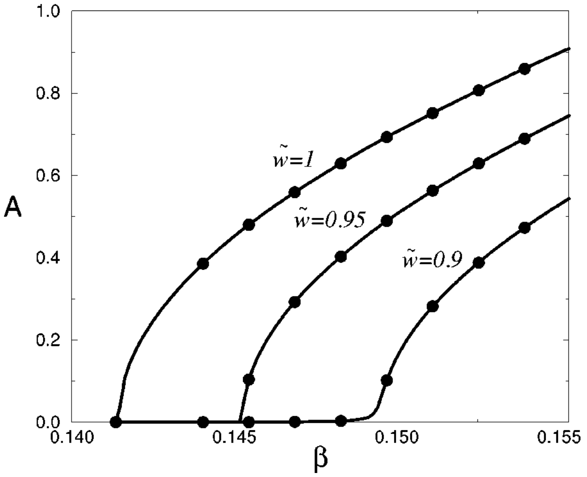

FIG. 3. Result of simulations for MLN with one hidden layer 100:3:1, $\tilde{w}_1=1,\tilde{w}_2=0.95,\tilde{w}_3=0.9$ and with tanh activation function. The amplitude obtained from simulations with $N=100$ (●) and analytically Eqs. (12)–(16) (solid lines).

$$S^t=\tanh\left\{\beta\sum_{i=1}^{3}\tilde{w}_i\tanh\left[A_i\cos\left(\frac{2\pi K_i t}{N}\right)\right]\right\}. \qquad (12)$$

This solution can be verified by the expansion of Eqs. (10) and (12) in power series of A_i. Since the presentation of the three coupled equations for A_i are involved, we present the solution only for the case where $R_i=1$. The constant A_i $(i=1,2,3)$ depends on β through the equation

$$A_i=\beta N\sum_{\mu=1}^{\infty}C_\mu\beta^{2\mu-1}\sum_{s=0}^{\mu-1}\sum_{m=0}^{\mu-s-1}\binom{2\mu-1}{2s}\binom{2(\mu-s)-1}{2m}\times D^{1}_{2(\mu-s-m)-1}(i)D^{0}_{2s}(j)D^{0}_{2m}(k), \qquad (13)$$

where i, j, and k are three different integers representing the three hidden units and

$$D^{x}_{m}(i)=\tilde{w}_i{}^{m}\sum_{\nu_1,\nu_2,\dots,\nu_m=-\infty}^{\infty}\prod_{r=1}^{m}Z_{\nu_r}\delta\left(\sum_{r=1}^{m}\nu_r-x\right), \qquad (14)$$

$$Z_\nu(i)=\sum_{\rho=1}^{\infty}\gamma_i(\rho)\sum_{n=0}^{2\rho-1}\binom{2\rho-1}{n}\delta(2(\rho-n)-1-\nu), \qquad (15)$$

$\gamma_i(\rho)=2A_i^{2\rho-1}(2^{2\rho}-1)B_{2\rho}/(2\rho)!$, $C_\rho=2^{2\rho}(2^{2\rho}-1)B_{2\rho}/(2\rho)!$ and B_ρ are the Bernoulli numbers [10]. This solution is exact for any system size N and a positive integer wave number K. We find that in a small gain regime

$$\beta<\beta_c^i=\sqrt{\frac{2}{N\tilde{w}_i}} \qquad (16)$$

the only solution is the trivial fixed point $S^t=0$. At β_c this solution becomes unstable, and the system undergoes a Hopf bifurcation to a periodic orbit of length N $(S^{t+N}=S^t)$ characterized by a nonzero amplitude A. Numerical solutions of Eqs. (12)–(15) are presented in Fig. 3 for $N=100$ with $\tilde{w}_1$

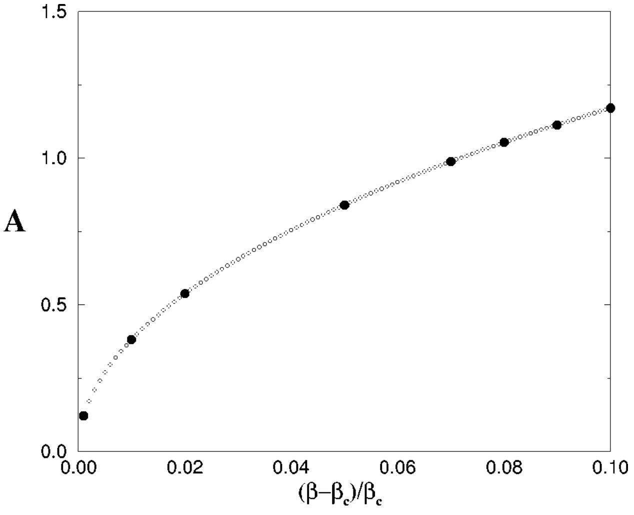

FIG. 4. Result of simulations for MLN with one hidden layer N:3:1 and with sin activation function. The amplitude obtained from simulations with $N=100$ (●) and analytically, Eq. (18), (○).

$=1$, $\tilde{w}_2=0.95$, and $\tilde{w}_3=0.9$. Results are found to be in agreement with the stationary amplitude observed in simulations of the same system (see Fig. 3). The system undergoes three Hopf bifurcation transitions, following Eq. (16), where in each one of them one of the three hidden units becomes greater than zero ($A_i>0$). Note that the critical gain β_c^i scales with $N^{-1/2}$ whereas for the perceptron $\beta_c \propto 1/N$.

The origin of mathematical complication of the above solution is the use of the tanh activation function. From Eq. (12) one can see that the *stationary solution evolves as a* tanh *acting over a sum of* tanh *and, unfortunately, no elegant way exists to expand in power series of A such an expression.* Since we would like to solve more structured networks we observed that sin activation function should simplify the calculations. The idea is that sin[sin(x)+sin(y)] can be written as sin[sin(x)]cos[sin(y)]+cos[sin(x)]sin[sin(y)] where now each term can be easily expanded using the Bessel functions [11]. More precisely, the stationary solution Eq. (12), for the sin activation function, is now replaced by

$$S^t=\sin\left\{\beta\sum_{i=1}^{3}\tilde{w}_i\sin\left[A_i\cos\left(\frac{2\pi K_i t}{N}\right)\right]\right\}. \tag{17}$$

For simplicity, we take $\tilde{w}_i=1$ and $R_1=1$ [Eq. (11)] and therefore $A_i=A$. The constant A now depends on β through the equation

$$A=2\beta N\left[\sum_{p=0}^{\infty}J_{2p+1}(\beta)/J_1((2p+1)A)\right] \times\left[J_0(\beta)+2\sum_{p=1}^{\infty}J_{2p}(\beta)J_0(2pA)\right]^2, \tag{18}$$

where $J_p(x)$ is the Bessel function of the first kind of order p. Again for $\beta<\beta_c$ the only solution is the trivial fixed point $S^t=0$. At β_c [given by Eq. (16)] this solution becomes unstable, and the system undergoes a Hopf bifurcation to a periodic orbit of length N ($S^{t+N}=S^t$) characterized by a nonzero amplitude A. Numerical solutions of Eqs. (18) are presented in Fig. 4 for $N=100$. Results are found to be in agreement with the stationary amplitude observed in simulations of the same system (see Fig. 4).

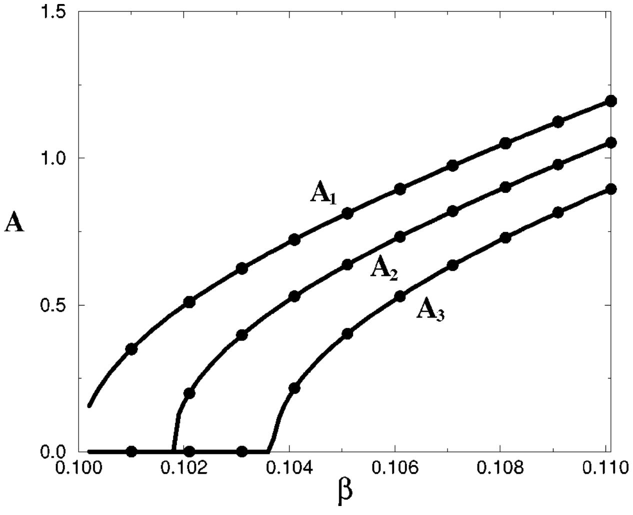

FIG. 5. A_i vs β for the architecture N:3:2:1 with tanh activation function. The weights from the first hidden layer to the second one, $\{w_{ij}\}$, are given by $w_{1j}=1$,$w_{21}=1$,$w_{22}=0.9$,$w_{23}=0.8$. Analytical solution for A_i, Eqs. (21)–(24) (solid lines) and simulations of this network with $N=1000$ (●).

V. A MLN WITH TWO HIDDEN LAYERS

In this section we present the results for the architecture N:3:2:1 (Fig. 2), which is a prototype MLN with two hidden layers. In order to simplify the presentation of the analytical treatment we assume again that $\tilde{w}_i=1$. The dynamical evolution of the network with tanh activation function is given by

$$S^t=\tanh\left\{\beta\sum_{i=1}^{2}\tanh\left(\beta_2\sum_{j=1}^{3}w_{ij}\tanh\left[\beta_1\sum_{m=1}^{N}W_{jm}S^{t-m}\right]\right)\right\}. \tag{19}$$

For the case where the weight vector for each one of the hidden units consists of a single Fourier component, relatively prime [see Eq. (11)] and with $R_i=1$ the dynamical solution has the following form:

$$S^t=\tanh\left[\beta\sum_{i=1}^{2}\tanh\left\{\beta\sum_{j=1}^{3}w_{ij}\tanh\left[A_j\cos\left(\frac{2\pi K_j t}{N}\right)\right]\right\}\right] \tag{20}$$

Although $R_i=1$ the solution is more involved, since the weights between the first and the second hidden units, $\{w_{ij}\}$, are not identical and therefore $A_i\neq A_j$. The self-consistent equation for A_1 is given explicitly by

$$A_1=\beta N\sum_{\delta=1}^{\infty}C_\delta\beta^{2\delta-1}\prod_{i=1}^{2\delta-1}\sum_{\{p_i^j=-\infty\}}^{\infty}L(p_i^1,p_i^2,p_i^3) \times\delta\left(\sum_{i=1}^{2\delta-1}p_i^1-1\right)\delta\left(\sum_{i=1}^{2\delta-1}p_i^2\right)\delta\left(\sum_{i=1}^{2\delta-1}p_i^3\right), \tag{21}$$

and

$$L(p_1,p_2,p_3)=\sum_{i=1}^{2}\sum_{\mu=1}^{\infty}C_\mu\beta^{2\mu-1}\sum_{s=0}^{2\mu-1}\sum_{m=0}^{2\mu-1-s}\binom{2\mu-1}{s}\binom{2\mu-s-1}{m}D_s^{p_1}(1,i)D_m^{p_2}(2,i)D_{2\mu-s-m-1}^{p_3}(3,i), \tag{22}$$

$$D_m^p(j,i)=\sum_{\nu_1,\nu_2,\dots,\nu_m=-\infty}^{\infty}\prod_{l=1}^{m}Z_{\nu_l}(j,i)\,\delta\left(\sum_{q=1}^{m}\nu_q-p\right), \tag{23}$$

$$Z_\nu(j,i)=\sum_{\rho=1}^{\infty}\gamma(\rho,j,i)\sum_{p=0}^{2\rho-1}\binom{2\rho-1}{p}\delta(2(\rho-p)-1-\nu). \tag{24}$$

$\gamma(\rho,j,i)=w_{ij}2A_j^{2\rho-1}(2^{2\rho}-1)B_{2\rho}/(2\rho)!$. This solution is again exact for any system size N and positive wave numbers. For a small gain

$$\beta<\beta_c=\frac{1}{N^{1/3}}\min\{[2/(w_{11}+w_{21})]^{1/3},[2/(w_{12}+w_{22})]^{1/3},[2/(w_{13}+w_{23})]^{1/3}\} \tag{25}$$

the only trivial solution is $S^t=0$, where at β_c the system undergoes a Hopf bifurcation to a periodic orbit of length N, characterized by a nonzero amplitude of at least one of the $\{A_i\}$. The critical gain in which each one of the hidden units in the first hidden layer becomes nonzero scales with $N^{-1/3}$, but the prefactor is a function of the weights connecting the hidden units to the output unit. As an example, the critical gain for the first hidden unit is $N^{-1/3}[2/(w_{11}+w_{21})]^{1/3}$.

The numerical solution of Eqs. (20)–(24) is not an easy task, since Eq. (21) and Eq. (23) consist of multiple summations over many variables which obey only one global restriction. We are able to solve numerically this set of equations perturbatively, keeping terms up to the cubic terms, A_i^3. This approximation is valid only near the first bifurcation, and was confirmed by simulations. The examination of whether the dynamical solution, Eq. (20), is valid for a wider range of the gain factor β can be answered by solving simultaneously Eqs. (19) and (20). This can be summarized by the following set of three ($j=1,2,3$) coupled iterative equations

$$A_j^{q+1}\cos\left(\frac{2\pi K_j t}{N}\right)=\beta\sum_{i=1}^{N}\cos\left(\frac{2\pi K_j i}{N}\right)\tanh\left[\beta\sum_{m=1}^{2}\tanh\left\{\beta\sum_{l=1}^{3}w_{ml}\tanh\left[A_l^q\cos\left(\frac{2\pi K_l(t-i)}{N}\right)\right]\right\}\right]. \tag{26}$$

For large q, A_j^q converges to a constant independent of the wave numbers $\{K_j\}$. The asymptotic fixed point solution of these equations as a function of β is given in Fig. 5, and is in an agreement with the stationary solution found in simulations on finite systems. Note that the system undergoes three transitions, each one of them corresponding to a transition of one of the hidden units in the first hidden layer.

In order to be able to solve explicitly this architecture for any given β we now replace the tanh by sin activation function and, as we explained above, this modification should simplify the calculations. Similarly to Eq. (20), the stationary solution in this case is given by

$$S^t=\sin\left[\beta\sum_{i=1}^{2}\sin\left\{\beta\sum_{j=1}^{3}w_{ij}\sin\left[A_j\cos\left(\frac{2\pi K_j t}{N}\right)\right]\right\}\right] \tag{27}$$

and the coefficient A_1, for instance, is given by

$$A_1=\frac{\beta N}{2}\left\{\gamma_1(1)+\gamma_1(2)+4\sum_{\delta_1\delta_2=0}^{\infty}J_{2\delta_1+1}(\beta)J_{2(\delta_2+1)}(\beta)[D^1(x)+D^1(\overline{x)}]\right\}, \tag{28}$$

where $D^1(x)$ is defined by

$$D^1(x)=T_+^1(C_+^2C_+^3+C_-^2C_-^3)-T_-^1(C_-^2C_+^3+C_+^2C_-^3) \tag{29}$$

and $C_\pm^\alpha$ and $T_\pm^\alpha$ are given explicitly as a function of x by

$$C_\pm^\alpha=\frac{1}{2}\left[J_0(x_\pm^\alpha)\pm J_0(x_\mp^\alpha)+2\sum_{k=1}^{\infty}[J_{2k}(x_\pm^\alpha)\pm J_{2k}(x_\mp^\alpha)]J_0(2kA_\alpha)\right], \tag{30}$$

$$T^{\alpha}_{\pm}=2\sum_{k=0}^{\infty}\left[J_{2k+1}(x^{\alpha}_{+})+J_{2k+1}(\pm x^{\alpha}_{-})\right]J_1[(2k+1)A_{\alpha}], \tag{31}$$

$$\gamma_{\alpha}(i)=8J_0(\beta)\sum_{\delta=0}^{\infty}J_{2\delta+1}(\beta)\left[\sum_{k_1=0}^{\infty}J_{2k+1}[\beta w_{i\alpha}(2\delta+1)]J_1[(2k+1)A_{\alpha}]\right]\pi^{3}_{j\neq\alpha}\{J_0(\beta w_{ij})[2\delta+1]$$

$$+2\sum_{k=1}^{\infty}J_{2k}[\beta w_{ij}(2\delta+1)]J_0(2kA_j)\} \tag{32}$$

and $x^{\alpha}_{\pm}=\beta[(2\delta_1+1)w_{1\alpha}\pm 2(\delta_2+1)w_{2\alpha}]$. The definition of $D(\overline{x})$ [see Eq. (28)] is similar with $\overline{x^{\alpha}_{\pm}}=\beta[(2\delta_1+1)w_{2\alpha}\pm 2(\delta_2+1)w_{1\alpha}]$. Although the form of Eqs. (28)–(31) seems to be complicated, they are much simpler than Eqs. (22)–(24) for the tanh activation function. The difference is that the equations for the sin activation function consist at most of only *three* summations whereas for the tanh activation function the multiple summation is unbounded. A solution of Eqs. (28)–(31) for a particular set of $\{w_{ij}\}$ is presented in Fig. 6, and an agreement between simulations and the analytical treatment is observed.

Note that lifting the degeneracy among the Hopf bifurcation transitions of the hidden units in the first layer, $\beta_c^i \neq \beta_c^j$, can be achieved in the following two ways: (a) lifting the degeneracy in the coming weights to these units, $R_i \neq R_j$ [Eq. (11)], (b) lifting the degeneracy in the outcoming weights from these units, $\tilde{w}_i \neq \tilde{w}_j$ in the case of an $N{:}M{:}1$ architecture (see Fig. 3) or by choosing $w_{ij} \neq w_{kl}$ in the case of $N{:}M{:}P{:}1$ (see Figs. 5 and 6).

VI. MORE STRUCTURED WEIGHTS

The extension of the analytical results from one-component weights between the input units and the first hidden layer [see Eq. (11)] seems to be possible in some limited cases. However, the full analytical treatment for any set of weights, W_{ij}, and for any gain factor is beyond our ability and was examined mainly numerically and only within the framework of tanh activation functions. In order to simplify the following discussion let us distinguish between the following two major classes of $N{:}M{:}1$ systems (with $\tilde{w}_i=1$).

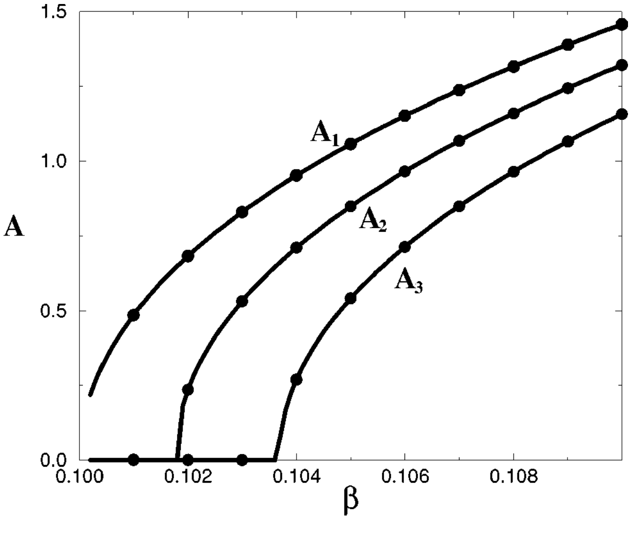

FIG. 6. A_i vs β for the architecture $N{:}3{:}2{:}1$ with sin activation function. The weights from the first hidden layer to the second one, $\{w_{ij}\}$, are given by $w_{1j}=1, w_{21}=1, w_{22}=0.9, w_{23}=0.8$. Analytical solution for A_i, Eqs. (27)–(32) (solid lines) and simulations of this network with $N=1000$ (●).

A. Nonoverlapping power spectrum

The power spectrum of the weights of any pair of hidden units does not contain a common wave number with a nonzero amplitude (or even almost the same nonzero wave number). More precisely, let us define the power spectrum of the weights to the rth hidden units to be diluted and to consist of only the following r_m nonzero components $\{K_{r_1},K_{r_2},\ldots,K_{r_m}\}$ with the following constraint: $|K_{r_m}-K_{s_n}|\gg 1$ for any pair of hidden units r and s (and also for $r=s$).

The prototypical case of this class is the architecture $N{:}M{:}1$ where the weights for each one of the hidden units consist of only one nonzero component in the power spectrum

$$W_{ij}=R_i\cos\left[\frac{2\pi K_i j}{N}-\pi\phi_i\right], \tag{33}$$

which is the generalization of the pure cos case, Eq. (11). The wave numbers $\{K_i\}$ are chosen to be relatively prime.

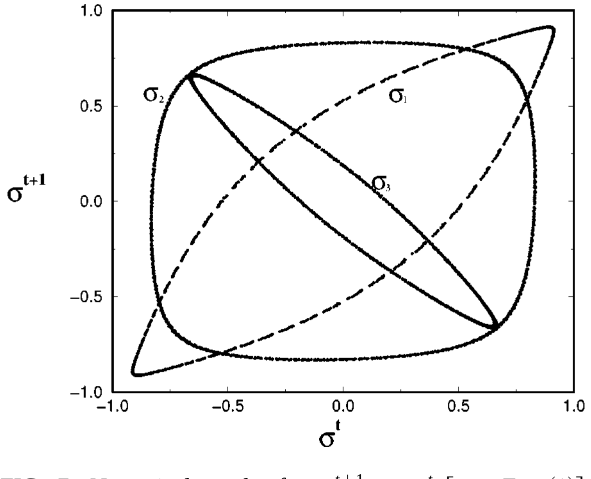

FIG. 7. Numerical results for σ^{t+1} vs σ^t [see. Eq. (1)] for $N{:}3{:}1$ with $N=500$, $\beta_1 \sim 1.65\beta_c$ and $K_i=31,117,231$, $\phi_i\sqrt{2}=0.1,0.05,0.01$, $R_i=1.0,0.9,0.8$ for $j=1,2,3$.

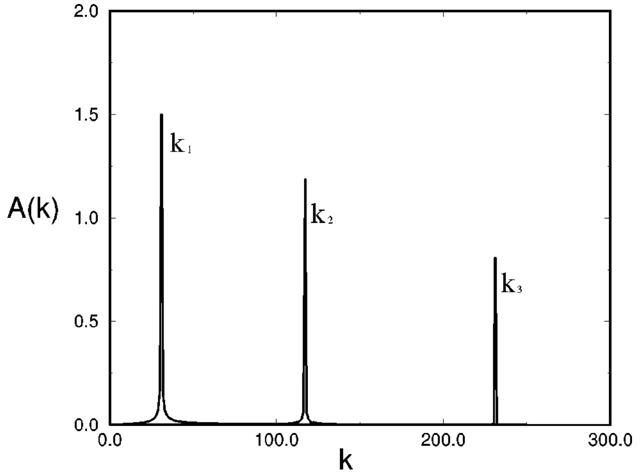

FIG. 8. The power spectrum of the output of N:3:1 defined in Fig. 7.

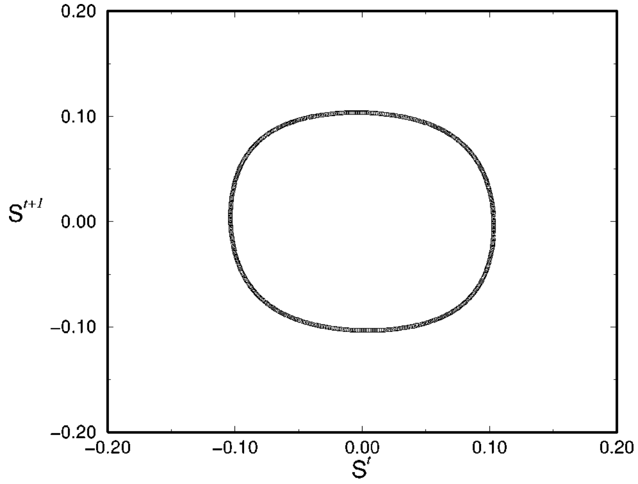

FIG. 9. The N:2:1 networks are classified in the two dimensional space $D_1=R_{21}/R_{11}$ and $D_2=R_{12}/R_{22}$. The presented simulations are for $D_1=D_2=0.6$, $N=512$, $\beta_1=1.1\beta_c$, $K_1=177$, $K_2=131$.

Results for the stationary solution are similar to that of Eq. (12), but with the following modifications. The critical gain, for each one of the hidden units, is a function of both the amplitude and the phase, $\beta_c^i=\beta_c^i(R_i,\phi_i)$. For the case N:3:1, for instance, one can show that

$$\beta_c^i=\sqrt{\frac{2}{NR_i}\,\frac{\pi\phi_i}{\sin(\pi\phi_i)}}, \tag{34}$$

where in general the critical gain increases with the absolute value of ϕ. Second, for $K\gg 1$ one can show that a phase shift, ϕ, in the weights results in a frequency shift,

$$K\rightarrow K-\phi \tag{35}$$

in Eq. (12) [with some higher harmonic corrections of $O(1/K)$]. Since a random ϕ is irrational, the flow is now quasiperiodic, $d_A=1$, instead of periodic as for the $\phi=0$ case, $d_A=0$. Each one of the hidden units becomes nonzero at a different gain following Eq. (34), and acts as an independent oscillator. Note that since there is an interplay between the phase and the amplitude, $\beta_c=\beta_c(R,\phi)$, the first hidden unit that undergoes a transition *is not necessarily the one with the largest amplitude*. Numerical results for N:3:1 with $N=500$, and $K_i=31,117,231$, $\phi_i\sqrt{2}=0.1,0.05,0.01$ and $R_i=1.0,0.9,0.8$ for $i=1,2,3$ are presented in Figs. 7 and 8. For each one of the hidden units the $d_A=1$. This attractor is characterized by a dominating peak of the power spectrum at $K_i-\phi_i$ with additional higher harmonic terms.

Note that since the power spectrum, P_K, of $\cos[2\pi(K_j-\phi)/N]$ decays asymptotically as $P_{K-K_j}\propto 1/|K-K_j|$, the constraint that $|K_i-K_j|\gg 1$ is necessary for each hidden unit to behave as an independent oscillator. This is indeed the case for finite M, $N\rightarrow\infty$ and where the power spectrum of each one of the hidden units consists of only a *finite random number* of components with nonzero amplitudes. In such a realization both K_{r_m} and $K_{r_m}-K_{s_n}$ are of $O(N)$.

B. Overlapping power spectrum

The power spectrum of at least one pair of hidden units has some common components with nonzero amplitudes. It is clear that the case of random weights belongs to this class. However, let us first analyze analytically the following prototypical case. The architecture is N:2:1 and the weights for each one of the two hidden units consist of only two nonzero pure cos [see Eq. (11)] with the wave numbers K_1 and K_2, which are relatively prime. The four amplitudes are R_{ij}, where the index i labels the hidden unit and j indicates the wave number. One can show that the critical gains for the two hidden units are given by

$$\beta_c^1=\sqrt{\frac{2}{N(R_{11}+R_{21})}},\quad \beta_c^2=\sqrt{\frac{2}{N(R_{22}+R_{12})}}. \tag{36}$$

It is clear that in the case that K_1, for instance, dominates the power spectrum of both the weights for the first and the second hidden units, the power spectrum of the time series generated by the network consists of only one nonzero component K_1 (plus higher harmonic terms). Both hidden units are locked onto K_1. For general amplitudes R_{ij}, one can run iteratively the equations for the amplitudes of the solutions, $\{A_{ij}^{q+1}\}$ as a function of $\{A_{ij}^q\}$, similar to Eqs. (10)–(12). More precisely, the time series is given by

$$S^t=\tanh\left\{\beta\sum_{i=1}^{2}\tanh\left[\sum_{j=1}^{2}A_{ij}\cos\left(\frac{2\pi K_{ij}t}{N}\right)\right]\right\} \tag{37}$$

and the iterative equations for the A_{ij} are given by

$$A_{mn}^{q+1}\cos(K_n t)=\beta R_{mn}\sum_{j=1}^{N}\cos\left(\frac{2\pi K_n j}{N}\right)\tanh\left\{\beta\sum_{i=1}^{2}\tanh\left[\sum_{l=1}^{2}A_{il}^{q}\cos\left(\frac{2\pi K_l(t-j)}{N}\right)\right]\right\}. \tag{38}$$

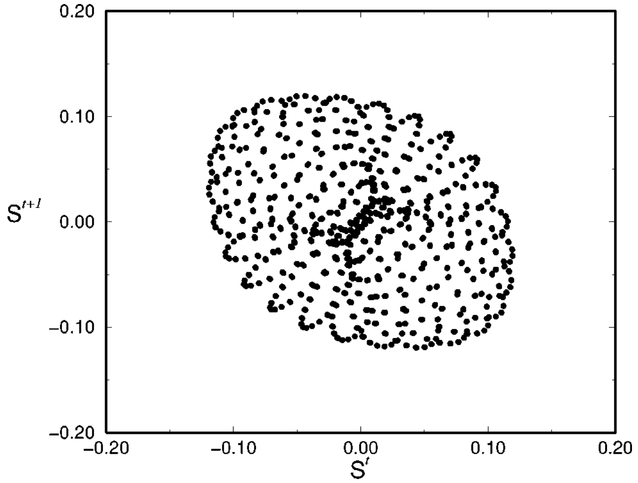

FIG. 10. The same as Fig. 9, but with $D_1 = D_2 = 0.3$.

The iterative solution of Eq. (38) indicates that the two-dimensional space $D_1 = R_{21}/R_{11}$ and $D_2 = R_{12}/R_{22}$ splits into the following two regimes. In the first regime there is only one attractor in which the two hidden units (and the output) are locked onto one of the components, K_1 or K_2. Hence, the number of nonzero components in the power spectrum of each one of the hidden units and that of the output is equal to one. In the second regime each one of the hidden units follows both K_1 and K_2, and hence there are two nonzero components in the power spectrum. (Note that in simulations in a subspace of the second regime it was found that both of the attractors with one or two nonzero components exist.) A result of a simulation of 512:2:1 in the first regime is presented in Fig. 9, where the power spectrum of the time series generated by the output consists of one nonzero component. A result of simulation in the second regime is presented in Fig. 10, where the power spectrum of each one of the hidden units consist of two nonzero components.

A similar picture occurs where a pure cos is replaced by one component in the power spectrum, Eq. (33). For the regime where both the hidden units are locked onto one of the components the d_A is equal to one, and in the second regime the d_A of both the hidden units and the output unit (the time series) is equal to two. Note that in contrast to the nonoverlapping case where each hidden unit behaves as an independent oscillator with $d_A = 1$, here the $d_A = 2$ for each one of the hidden units and for the output, and furthermore the hidden units undergo a transition to a nonzero amplitude *simultaneously at the same gain*. Results of simulations for N:2:1 with $N = 512$, $K_i = 131{,}177$, $\phi_{11}\sqrt{2} = 0.1$, $\phi_{12}\sqrt{2} = 0.9$, $\phi_{21}\sqrt{2} = 0.2$, $\phi_{22}\sqrt{2} = 0.4$, $\beta = 5\beta_c$ for two different sets of amplitudes are presented in Fig. 11 and Fig. 12. In Fig. 11 the $d_A = 2$ for each one of the two hidden units, where in Fig. 12 the $d_A = 1$.

Note that the critical gain is given by $\beta\beta_1 \propto 2/N$ [see Eq. (16)], where in simulation β was fixed to be of $O(1)$ and only β_1 was increased. This was done in order to enlarge the regime of the gain where the output is far from saturation, *output*$\to 1$, which is the bit-generator case [12]. Simulations indicate that the above picture, that the maximal $d_A = 2$, holds for $\beta_1 \gg 50\beta_c$, for $N = 500$.

The generic time series generated by the output of a network N:M:1 with *random* weights W_{ij} (without bias $P_0 = 0$) is similar to the overlapping two-component case in the following sense. As the gain β increases, some of the hidden units undergo a transition to their common dominated wave number K_1, for instance. The d_A of the output is one. (The equation for the critical gain is similar to Eq. (36) but the effect of ϕ and that of higher harmonic terms in the weights have to be taken into account). As the gain increases, it is plausible that a second wave number is taking place and the $d_A = 2$. Note that the scenario in which each one of the hidden units acts as an independent oscillator is found to be very rare in the case of random weights.

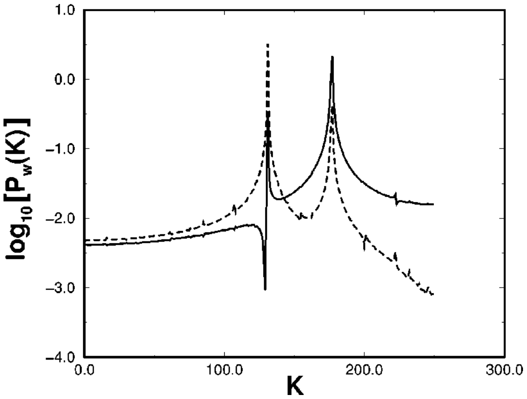

FIG. 11. The logarithm of the power spectrum of the local fields $\beta\Sigma_{j=1}^{N} W_{ij} S_j$, [see Eq. (1)], measured in simulations of N:2:1 with $N = 500$, $K_i = 131{,}177$, $\phi_{11}\sqrt{2} = 0.1$, $\phi_{12}\sqrt{2} = 0.9$, $\phi_{21}\sqrt{2} = 0.2$, $\phi_{22}\sqrt{2} = 0.4$, $R_{11} = 1.0$, $R_{12} = 0.2$, $R_{21} = 0.1$, $R_{22} = 1.0$, and $\beta = 5\beta_c$. The dashed line is for the first hidden unit and the full line is for the second one.

VII. CONCLUSIONS

The properties of time series generated by multilayer networks consist of one and two hidden layers, are studied analytically and numerically. The detailed analytical treatment is limited to the architectures N:2:1, N:3:1, and N:3:2:1. The main results at high gains but far from saturation where the

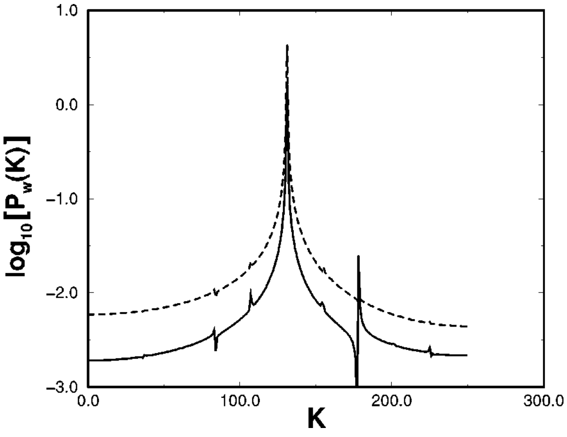

FIG. 12. The logarithm of the power spectrum of the local fields $\beta\Sigma_{j=1}^{N} W_{ij} S_j$, [see Eq. (1)], measured in simulations for N:2:1 with $N = 500$, $K_i = 131{,}177$, $\phi_{11}\sqrt{2} = 0.1$, $\phi_{12}\sqrt{2} = 0.9$, $\phi_{21}\sqrt{2} = 0.2$, $\phi_{22}\sqrt{2} = 0.4$, $R_{11} = 0.8$, $R_{12} = 0.2$, $R_{21} = 0.3$, $R_{22} = 1.0$, and $\beta = 5\beta_c$. The dashed line is for the first hidden unit and the full line is for the second one.

output is almost 1 are as follows:

(a) The d_A is only a function of the number of hidden units in the first hidden layer. More precisely, the d_A increases with the gain β and is bounded by the number of hidden units in the first layer (at least far from saturation). (b) Translating solution schematically mirrors the architecture of the network itself: weights in the lowest level are acted upon by the activation function of the first hidden units, and in turn acted upon by the activation function of the second hidden units, etc. (c) Increasing the number of hidden layers changes the critical gain dramatically. In general, the critical gain is $\propto N^{-1/(\delta+1)}$, where δ is the number of hidden layers. (d) In the case of nonoverlapping power spectrum, each hidden unit is an independent oscillator with a $d_A=1$. The d_A of the whole network is equal to the sum of independent oscillators. (e) For overlapping power spectrum there are two possible scenarios. In the first scenario, the hidden units are locked onto one of the dominated common components of the power spectrum such that the d_A of each hidden unit and that of the output is equal to one. In the second scenario, the d_A of each hidden unit and that of the output is equal to two (besides a plausible attractor with $d_A=1$).

There are still many questions to be answered, in particular the nature of the solution at high β near saturation and in particular with periodic activation functions like sin.

ACKNOWLEDGMENTS

We thank D. Kessler and W. Kinzel for fruitful discussions. The support of the Israel Academy of Sciences is acknowledged.

[1] T. L. H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).

[2] W. Kinzel and M. Opper, in *Physics of Neural Networks III*, edited by E. Domany, J. L. van Hemmen, and K. Schulten (Springer, Berlin, 1996).

[3] *Time Series Prediction*, edited by A. S. Weigand and N. A. Gershenfeld (Addison-Wesley, New York, 1994).

[4] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control* (Prentice-Hall, Englewood Cliffs, NJ, 1994).

[5] E. Eisenstein, I. Kanter, D. Kessler, and W. Kinzel, Phys. Rev. Lett. **74**, 6 (1995).

[6] M. Schroeder, W. Kinzel, and I. Kanter, J. Phys. A **29**, 7965 (1996).

[7] I. Kanter, D. Kessler, A. Priel, and E. Eisenstein, Phys. Rev. Lett. **75**, 2614 (1995).

[8] F. Takens, in *Dynamical Systems and Turbulence*, edited by D. A. Rand and L. S. Young, Lecture Notes in Mathematics No. 898 (Springer-Verlag, Berlin, 1981).

[9] T. Sauer, A. Yorke, and M. Casdagli, J. Stat. Phys. **65**, 579 (1991).

[10] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover Publications, New York, 1970), p. 804.

[11] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Ref. [10]), p. 361.

[12] M. Schroeder and W. Kinzel (unpublished).