Annals of Mathematics and Artificial Intelligence **39:** 315–332, 2003.

Time series generation by recurrent neural networks

A. Priel [a,b,*] and I. Kanter [a]

[a] *Department of Physics and Minerva Center, Bar-Ilan University, 52900 Ramat-Gan, Israel*
E-mail: priel@mail.biu.ac.il; ido@kanter1.ph.biu.ac.il
[b] *Center for Brain Sciences, Bar-Ilan University, Ramat-Gan, Israel*

The properties of time series, generated by continuous valued feed-forward networks in which the next input vector is determined from past output values, are studied. Asymptotic solutions developed suggest that the typical stable behavior is (quasi) periodic with attractor dimension that is limited by the number of hidden units, independent of the details of the weights. The results are robust under additive noise, except for expected noise-induced effects – attractor broadening and loss of phase coherence at large times. These effects, however, are moderated by the size of the network N.

Keywords: recurrent neural networks, time series, asymptotic properties, non-linear dynamical systems, attractor dimension, stochastic processes

1. Introduction

The construction of neural networks with complex time dynamics poses an important and interesting challenge. Progress in this field is crucial for understanding real-life intelligence and information processing [21]. Comprehensive studies of such networks have begun only recently and very little knowledge of an analytic nature has as yet been achieved.

In the well-studied cases of symmetric and feed-forward networks, there is no non-trivial dynamic behavior [6,20]. The central issue is rather the dependence of the output (or equilibrium) state on a set of random inputs. Here the input to the network is constantly changing in response to the current state of the network and so the issue of the dynamic behavior of the network rises to the fore. Hence a new repertoire of analytical techniques and concepts is required to explore the implications of different architectures and feedback schemes for the behavior of the network.

In this paper we shall review some of the results of an analytic study of a particular class of networks, the Sequence Generator (SGen), which is in essence a perceptron (or multi-layer network) with a continuous output, whose 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. Denoting the N inputs at time t by the vector $\mathbf{S}^t = (S_1^t, \dots, S_N^t)$,

* URL: http://www.biu.ac.il/faculty/priel

where $S_i^t \in \mathcal{R}$ and $\mathbf{S}^t \in \mathcal{R}^N$, the process can be summarized as follows:

$$S_1^{t+1} = S_{\text{out}}^t; \qquad S_i^{t+1} = S_{i-1}^t, \quad i = 2, \ldots, N. \tag{1}$$

Here, S_{out}^t, the output at each time step is determined from the inputs at that time and the fixed weight vector $\mathbf{W} \in \mathcal{R}^N$, via non-linear activation function $\mathcal{F}$. We restrict the discussion to bounded, symmetric non-linear transfer functions,

$$f : \mathcal{R} \to \mathcal{R}, \qquad |f(x)| < \infty \quad \forall x \in \mathcal{R},$$

where in the case of a perceptron we get

$$S_{\text{out}}^t = \mathcal{F}\left[\beta\left(\mathbf{W} \cdot \mathbf{S}^t\right)\right], \quad \beta \in \mathcal{R}. \tag{2}$$

In the sequel we shall assume a hyperbolic-tangent function as well as other functions. In the limit where the coupling ($\beta > 0$) goes to ∞, this network reduces to the case of the bit-generator, with discrete binary inputs and outputs, ± 1, see [5].

The rational behind using time delayed vectors as inputs is based on the theory of state space reconstruction of a dynamic system using delay coordinates [15]. This theory addresses the problem of reproducing a set of states associated with the dynamic system using vectors obtained from the measured time series, and is widely used for time series analysis.

The dynamics described above is also known as the 'non-linear autoregressive model' (NAR) used for time-series prediction and analysis. In fact, the NAR model includes an additive noise, see section 5. Some of their stationary statistical properties have been analyzed by H. Tong [18] and others. The computational capabilities of such models were found to be equivalent to fully recurrent networks [17]. A similar dynamical system, the generalized shift map, can be viewed as a special case of the noiseless NAR model in which piecewise linear mappings are used, e.g., [9]. It has also been shown that this class of models may exhibit a super-Turing computational power. Nevertheless, these papers and others do not discuss the long-term behavior of such models or its implications. A similar architecture incorporating time delays is the TDNN – time-delay neural network with recurrent loops [19]. This type of network is known to be appropriate for learning temporal sequences, e.g., speech signal. In the context of time series, it is mostly used for short term predictions. The aim of our investigation is the understanding and characterization of the long term sequences generated by such architectures, and the time scale to reach this asymptotic behavior.

The class of SGen networks has been analyzed extensively, mainly at asymptotic time. The SGen exhibits quasi-periodic dynamics, i.e., a limit cycle with irrational (or non-commensurate) phase that leads to a cycle of infinite length. This dynamics is governed by a trajectory (a sequence of states) whose limit is called an attractor. Attractors are typically classified into four major classes: a fixed point, a limit cycle (periodic orbit, or fixed point of order n), a quasi-periodic orbit and a chaotic attractor. An important quantity that characterizes the attractor is its dimension. This quantity has several definitions, see [1] for a review. Non-chaotic dynamics is characterized by an integer attractor dimension, where the strictly periodic attractor has dimension zero. Quasi-periodic attractors have a positive dimension whose measure depends on the number of independent variables of the system. The attractor dimension of a sequence generated by a SGen depends on the architecture and control parameters [4,7,10,13]. It may also generate chaotic behavior whose nature depends on the activation function [11,12]. It has been shown that monotonic functions, e.g., hyperbolic-tangent, generate fragile chaos only, i.e., sensitive to deviations in parameter space, whereas non-monotonic functions may generate high-dimensional robust (stable) chaos. We shall not elaborate on this behavior here. Recent studies also concentrated on the issue of the dynamics of interacting neural networks [8].

In the next section we define the architectures and their dynamic rules. In sections 3, 4 we report the main results regarding the asymptotic behavior of the time-series generated for perceptron-SGen and MLN-SGen. We will also explore extensions of this SGen to other activation functions. In section 5 we present the results in the case of a noise added to the output of the model. Finally, we shall discuss the implications of the results on the issue of time series prediction in general and the relation between dynamics and computation in particular.

2. The model

In the following we examine both single and multi-layer perceptron-SGen's. The architectures differ in the number of feed-forward layers between the input and the output. The only feedback considered is described in equation (1). We may think of the network as generating an infinite sequence of output numbers (see figure 1), S^t, starting from the initial inputs, $S_i = S_i^0$ via the dynamic equation (for the perceptron)

$$S^t = \mathcal{F}\left[\beta\left(\sum_{j=1}^{N}(W_j + W_0)S^{t-j}\right)\right], \tag{3}$$

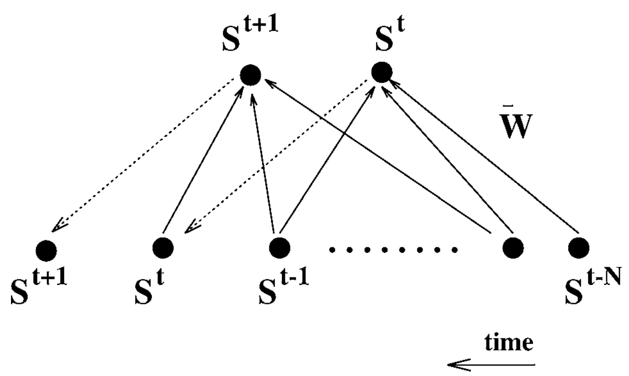

Figure 1. Perceptron-SGen generating a time series.

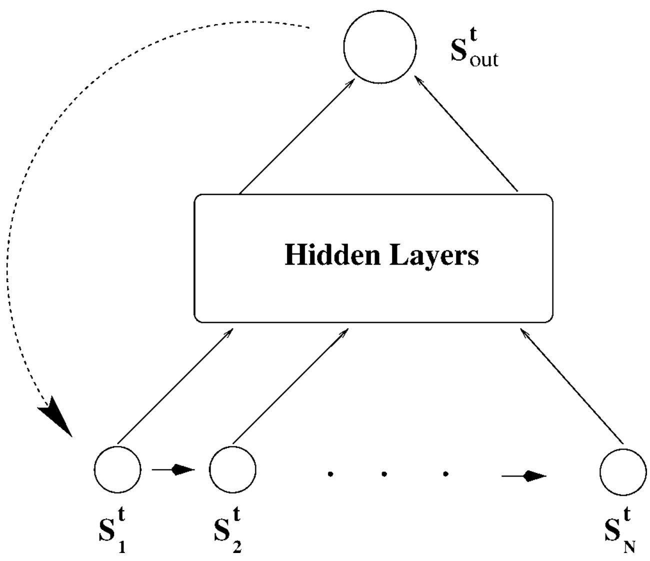

Figure 2. MLN-SGen generating a time series.

where W_0 is a possible bias of the weights. This is actually the origin of the name ‘sequence generator’ given to this machine.

In the more general case of a two-layer network $N : M : 1$ the dynamic equation is given by:

$$S^t = \mathcal{F}\left(\beta \sum_{h=1}^{M} (w_h + w_0)\mathcal{F}\left[\beta_1 \left(\sum_{j=1}^{N} (W_{hj} + W_{h0}) S^{t-j}\right)\right]\right), \tag{4}$$

assuming the same activation functions for all units; the second layer’s function may be linear. Figure 2 depicts the general architecture with an arbitrary number of hidden layers.

Throughout the paper we use interchangeably the notation S^{t-j} and S_j^t to denote the time series and the state of the jth input, respectively.

3. Single layer perceptron-SGen

3.1. Hyperbolic-tangent function

We begin with a study of the simplest feed-forward network, the single layer perceptron. The strategy for developing the stationary solution (or attractor) for the system described below is based on the method of linear stability analysis (see, e.g., [3]). In this method, each type of dynamics assumed is analyzed in the vicinity of its onset where dynamic stability is tested.

Following equation (2) we analyze a perceptron with a monotonic activation function ($\mathcal{F} = \tanh$) whose output S_{out} at time step t is given by:

$$S^t_{\mathrm{out}} = \tanh\left(\beta\left[\sum_{j=1}^{N}(W_j + W_0)S^t_j\right]\right), \tag{5}$$

where β is a gain parameter, N is the input size and W_0 is the bias term. Since the (stable) asymptotic behavior is quasi-periodic, its mathematical description is influenced primarily by one of the larger Fourier components in the power spectrum of the weights vector W of the perceptron. This observation motivates the following representation of the vector W. Let us start with the case of a vector that consists of a single Fourier component of the form:

$$W_j = a\cos(2\pi Kj/N), \quad j = 1, \ldots, N; \qquad W_0 = 0, \tag{6}$$

where a is a constant and K is a positive integer. Note that the vector $\mathbf{W}$ can always be represented as a Fourier decomposition of its values. It is straightforward to show that the only consistent fixed point (f.p.) solution is $S_j = S^* = 0\ \forall j$ by inserting $S_j = S^*$ back to equation (5). To study the linear stability of this f.p., namely, the onset of a periodic attractor, we rewrite the dynamic equation (equation (5)) as an N-dimensional map, as follows:

$$\left\{\begin{array}{l} S^{t+1}_1 = f_1(\mathbf{S}^t, \mathbf{W}) = \tanh\left(\beta\left(\sum_{j=1}^{N} W_j S^t_j\right)\right); \\ S^{t+1}_2 = f_2(\mathbf{S}^t, \mathbf{W}) = S^t_1; \\ \quad\vdots \\ S^{t+1}_N = f_N(\mathbf{S}^t, \mathbf{W}) = S^t_{N-1}; \end{array}\right\} \tag{7}$$

for each of the variables, S_j, from which we derive the Jacobian matrix at the fixed point:

$$\left[\frac{\partial f_i}{\partial S_j}\right]_{S_j=S^*} = \begin{bmatrix} \beta W_1 & \beta W_2 & \ldots & \beta W_{N-1} & \beta W_N \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}. \tag{8}$$

Noting that the eigenvalues of this matrix determine the stability of the f.p., we can write the characteristic equation of the Jacobian matrix, where for the above weights we obtain:

$$(-1)^{(N-1)}\lambda^N\left[\beta\sum_{n=1}^{N}\frac{a\cos(2\pi Kn/N)}{\lambda^n} - 1\right] = 0, \tag{9}$$

where λ is the eigenvalue. The transition to a periodic phase occurs when $|\lambda| = 1$ or $\lambda = e^{i\varphi}$ which gives rise to two coupled equations that determine φ and β_c – the value of β at the transition point

$$\sum_{n=1}^{N} \cos\left(\frac{2\pi}{N} Kn\right) \sin(n\varphi) = 0, \qquad a\beta_c \sum_{n=1}^{N} \cos\left(\frac{2\pi}{N} Kn\right) \cos(n\varphi) = 1. \tag{10}$$

Following the solution of the coupled equations, we find the onset of an N-cycle:

$$\varphi = \pm \frac{2\pi}{N} K \quad \Rightarrow \quad \beta_c = \frac{2}{aN}. \tag{11}$$

This analysis gives us only the linear term, i.e., near the transition. In order to continue, we can write a general periodic function based on a linear superposition of powers of the basic periodic term, $\cos(\frac{2\pi}{N} Kt)$ and find its coefficients, however this method becomes complicated for orders higher than three. Hence, we shall elaborate on a solution which is based on the above analysis up to the third order term, whereas to obtain the higher order terms we assume the following relation:

$$S^t = \tanh\bigl[A(\beta) \cos(2\pi Kt/N)\bigr], \tag{12}$$

where the coefficients are determined from the demand for self-consistency in equation (5), rewritten in the form of an infinite sequence:

$$S^t = \tanh\left(\beta \sum_{j=1}^{N} W_j S^{t-j}\right). \tag{13}$$

Plugging the proposed solution, equation (12), in the right-hand side of equation (13) and expanding the innermost hyperbolic-tangent function in powers of its argument, we obtain an equation for $A(\beta)$, the amplitude of the resulting sequence S^t, as follows:

$$A = \frac{Na}{2} \beta \sum_{\rho=1}^{\infty} M(\rho)(A/2)^{2\rho-1}(\rho!)^{-2}, \quad M(\rho) = 2^{2\rho}\bigl(2^{2\rho} - 1\bigr)\mathcal{B}_{2\rho}, \tag{14}$$

where $\mathcal{B}_{2\rho}$ are the Bernoulli numbers (see [2], pp. 804–806). This equation is a direct result of the demand for self-consistency. Analysis of equation (14) reveals the onset of the periodic phase, equation (11), as well as the exact behavior of the (asymptotic) sequence generated by the SGen for a given choice of the parameters a, β. Figure 3 depicts the analytical result (solid line) and simulation (circles) of this network for the amplitude $A(\beta)$. The simulations where done for $N = 100$, $K = 17$, $a = 1.5$.

We note that this behavior of the dynamics, namely the transition from a fixed point to a limit cycle, is known as Hopf bifurcation [3].

Let us continue with the case of a vector that consists of a single biased Fourier component of the form described in equation (6) but taking $W_0 = b$, i.e.,

$$W_j = a\cos(2\pi Kj/N), \quad j = 1, \ldots, N; \qquad W_0 = b. \tag{15}$$

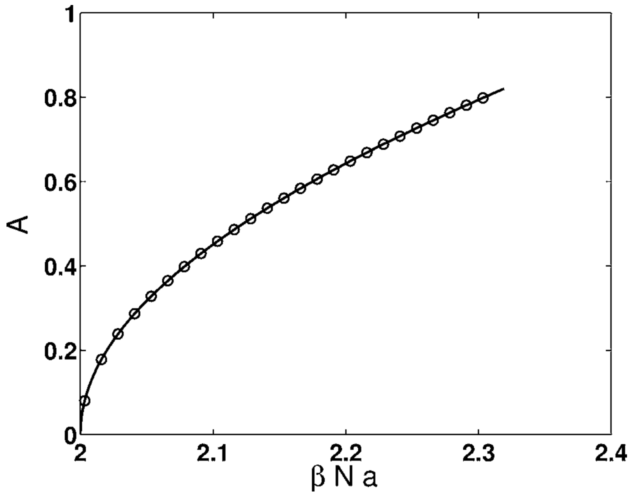

Figure 3. The amplitude A vs. the scaled β obtained from equation (14) (solid line) and from simulating the perceptron SGen (circles) whose weights follow equation (6) and $N = 100$, $K = 17$, $a = 1.5$.

Since the analysis is similar to that described above, we shall not repeat the derivation (one can refer to [10], ch. 2), but rather present the main results. The stationary solution for the sequence (S^t) for this choice of the weights (equation (15)), can be shown to be of the form:

$$S^t = \tanh\bigl[A(\beta)\cos(2\pi Kt/N) + B(\beta)\bigr]. \tag{16}$$

There are two possible stable non-zero solutions for the variables (A, B):

$$\begin{aligned} A &= \frac{1}{2}\beta Na\sum_{\rho=1}^{\infty} M(\rho)(A/2)^{2\rho-1}(\rho!)^{-2}; \qquad B = 0,\\ B &= \beta Nb\sum_{\rho=1}^{\infty} M(\rho)B^{2\rho-1}\bigl((2\rho)!\bigr)^{-1}; \qquad A = 0. \end{aligned} \tag{17}$$

Analysis of equations (17) reveals the following behavior as a function of the parameter β. Each of the variables is the amplitude of an attractor. The attractor represented by $(A \neq 0,\ B = 0)$ is the same limit cycle described above, while the attractor represented by $(B \neq 0,\ A = 0)$ is a fixed point of the dynamics. The onset of each of the attractors A (B) is at $\beta_{c1} = 2(aN)^{-1}$ $(\beta_{c2} = (bN)^{-1})$, respectively. One can identify three regimes: (1) $\beta < \beta_{c1,c2}$ – the stable solution is $S^t = 0\ \forall t$; (2) $\min(\beta_{c1}, \beta_{c2}) < \beta < \max(\beta_{c1}, \beta_{c2})$ – the system flows for all initial conditions into the attractor whose β_c is smaller; (3) $\beta > \beta_{c1,c2}$ – depending on the initial condition of the input vector, the system flows into one of the attractors, namely, the stationary state is either a fixed point or a periodic flow.

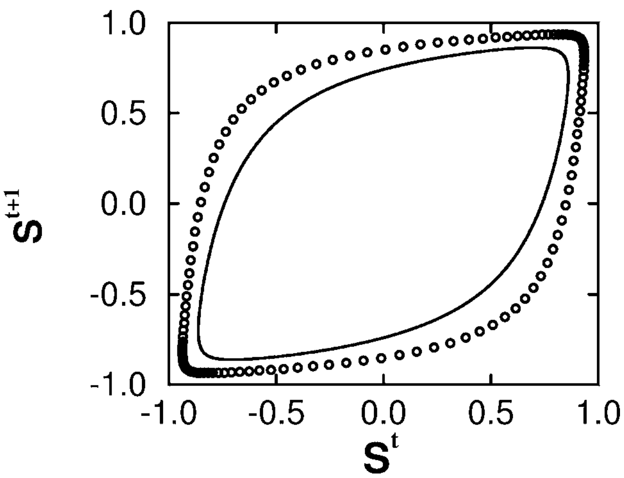

Figure 4. Two-dimensional embedding of a sequence generated by a perceptron whose weights follow equations (15), (18). The parameters for the periodic sequence are $N = 128$, $K = 17$, $b = 0.3$, $\beta = 1/40$ and for the quasi-periodic $K = 17$, $\phi = 0.123$, $\beta = 1/45$, respectively.

Next we discuss the more general case where the weights of equation (15) include an arbitrary phase shift of the form:

$$W_j = a\cos(2\pi Kj/N - \pi\phi), \quad \phi \in (-1, 1). \tag{18}$$

In this case, we derive only an approximate solution. The leading term of the stationary solution in the limit $N \gg 1$ is of the form:

$$S^t = \tanh\bigl[A(\beta)\cos\bigl(2\pi(K-\phi)t/N\bigr) + B(\beta)\bigr], \tag{19}$$

where the higher harmonic corrections are of $\mathcal{O}(1/K)$. A note should be made here that the phase shift in the weights is manifested as a frequency shift in the solution. In addition, the attractor associated with $A \neq 0$ is now a quasi-periodic flow in the generic case when ϕ is irrational. The onset value of the fixed point (β_{c2}) is the same as before, however the onset of the quasi-periodic orbit is $\beta_{c1} = \frac{\pi\phi}{\sin(\pi\phi)}2(aN)^{-1}$. The variables A, B are similar to equations (17):

$$\begin{aligned} A &= \beta Na\frac{\sin(\pi\phi)}{\pi\phi}\sum_{\rho=1}^{\infty} M(\rho)(A/2)^{2\rho-1}(\rho!)^{-2}; \qquad B = 0,\\ B &= \beta Nb\sum_{\rho=1}^{\infty} M(\rho)B^{2\rho-1}\bigl((2\rho)!\bigr)^{-1}; \qquad A = 0. \end{aligned} \tag{20}$$

The three regimes discussed above appear in this case as well. Figure 4 shows the attractor associated with ($A \neq 0$, $B = 0$) for the two cases where the series generated by the output is embedded as a sequence of two-dimensional vectors (S^{t+1}, S^t).

The general weights can be written as a combination of their Fourier components with different K's and ϕ's:

$$W_j = \sum_{i=1}^{m} a_i \cos(2\pi K_i j/N - \pi\phi_i), \quad \phi_i \in (-1, 1). \tag{21}$$

When the different K's are not integer divisors of each other, the general solution is similar to that described above:

$$S^t = \tanh\left[\sum_{i=1}^{m} A_i(\beta)\cos\bigl(2\pi(K_i - \phi_i)t/N\bigr) + B(\beta)\right], \tag{22}$$

where m is the number of relevant Fourier components. As above, the variables A_i, B are coupled via self-consistent equations. Nevertheless, the generic stationary flow is one of the possible attractors, depending on β and the initial condition; i.e., ($A_q \neq 0$, $A_i = 0$ $\forall i \neq q$, $B = 0$) or ($B \neq 0$, $A_i = 0$). At this stage we can conclude that the stable flow for the perceptron is one of three: a fixed point, periodic cycle or quasi-periodic flow. The first two have a zero dimension while the last describes a one-dimensional flow. We stress that more complex flows are possible even in our solution (e.g., equation (22)), however they require a special relation between the frequencies and a very high value of β, typically more than an order of magnitude greater than the bifurcation value [16]. As mentioned, a chaotic behavior generated by the dynamic system of the type described in equation (5) is described in [11,12].

3.2. *Other activation functions*

The results drawn above were extended to other odd functions, e.g., $\mathcal{F} = \sin$ in equation (2). The general results are similar although the details differ, see [7,10]. From the mathematical point of view, this function is more convenient to handle.

Figure 5 depicts the embedded sequence generated by the perceptron-SGen with sine function for the two gain values. Unlike the monotonic functions, e.g., the hyperbolic-tangent, the non-monotonic functions do not exhibit saturation as β increases, but rather the solution folds back in the bounded region. At higher β values one observes a period doubling route to chaos, even without a bias term in the weights [11,12] (unlike monotonic functions that require a non-vanishing bias term to generate chaos).

We note that the family of even functions is not considered here as they do not exhibit a Hopf bifurcation due to lack of a linear term; instead, one observes a sequence of bifurcations that may lead to chaotic behavior.

4. Multi-layer perceptron-SGen

The analysis of the perceptron-SGen in the previous section reveals a major drawback of this architecture that is manifested in the attractor dimension of the sequence

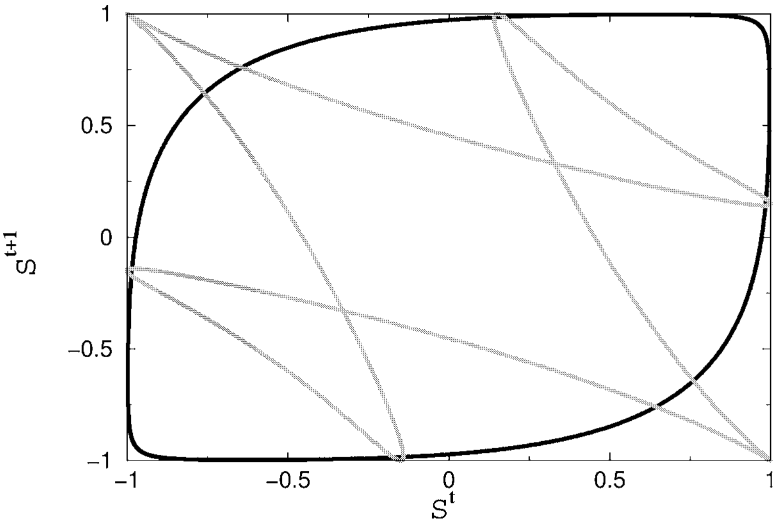

Figure 5. Embedding of a sequence generated by a perceptron-SGen with sine transfer function and weights defined by equation (18). Solid black line for $\beta = 1/25 \approx 1.35\beta_c$ and solid gray line for $\beta = 2/15 \approx 4.5\beta_c$. $N = 100$, $K = 17$, $\phi = 0.471$.

generated. The attractor dimension is one for a typical weight vector and is a characteristic of the architecture and not the complexity of the weights. The question on which we focus in this section is the interplay between the architecture, i.e., number of hidden units and layers, and the attractor dimension of the sequences generated. In the following, we shall concentrate on a 2LP-SGen with a linear output and hyperbolic-tangent transfer function in the hidden units. Applying the sine function or additional hidden layers does not change the fundamental results (see [4]).

The dynamics is given by equation (4) with the output defined as follows:

$$S^t_{\text{out}} = \sum_{m=1}^{H} R_m \tanh\left[\beta\left(\sum_{j=1}^{N} W_{jm} S^t_j\right)\right], \tag{23}$$

where H denotes the number of hidden units and the weights are bias-less. The time series generated by the network is similar to equation (13), and is given by

$$S^t = \sum_{m=1}^{H} R_m \tanh\left(\beta \sum_{j=1}^{N} W_{jm} S^{t-j}\right), \tag{24}$$

assuming the same gain β is used for all hidden units. One can easily extend the following analysis to different gain variables for each unit.

In the case of weight vectors that consist of a single component in the power spectrum of the form

$$W_{jm} = a_m \cos\left(\frac{2\pi}{N} K^m j\right), \quad j = 1, \ldots, N, \tag{25}$$

(where m labels the hidden unit and K^m denotes the wave number in the weight vector of the mth hidden unit), and under the assumption that each wave number is incommensurate to the other, we may treat each perceptron-SGen separately; hence, the asymptotic solution would be a linear combination of perceptron-SGen solutions type. The formal stationary solution to equation (24) with weights given by equation (25) (see [4,10]) is given by

$$S^t = \sum_{m=1}^{H} R_m \tanh\left[A_m \cos\left(\frac{2\pi}{N} K^m t\right)\right], \tag{26}$$

where A_m is the amplitude of the mth wave number. Inserting this equation in both sides of equation (24) and expanding the expression in the right-hand side gives, again, self-consistent equations for the amplitudes A_m, however they now become independent of each other (in contrast to the case of the perceptron-SGen)

$$A_m = \beta a_m N R_m \sum_{\rho_m=1}^{\infty} M(\rho_m)\left(\frac{A_m}{2}\right)^{(2\rho_m-1)} (\rho_m!)^{-2}. \tag{27}$$

Unlike the solution for the perceptron-SGen with general weights, all the H attractors can be stable simultaneously, (depending on the value of β) possibly leading to an H-dimensional attractor. Taking the limit $A_m \to 0$ in equation (27), we find the onset of each of the attractors to be as follows:

$$\beta_c^m = \frac{2}{a_m R_m N}. \tag{28}$$

Note again that each attractor's transition from a zero fixed point to a periodic attractor occurs independently (under the assumptions described above) of the other attractors, depending on the parameters of its associated hidden unit, a_m, R_m.

An example of the coexistence of two attractors is depicted in figure 6. The network consists of two hidden units, each with a weight vector of a single, phase-less Fourier component. The parameters were chosen such that both hidden units are above the bifurcation point. The dimension of the attractor is zero since the sequence is periodic; however, it is embedded on the surface of a 3D-torus.

In the case of weights that consist of a single (possibly biased) component with phase, i.e.,

$$W_{jm} = a_m \cos\left(\frac{2\pi}{N} K^m j + \pi\phi^m\right); \quad m = 1, \dots, H, \tag{29}$$

the stationary solution becomes a combination of the perceptron-like solutions:

$$S^t = \sum_{m=1}^{H} \tanh\left[A_m(\beta) \cos\left(\frac{2\pi}{N}\left(K^m + \phi^m\right)t\right) + B_m(\beta)\right], \tag{30}$$

where B_m is associated with the (possible) bias term. The variables A_m, B_m are the solution of the self-consistent coupled equations; however, by contrast with the single

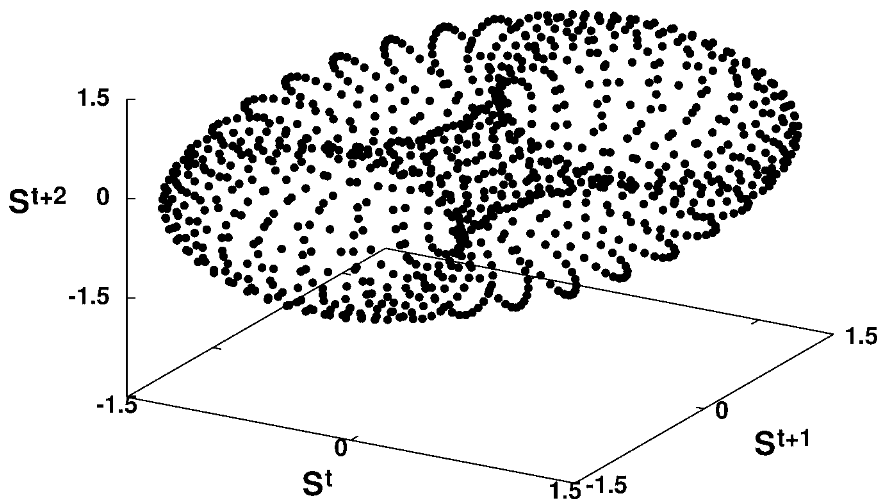

Figure 6. A 2LP-SGen with two hidden units generating a periodic attractor. The attractor is a combination of the corresponding two perceptron-SGen attractors. $N = 1024$, $K^1 = 71$, $K^2 = 211$, $\beta = 1.2\beta_c$.

perceptron, each hidden unit operates independently and can potentially develop an attractor of the type described above. The number of attractors depends on β with a maximum of H periodic attractors and possibly an additional fixed point attractor (for $W_0 \neq 0$). The number of non-zero A_m's defines the attractor's dimension in the generic case of irrational ϕ's associated with them.

Figure 7 shows the results of simulating the network at different gain values, $\beta = \frac{1}{40}, \frac{1}{33}$ and $\frac{1}{25}$. In the left-hand side of the figure, the attractor of the dynamics is embedded in a 2D-space, whereas in the right-hand side the associated power spectrum is given. As the gain increases, more components in the power spectrum have non-zero amplitude. Since these components have an irrational phase, the attractor fills the space, e.g., for $\beta = 1/25$ a 3D-attractor emerges.

Generalization of this result to more than a single biased Fourier component is straightforward. Each vector is of the form described in equation (21) plus an index for the hidden unit. The result is a combination of the general perceptron solution, equation (22). This solution is much more involved and the coupled equations are complicated, but their careful study reveals the same conclusion; namely, each hidden unit possesses a single dominant Fourier component (possibly with additional much smaller components associated with the other components in the vector). As the gain parameter β becomes larger, more components becomes available and the number of possible attractors increases. Particular extensions are discussed in [4].

5. Noisy SGen

Since a realistic time series is noisy, it is imperative that we understand the effect of noise on the output of the model. In this section we study the effect of an additive

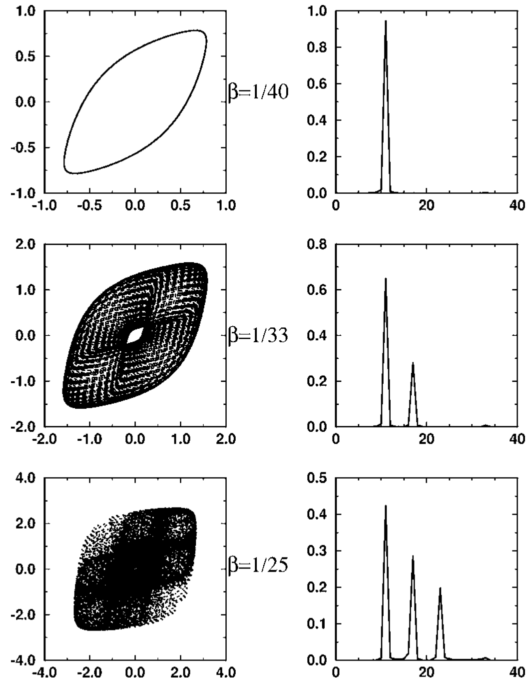

Figure 7. A 2LP-SGen with three hidden units and hyperbolic-tangent transfer function generating a quasi-periodic attractor. The attractor and the power spectrum of the sequence are presented for three values of the gain β. $N = 100$, $(K^1 = 11, \phi^1 = 0.176)$, $(K^2 = 17, \phi^2 = 0.431)$, $(K^3 = 23, \phi^3 = 0.585)$.

noise on the results presented in previous sections. The modified dynamic rule is given by

$$S_1^{t+1} = S_{\text{out}}^t + \text{noise}; \qquad S_j^{t+1} = S_{j-1}^t, \quad j = 2, \ldots, N, \tag{31}$$

where, as before, S_{out}^t is the network's output at time step t and S_j^t are the inputs at that time. Note that this is the complete NAR model mentioned earlier.

The problem of noise in a dynamic system is of great importance for the behavior of the system (e.g., stability), hence its implications on the time series measured from that system. In the classical theory of time series analysis (linear and non-linear), the prediction ability of a model is of interest when trained with noisy data. Since the intention is to use a SGen to reproduce noisy data, it is important to understand how noise affects the output of a generic SGen. In particular, it is crucial that the SGen be robust with the addition of noise, which is non-trivial given the non-linear feedback dynamics of the SGen.

The results described below were obtained for two kinds of updating rules, sequential and parallel, and are similar up to time rescaling. In the sequential scheme, noise is presented to the system's output in the following way:

$$S_1^{t+1} = S_{\text{out}}^t + \eta^t, \tag{32}$$

where the first two moments of η are:

$$E[\eta^t] = 0, \quad E[\eta^t \eta^{t'}] = \sigma^2 \delta_{tt'}. \tag{33}$$

In this way, noise is added only to the first unit in each iteration of the dynamic rule. In the parallel scheme, the system is actually composed of $N+1$ units since each input unit also functions as an output – the weight matrix has rank $N+1$ with zero diagonal ($W_{ii}=0$). The noise, then, is represented by a vector with $N+1$ independent components η^t, which is added to all units simultaneously in each iteration:

$$S_i^{t+1} = \tanh\left(\beta \sum_{j=1}^{N+1} W_{ij} S_j^t\right) + \eta_i^t, \tag{34}$$

where in this case $E[\eta_i^t \eta_{i'}^{t'}] = \sigma^2 \delta_{ii'} \delta_{tt'}$.

In the following we describe the results obtained for perceptron-SGen, though the general properties hold with an additional hidden layer. Let us consider the properties of a perceptron-SGen with a weight vector that consists of a single Fourier component with an arbitrary phase (ϕ) of the form given by equation (18). When no noise is added to the dynamic equation, equation (5), the generic stable solution is found to be a quasi-periodic orbit, e.g., equation (19) and figure 4. When noise is added, equation (32), the orbit is broadened, however the trajectory is confined in phase space. A characteristic quantity is the noise induced width of the broadened attractor which depends on the size of the system, N. Another important issue is that of phase coherence. A periodic system in the presence of noise typically exhibits a loss of phase coherence, manifested by a broadening of the power spectrum of the time series generated by the system.

To demonstrate the broadening effect we embed the data generated by a noisy perceptron-SGen in a two-dimensional space and measure the extent perpendicular to the local tangent, see figure 8. We found a power-law scaling between the average width and the size of the system of the form $\langle\omega\rangle \propto A/\sqrt{N}$, where A is a constant (see [14], equations (15)–(17)).

To understand the behavior of the phase, let us use the following definition. The temporal phases of a noisy system, denoted by $\Phi_n(t)$, and of an identical, noise-free, system by $\Phi_c(t)$, are defined by:

$$\Phi_c(t) = \sum_{i=0}^{t} \big(\phi_c(i+1) - \phi_c(i)\big), \qquad \phi_c(0) = \phi_n(0), \quad \Phi_n(t) = \sum_{i=0}^{t} \big(\phi_n(i+1) - \phi_n(i)\big), \tag{35}$$

where the phases $\phi_c(i)$, $\phi_n(i)$ are the relative phases of the ith clean/noisy embedded vectors in an arbitrary, but fixed, coordinate system (the geometrical interpretation of the phase is just the polar angle of an embedded 2D-vector, (S^t, S^{t-1})). Assuming identical

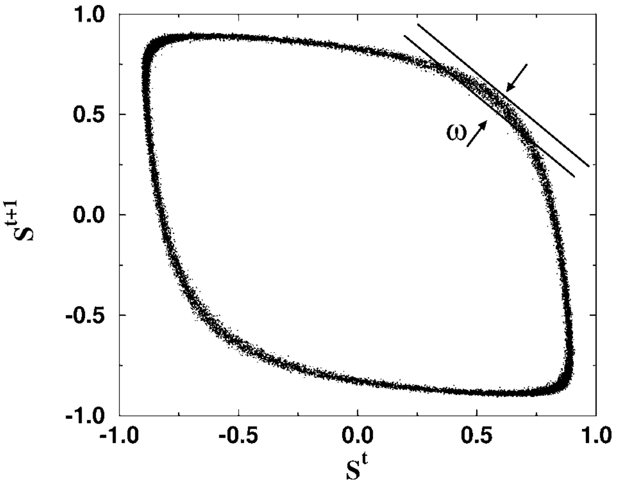

Figure 8. A broadened attractor generated by a perceptron-SGen, $N = 50$, $K = 17$, $\beta = 1/17$, $\phi = 0.123$ with a uniformly distributed noise of amplitude ± 0.1 added, equation (32). $\langle\omega\rangle$ denotes the width perpendicular to the attractor.

initial conditions of both systems, and using the fact that $E[\Phi_c(t) - \Phi_n(t)] = 0$, we define the statistical variance of the phase due to the added noise as follows:

$$\langle \Delta\Phi^2(t)\rangle = E\big[\big(\Phi_c(t) - \Phi_n(t)\big)^2\big], \tag{36}$$

where the expectation is assumed as an average over all samples taken after time t. Analysis of this quantity reveals a diffusive behavior (see [14], equations (20) and arguments thereafter), i.e., the variance of the phase scales linearly with time. As depicted in figure 9, we found in our system that the diffusion coefficient inversely scales with the size of the network squared,

$$D \sim \frac{1}{N^2}. \tag{37}$$

We conclude that the time scale associated with this process increases with the size of the system, therefore the system remains coherent over increasingly (quadratic) long times as N increases.

The above-mentioned results present several properties of the dynamics in the neighborhood of a single attractor. When there are multiple attractors available to the system, even for a perceptron-SGen, one expects noise-induced transitions between the domains (basins) of attraction of the possible attractors having local stability. One of the quantities of interest in this scenario is the average time it takes to escape from the domain of one of the attractors. We derived an analytical expression for the mean-first-passage-time (MFPT) in the limit of weak noise and a weakly non-linear function, supported by a series of simulations. We shall not elaborate on this issue further and one can refer to [10, chapter 4] and [14] for details. It has been found that the typical transition time between attractors scales exponentially with $\sqrt{N}$.

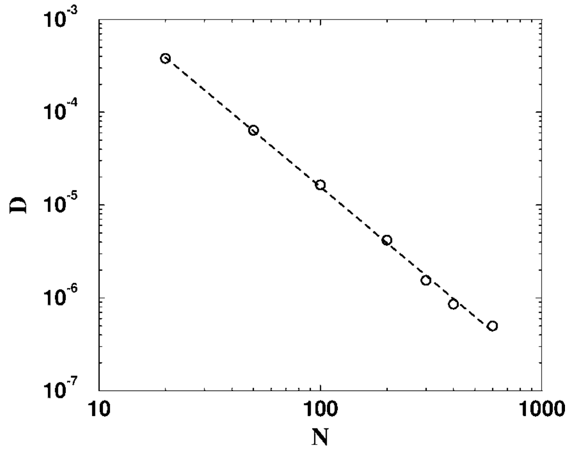

Figure 9. Diffusion coefficient obtained by simulating the system for several system sizes (over a decade). The linear regression (dashed line) is $D = 0.154/N^a$ with $a = 2 \pm 0.036$.

6. Discussion

This paper addresses several aspects arising from the dynamics of a special kind of recurrent neural network – a feed-forward network with a recurrent loop from the output to the input layer. In sections 3, 4 we introduce the theoretical results of the SGen in the stable regime of the parameters. We show that the generic dynamics is basically quasi-periodic. We reveal an interesting interplay between the dimension of the attractor and the architecture; more hidden units can generate higher dimensional attractors.

Testing the genericity of the results obtained for the clean system when noise is added to the output of the model is described in section 5. We found that the attractor is broadened by a factor which is inversely proportional to $\sqrt{N}$ whereas phase coherence is maintained for time length that linearly scales with the size of the network. We note that in the case of more than a single attractor, the typical escaping time scales exponentially with square root the size of the network ($\sqrt{N}$) and the amplitude of the noise. Since the relaxation time needed to converge to the vicinity of an attractor scales linearly with N (see [10,13]), this means that the system will spend most of the time near attractors, occasionally jumping to other available attractors.

We propose the following implications of these results on the prediction ability of such networks. Since neural networks have universal approximation capabilities, they should be capable of approximating the prediction function of the delay coordinate vector (which is a necessary condition for reproducing the time series). This, however, holds only for short times in general. As we have shown, long term (asymptotic) behavior of these networks is governed by the attractor dynamics described above. Therefore, the initial condition imposed on the network by a given time series is initially mapped to the vicinity of the original manifold created by the network. As the trajectory evolves according to the dynamic rules (section 2), it flows to one of the attractors of the dynamics whose characteristics were obtained analytically. In other words, any initial condition will converge to one of the possible 1D-attractors (for perceptron-SGen) or to a higher dimensional attractor (for MLN-SGen). We conclude that the asymptotic behavior for these networks is dictated by the architecture and not by the details of the weights (learning).

Interestingly, noise has a limited effect on the dynamics. This means that although one cannot predict the actual attractor to which the system will converge for a generic initial condition, the time that the system will spend in the basin of each attractor scales exponentially with $\sqrt{N}$, occasionally escaping (diffusing) to another attractor.

We note that prediction of chaotic behavior is in general short term in its essence and was not treated here. Moreover, architectures that contain internal delays (full TDNN) may exhibit more complicated behavior with longer relaxation times. Also, existence of external inputs can completely alter the results since one may initialize the inputs (or a fraction of them) to a given state at any time.

Another related issue of interest that arises in this context is the implication on computation. Several authors (e.g., [17]) have shown that recurrent neural networks (as other dynamical systems) can simulate Turing machines. Our analysis has shown that the asymptotic behavior may be predicted for a given architecture and initial condition. Namely, the set of attractors can be obtained, therefore one can predict the trajectory (up to a constant phase shift) in the stable regime and noiseless system. This behavior is, to some extent, different from dynamical systems based only on a simple functional mapping (usually polynomial), since the weight matrix of the network behaves as a filter that chooses (or amplifies) only certain modes. Note again, that even additional noise destroys phase coherence only after time that scales linearly with the network's size.

Acknowledgement

IK acknowledges partial support from the GIF.

References

[1] H.D.I. Abarbanel, R. Brown, J.J. Sidorowich and L.S. Tsimring, The analysis of observed chaotic data in physical systems, Reviews of Modern Physics 65 (1993) 1331.

[2] M. Abramowitz and C.A. Stegun (eds.), *Handbook of Mathematical Functions* (Dover, 1972).

[3] R.L. Devaney, *An Introduction to Chaotic Dynamical Systems* (Addison-Wesley, 1985).

[4] L. Ein-Dor and I. Kanter, Time series generation by multilayer networks, Physical Review E 57 (1998) 6564.

[5] E. Eisenstein, I. Kanter, D.A. Kessler and W. Kinzel, Generation and prediction of time series by a neural network, Physical Review Letters 74 (1995) 6.

[6] J.A. Hertz, A. Krogh and R.G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, 1991).

[7] I. Kanter, D.A. Kessler, A. Priel and E. Eisenstein, Analytical study of time series generation by feed-forward networks, Physical Review Letters 75 (1995) 2614.

[8] R. Metzler, W. Kinzel and I. Kanter, Interacting neural networks, Physical Review E 62 (2000) 2555.

[9] C. Moore, Generalized shifts: undecidabilty and unpredictability in dynamical systems, Nonlinearity 4 (1991) 199.

[10] A. Priel, Dynamic and static properties of neural networks with feedback, Ph.D. Thesis, Bar-Ilan University, Israel (1999). See URL: http://www.biu.ac.il/faculty/priel
[11] A. Priel and I. Kanter, Long-term properties of time series generated by a perceptron with various transfer functions, Physical Review E 59 (1999) 3368.
[12] A. Priel and I. Kanter, Robust chaos generation by a perceptron, Europhysics Letters 51 (2000) 230.
[13] A. Priel, I. Kanter and D.A. Kessler, Analytical study of the interplay between architecture and predictability, in: *Proceedings of Neural Information Processing Systems 10*, eds. M.I. Jordan, S.A. Solla and M. Kearns (MIT Press, 1998).
[14] A. Priel, I. Kanter and D.A. Kessler, Noisy time series generation by feed-forward networks, Journal of Physics A 31 (1998) 1189.
[15] T. Sauer, J.A. Yorke and M. Casdagli, Embedology, Journal of Statistical Physics 65 (1991) 579.
[16] M. Schroder and W. Kinzel, Limit cycles of a perceptron, Journal of Physics A 31 (1998) 2967.
[17] H.T. Siegelmann, B.G. Horne and C.L. Giles, Computational capabilities of recurrent NARX neural networks, IEEE Transactions on Systems, Man and Cybernetics 27 (1997) 208.
[18] H. Tong, *Non-Linear Time Series* (Oxford University Press, 1990).
[19] A. Waibel, Modular construction of time-delay neural networks for speech recognition, Neural Computation 1 (1989) 39.
[20] T.L.H. Watkin, A. Rau and M. Biehl, The statistical mechanics of learning a rule, Reviews of Modern Physics 65 (1993) 499.
[21] A.S. Weigand and N.A. Gershenfeld (eds.), *Time Series Prediction*, Santa Fe Proceedings (Addison-Wesley, 1994).