RC24112 (W0611-132) November 27, 2006
Mathematics

IBM Research Report

Learning and Representation Capabilities of Echo State Networks

Ralph Linsker, Geoffrey Grinstein
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Learning and Representation Capabilities of Echo State Networks

Ralph Linsker, Geoffrey Grinstein

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

Correspondence to: Dr. Ralph Linsker, Rm. 36-110, IBM T. J. Watson Research Center, 1101 Kitchawan Road & Route 134, P.O. Box 218, Yorktown Heights, NY 10598. Tel: 914-945-1077; Fax: 914-945-2141; Email: linsker@us.ibm.com

Learning and Representation Capabilities of Echo State Networks

Abstract

We present results on the capability of echo state networks (ESNs) to generate and learn sinusoidal oscillations, and on the sufficiency condition for ESNs to exhibit the echo state property. In particular, we show analytically and verify numerically that, provided $N \geq 2K$, the adjustable weights of a linear ESN with N reservoir nodes can be chosen so that the ESN generates a linear superposition of K sinusoidal oscillations having arbitrary prescribed periods, including periods long compared with the network dynamics, and multiple incommensurate periods. These weights can be learned using the usual MSE minimization training procedure; however, the accuracy with which they are learned is limited by numerical round-off error. When the network activities (after training) are temporarily perturbed by random noise, the periods of the individual oscillations comprising the signal are recovered correctly by the ESN, but their amplitudes and relative phases are not. We also present numerical results on the ability of nonlinear ESNs to learn individual sinusoidal oscillations and their superpositions. Finally, we demonstrate through explicit example that not all ESNs having reservoir connection matrices with spectral radii less than unity possess the echo state property. The probability of encountering such counterexamples decreases with increasing N, consistent with a conjecture by Jaeger.

Key words: Echo state network (ESN); Echo state property; Learning

1 Introduction

The "echo state network" (ESN) (Jaeger & Haas, 2004), and the closely-related "liquid state machine" (Maass, Natschläger, & Markram, 2002), are recurrent neural network architectures in which only a small fraction of the connections have modifiable weights. A general ESN comprises (a) input nodes I, "reservoir" nodes R, and output nodes O; (b) feedforward connections (specified by a weight matrix W^I) from I to R, feedback connections (matrix W^F) from O to R, and a typically sparse set of recurrent connections (matrix W) from R to R, where all of these weight matrices are randomly chosen and fixed; and (c) adjustable-weight feedforward connections W^{IO} from I to O, and W^L from R to O, whose values are to be learned by a training process. The input nodes (and their connections), and the feedback connections W^F, are optional. Each node i, at each discrete time step t, has a real-valued activity $x_i^{\{I,R,O\}}(t)$; the corresponding activity (column) vector for each set of nodes is denoted $x^{\{I,R,O\}}(t)$. All weights are real-valued. The goal of the training process is to learn adjustable weights so that a measure of the discrepancy between $x^O(t)$ and a prescribed teacher function $z(t)$ is minimized. To function as an ESN, a network must satisfy the "echo state property" (ESP, defined below).

The dynamics of ESN activity are given by (Jaeger & Haas, 2004)

$$\begin{aligned} x^O(t) &= f[W^L x^R(t) + W^{IO} x^I(t)] \quad , \\ x^R(t+1) &= f\{W x^R(t) + W^I x^I(t) + W^F [a z(t) + (1-a) x^O(t)] + n(t)\} \end{aligned} \tag{1}$$

where $a = 1$ when the teacher function provides the feedback input to the network, $a = 0$ when the network is "free-running" without teacher feedback,

and $n(t)$ denotes an optional noise term.

The invertible function f operates separately on each component of its vector argument; it is typically a nonlinear monotonic sigmoidal or "squashing" function such as tanh, but is the identity function for the special case of a linear ESN. The adjustable weights can be learned, e.g., by minimizing the "training error"

$$\text{MSE} = \langle \| W^L x^R(t) + W^{IO} x^I(t) - f^{-1}[z(t)] \|^2 \rangle \tag{2}$$

where $\| \cdot \|$ denotes the L2 norm, and $\langle \cdot \rangle$ denotes an average over the training time interval.

The training and use of an ESN comprises several stages in sequence:

(1) A startup interval of length T_{start}, starting with arbitrary $x^R(0)$, during which $a = 1$ and prescribed $\{x^I(t), z(t)\}$ are provided to the network. Since $a = 1$, $x^O(t)$ is not used, and the first of Eqs. 1 is not computed.

(2) A training interval of length T_{train}, with $a = 1$ and prescribed $\{x^I(t), z(t)\}$, during which the values of $\{x^{R,I}(t), f^{-1}[z(t)]\}$ are stored.

(3) A learning step in which the values of $W^{L,IO}$ are computed by minimizing MSE of Eq. 2. The training error is the minimum value of the MSE thus obtained.

(4) A "free-running" interval of length T_{free}, with $a = 0$ and prescribed $\{x^I(t)\}$, during which the learned $W^{L,IO}$ are used to compute outputs $\{x^O(t)\}$. The "test error" can be computed using Eq. 2.

The ESP is defined as follows, paraphrasing (Jaeger, 2002, p.29): Assume an untrained network with weights $\{W^I, W, W^F\}$ is driven by inputs x^I and

teacher signals z drawn from compact intervals. The network satisfies the ESP (with respect to these intervals) if the network state $x^R(t)$ is uniquely determined by any left-infinite sequence of input and teacher signals $\{z(-\infty), x^I(-\infty), \ldots, z(t-1), x^I(t)\}$. Informally, the ESP is satisfied if x^R at asymptotically long times is independent of the initial x^R.

Although interesting practical applications of ESNs have been described, the theoretical understanding of these networks is at an early stage. It has been suggested on empirical grounds (Jaeger, 2002, p.41) that it is "almost impossible ... to obtain ESN generators of very slow sinewaves" using standard sigmoidal networks; alternative ESN dynamics (e.g. a leaky integrator network having an adjustable time constant) have been invoked in order to generate slowly varying signals. It has also been noted that "in practice it was consistently found that when ... the spectral radius of the [reservoir] weight matrix" $\rho(W)$ (i.e., the maximum of the absolute values of the eigenvalues of W) is less than 1, the resulting network satisfies the ESP (Jaeger, 2002, p.30; Jaeger & Haas, 2004, Supporting online material, p.6). The best analytic bounds on the ESP are that (a) $\rho(W) < 1$ is a *necessary* condition (Jaeger, 2001) and (b) $\mu(W) < 1$ is a *sufficient* condition (Buehner & Young, 2006) for the ESP, where $\mu(W) \equiv \inf_D \sigma(DWD^{-1})$, $\sigma(Z)$ denotes the largest singular value of Z, D ranges over the set of diagonal matrices of the same size as W (this limitation on D arising because f is assumed to be a nonlinear "squashing" function acting on each component of its vector argument), and "inf" denotes the infimum. For specially structured W (including normal and triangular matrices), $\rho(W) = \mu(W)$, so the bound is tight. For general W, $\rho(W)$ is typically less than $\mu(W)$, so there is a gap between the necessity and sufficiency conditions. Jaeger (2002, p.30) has conjectured that, for arbitrary small δ and ϵ, there

exists a network size N such that a random $N \times N$ matrix W, scaled so that $\rho(W) = 1 - \delta$, satisfies the ESP with probability $1 - \epsilon$; this conjecture remains open.

In this paper we present analytic and numerical results concerning the above questions. We first show analytically that *linear* ESNs can generate sine waves of arbitrary prescribed period (including periods much greater than the time scale of the network dynamics, and greater than the training time interval), and can also generate superpositions of sine waves having arbitrary (including incommensurate) prescribed periods, by an appropriate choice of W^L. Once the correct W^L has been specified, the network can be perturbed by noise, and a superposition of sine waves having the same set of periods is again spontaneously generated after the noise has been turned off. The original amplitudes and phases of each sine wave component are, however, not remembered. We illustrate these results numerically for linear ESNs, and compare these behaviors with those found for nonlinear (sigmoidal) ESNs.

Finally, we explore numerically the question of whether $\rho(W) < 1$ is sufficient for W to satisfy the ESP for nonlinear ESNs (as it is for linear ESNs). We find that it is not sufficient: For random W matrices satisfying $\rho(W) = 0.99$ and $\mu(W) > \mu_0 (> 1)$, we find that a small fraction of such Ws violate the ESP. This fraction tends to decrease as the size of the reservoir is increased, consistent with Jaeger's conjecture (Jaeger, 2002).

2 Analytic results on generation of periodic states by a linear ESN

To understand how linear ESNs succeed in generating functions consisting of one or more sinusoidal oscillations of arbitrary prescribed period, we consider for simplicity a noiseless N-node ESN reservoir connected to a single output node (and having no input node), so that W^F is a (nonzero) $N \times 1$ matrix and W^L is a $1 \times N$ matrix. Equations 1 then yield $x^R(t+1) = Wx^R(t) + W^F z(t)$ for the teacher-forcing ($a = 1$) phase, and $x^R(t+1) = Mx(t)$ for the free-running ($a = 0$) phase, where $M \equiv W + W^F W^L$.

2.1 The case $N = 2$

Let us first specialize to the case $N = 2$ and a teacher signal $z(t)$ that is a single sine wave of unit amplitude: $z(t) = \sin(2\pi t/\tau)$, for arbitrary period τ. During the initial training phase, $z(t)$ induces in the two components of $x^R(t)$ a sinusoidal oscillation of period τ. Once transients have died away and this oscillation is established, the weights W^L are chosen to minimize the MSE over some number T_{train} of training time steps. Thereafter the teacher input is turned off ($a = 0$), and we want the free-running dynamics using the chosen W^L to spontaneously reproduce the desired sinusoidal oscillation in the output $x^O(t)$. This implies that $x^R(t)$ must oscillate with period τ, which in turn implies that W^L must be chosen to make the eigenvalues of the matrix M equal to $e^{\pm i\phi}$ where $\phi = 2\pi/\tau$. Solving the eigenvalue equation (which is linear in W^L) yields a solution for W^L provided the components of W and W^F satisfy

$$W_1^F W_2^F (W_{11} - W_{22}) + (W_2^F)^2 W_{12} - (W_1^F)^2 W_{21} \neq 0 \quad . \tag{3}$$

The accuracy with which the desired oscillation period can be reproduced by the ESN is limited by numerical round-off error (or other noise if present), provided that the training phase is long enough for the oscillation of the teacher signal to be accurately reflected in the behavior of the reservoir. Note also that, having learned the values of W^L necessary to generate a desired sinusoidal oscillation, a linear ESN will continue to generate a sine wave of the same period even if it is temporarily subjected to noise or other disturbance. Once the disturbance is removed, the sinusoidal oscillation will reappear, though its amplitude and phase will no longer be the same as that of the teacher signal it was designed to replicate. The neutral stability to changes in amplitude is of course a standard feature of linear systems, while the neutral stability to changes in phase is a standard feature of any system (such as the ESN with $a = 0$) having dynamics that are time-translation invariant but with an output that breaks the time-translation symmetry.

2.2 The case $N \geq 3$

We give first a general proof that a sine wave of arbitrary period, or a superposition of K sine waves of arbitrary periods τ_k, can be generated by a linear ESN that is generic (i.e., provided a singularity condition is avoided) and has at least $2K$ nodes. We then give an alternative, more intuitive explanation of this result. To generate such a superposition, we want to show that a W^L exists such that M has K conjugate pairs of eigenvalues $\exp(\pm i\phi_k)$ (denoted $\lambda_1 \cdots \lambda_{2K}$) where $\phi_k \equiv 2\pi/\tau_k$, and such that all other eigenvalues of M (denoted $\lambda_{2K+1} \cdots \lambda_N$) lie strictly within the unit circle. Choose arbitrary values for $\lambda_{2K+1} \cdots \lambda_N$ that satisfy $|\lambda_k| < 1$. By definition, each eigenvalue

satisfies $\det(P^{(k)}) = 0$ where $P^{(k)} \equiv M - \lambda_k I$. Note that each $\det(P^{(k)})$ is linear in W^L. [To see this, assume without loss of generality that $W_1^F \neq 0$, and subtract W_i^F/W_1^F times the first row of $P^{(k)}$ from the ith row (for all $i > 1$) of $P^{(k)}$. Then the resulting matrix has a term proportional to W_j^L in the jth column of row 1, and no W^L dependence in any other rows.] The system of linear equations $\det(P^{(k)}) = 0$ includes complex-conjugate pairs of equations; replace these by equations for the real and imaginary parts to obtain a real system $AW^L = B$ of N linear equations in N unknowns. Provided A is nonsingular, W^L exists. When $N = 2$, Eq. 3 is explicitly recovered.

Note that the above argument ensures that a suitable W^L exists and can be constructed in the generic (nonsingular) case whether or not the network satisfies the ESP, and that the free-running state having that W^L will generate a superposition of sines having the desired periods. However, for such a W^L to be *learned* by the process of minimizing the MSE during a training period, we require the ESP to be satisfied, so that the reservoir activity during the training period reflects the teacher signal, rather than an intrinsic activity consisting of modes whose eigenvalues have magnitude greater than one (in the absence of the ESP).

To understand the construction of W^L more intuitively, consider first the case of an N-node reservoir that we wish to use to generate a single sine wave of specified period. Assuming W satisfies the ESP, its eigenvalues must all lie within the unit circle. There exists a real matrix S such that the similarity transformation $W \to W' \equiv S^{-1}WS$ yields a real W' in block-diagonal form, with a 1×1 block corresponding to each real eigenvalue of W and a 2×2 block corresponding to each complex conjugate pair of eigenvalues. Thus the linear N-node ESN reservoir effectively is transformed into a "quasinode" network

that is a disjoint collection of 1- and 2-node subnetworks. Let $W'^F \equiv S^{-1}W^F$ be the transformed version of W^F; i.e., the feedback connection weights to the quasinodes. Let the row vector W'^L be the adjustable weights (to be determined) from the quasinodes to the output node. Set $W_i'^L = 0$ for all $i > 2$, and choose $W_1'^L$ and $W_2'^L$ (as in the $N = 2$ argument above) so that the upper 2×2 block of the matrix $M' \equiv W' + W'^F W'^L$ has eigenvalues $e^{\pm 2\pi i/\tau}$. These choices for W'^L guarantee that the first two eigenvalues of M' generate the $\sin(2\pi t/\tau)$ oscillation, while all other eigenvalues remain the same as those of W, i.e., less than unity in magnitude. Finally, defining $W^L \equiv W'^L S^{-1}$, we know by construction that the matrix M has the same eigenvalues as $M' = S^{-1}MS$.

In order to generate a superposition of sine waves having arbitrary periods (whether commensurate or not), note that applying the above method for one of the periods yields an M' that does not, in general, have the block-diagonal form of W', owing to $W_1'^L$ and $W_2'^L$ being nonzero. However, we can apply a second real similarity transformation to produce a matrix M'' that is again real, block-diagonal, and has the same eigenvalues as M'. Applying the method of the previous paragraph, with M'' now playing the role of W', and repeating this process as needed, yields the desired result.

3 Numerical results

3.1 Learning of single sine waves

We find that a sine-wave teacher function is readily learned by a *linear* ESN, even when its period τ is very long compared with the intrinsic time scale of

the dynamics, and long compared with the duration of the training interval. In an example run, an ESN having $N = 20$ reservoir nodes, 15% connectivity, and spectral radius $\rho(W) = 0.85$, learns a sine wave of $\tau = 4111$ essentially perfectly (i.e., with a test error of $< 10^{-13}$) in the absence of noise. The nodes' activities are initially random, and the run parameters (see Introduction for definition) are $T_{\text{start}} = 1000$, $T_{\text{train}} = 1000$ (i.e., only one quarter-period is used for training), and $T_{\text{free}} = 150000$. When i.i.d. noise $n(t)$ from a uniform distribution on $[-n_{\max}, n_{\max}]$ with $n_{\max} = 0.001$ is inserted for 50000 time steps during the free-running period, and free-running is then continued in the absence of noise, a sine wave having the trained period is regenerated, although the original training amplitude and phase are of course not recovered. Also, if a network is trained using a sine of period τ, and is then teacher-forced by a sine of different period (without retraining), the network's output rapidly reverts (during the free-running interval) to a signal of period τ.

If low-level noise ($n_{\max} = 10^{-10}$) is applied during the *training* interval, the learned period is approximately equal to τ, with an error whose s.d. is approximately 6%.

A *nonlinear* ESN (using $f = \tanh$) fails to learn a sine wave having long period τ, consistent with Jaeger's observations (Jaeger, 2002). Even when initial activities and teacher signals are scaled to lie within a weakly nonlinear regime [e.g., $x^R(0)$ and $z(t) \sim O(10^{-5})$ to $O(10^{-2})$], and T_{start} and T_{learn} are increased to 30000, the free-running behavior does not have the desired period, and depends upon the random initial conditions [W and $x^R(0)$]. In some cases the period of the learned oscillation is of the same order as the teacher period, and in other cases it is much shorter [e.g., of $O(10)$].

When the sine wave period is of moderate length (e.g., $\tau \sim 20$), an ESN operating in the strongly nonlinear regime [teacher signal and initial activities of $O(1)$] will learn the teacher signal accurately. If noise of $O(10^{-3})$ is added during free-running, then removed, both the amplitude and the period of the sine wave will rapidly return to their learned values. An ESN in the weakly nonlinear regime [e.g., teacher signal and initial activities of $O(0.01)$] can also accurately learn a sinusoidal teacher signal of moderate period, and will also recover its approximate learned amplitude, although more slowly than in the strongly nonlinear case. In neither case is the correct phase (i.e., that corresponding to a fully noiseless run) recovered.

3.2 Learning of sine wave superpositions

We consider the training of an ESN by a teacher signal that equals the sum of two moderate-period sines; the periods τ_1 and τ_2 may or may not be commensurate. A *linear* ESN learns to generate a superposition of both sine waves, regardless of period commensurability. Inspection of the eigenvalues of the learned matrix M (see "Analytic results" above) shows that two conjugate pairs of eigenvalues lie on the unit circle in the complex plane at angles of $\pm 2\pi/\tau_1$ and $\pm 2\pi/\tau_2$, and all other eigenvalues lie within the unit circle. When noise is applied during the free-running interval, then turned off, the two learned periods are recovered, but neither the relative amplitude nor the relative phase of the two sinusoidal components is recovered by the network. Therefore, the teacher signal pattern (i.e., the sum of the sinusoids, with the correct relative amplitudes and phases of the two components) is not recovered after a noisy interval, even if the LCM of the two periods is much shorter

than the training interval.

As K is increased (e.g., for $K = 4$ with $N = 20$), the eigenvalues of $M = W + W^F W^L$ that result from the minimum-MSE training procedure do not precisely yield $2K$ conjugate pairs lying on the unit circle and having the correct phases, owing to numerical round-off error. In this case even a small error, leading to a pair having modulus > 1, will result in unstable behavior even in the absence of intentionally added noise.

A *nonlinear* ESN often, but not always, learns the sum of two sines that have moderate (rather than long) periods that are commensurate (their LCM being small compared with T_{learn}), and recovers the correct pattern (i.e., including amplitudes and relative phases of the components) after an interval of noise during the free-running period. Whether or not learning and/or recovery occurs depends on the random choice of network. For example, we used parameters $N = 200$ with 5% connectivity, a teacher signal that is a sum of sines having periods 7 and 11, $T_{\text{start}} = T_{\text{train}} = 4000$ or 10000, $T_{\text{free}} = 30000$, and $\rho(W) = 0.85$, with $x^R(0)$ and $z(t)$ either of $O(1)$ or $O(0.05)$, and noise (during part of the free-running interval) of amplitude 0.0002 (relative to the signal). For different random networks, either (a) the pattern was learned and was recovered post-noise (except of course that the pattern was out-of-phase with the teacher signal), (b) the pattern was learned, but the post-noise output recovered neither the pattern nor the power spectrum of the teacher signal (e.g., high-frequency components were added), or (c) the pattern was not learned (i.e., even the pre-noise free-running output differed from the teacher signal).

When two sines have incommensurate periods, we find that their superposition is not learned, consistent with Jaeger's observations (Jaeger, 2002).

3.3 Exploration of the sufficiency condition for the echo state property

For a *linear* ESN in which the eigenvalues of W all lie within the unit circle (i.e., $\rho(W) < 1$), it is clear that the echo state property (ESP) is always satisfied. Given two runs, starting with different initial states $x(0)$ and $\tilde{x}(0)$, and using the same stream of teacher signals, the difference $y(t) \equiv x(t) - \tilde{x}(t)$ evolves according to $y(t) = Wy(t-1)$; i.e., $y(t) = W^t y(0)$. Thus $\lim_{t\to\infty} y(t) = 0$, satisfying the ESP.

For a *nonlinear* ESN, the situation is more complicated, and the best sufficiency condition for the ESP is that of Buehner and Young (2006): $\mu(W) < 1$. Also as noted in the Introduction, Jaeger (2002; Jaeger & Haas, 2004) had earlier noted that in all cases he had tried, an ESN having $\rho(W) < 1$ consistently satisfied the ESP, although this was purely an empirical observation.

We have explored numerically the intermediate regime $\rho(W) < 1 < \mu(W)$, within which the ESP has been empirically reported to be consistently satisfied, although no sufficiency condition for the ESP has been established. For nonlinear (tanh) ESNs, and reservoirs of various sizes N, we generated $2000 - 10000$ random reservoir matrices W (each element chosen from a uniform distribution on $[-1, 1]$, without imposing a sparseness condition), scaled each W so that $\rho(W) = 0.99$, and screened the Ws, keeping only those W for which $\mu(W) > \mu_0 = 1.2$, a value arbitrary chosen to ensure a significant separation between $\rho(W)$ and $\mu(W)$. For each of these Ws, we considered the simple case in which $x^I = z = 0$, and we set $\tilde{x}(0) = 0$ for convenience, so that $y(t) = x(t) = \tanh(Wx(t-1))$. The ESP condition (with respect to any compact interval that includes $x^I = 0$ and $z = 0$) then requires that

$\lim_{t\to\infty} y(t) = 0$. We chose a random $x(0)$ (uniform on $[-x_{\max}, x_{\max}]$, with $x_{\max} = 0.5$ unless otherwise stated) and iteratively computed $y(t)$ for enough time steps (400 to 10^6 in various runs) to determine whether (a) $y(t)$ converges to zero as required if W satisfies the ESP [we will refer to this as a contracting $y(t)$] or (b) W violates the ESP, with $y(t)$ either converging to a nonzero fixed point or limit cycle, or exhibiting apparently chaotic behavior.

For $N = 2$, approximately 30% of the rescaled random Ws satisfied the screening condition $\mu(W) > \mu_0$. Of those Ws, about 1.5% did not satisfy the ESP. An explicit example of a W that violates the ESP [in that $y(t)$ converges to a nonzero fixed point that depends on $x(0)$], is given by the 2×2 matrix $W = [3.6136, -1.9339; 4.3328, -2.0476]$ (for which $\rho(W) = 0.99$ and $\mu(W) = 5.8293$).

For $N = 4$, about 60% of rescaled random Ws satisfied $\mu(W) > \mu_0$, and about 0.5% of those Ws did not satisfy the ESP. In most of these non-ESP cases, the dynamics rapidly converged to a nonzero fixed point or to a limit cycle of period 2 or 4; a minority of cases exhibited apparently chaotic behavior.

For $N = 8$, about 90% of rescaled random Ws satisfied $\mu(W) > \mu_0$, and about 0.2% of those Ws did not satisfy the ESP.

When the maximum initial node activity $x_{\max}$ was small, so that the $x(t)$ dynamics at early t lie in the weakly nonlinear regime, a smaller fraction of W yielded noncontracting behavior than when $x_{\max} = 0.5$ as above. This is expected since, in the fully linear case, all W having $\rho(W) < 1$ show contracting behavior, and the only exceptions in the weakly nonlinear case are those in which $x(t)$ enters the strongly nonlinear regime by increasing to $O(1)$, which is less likely for smaller $x_{\max}$. For example, for $N = 2$, choosing

$x_{\max} = \{1, 0.1, 0.01\}$ led, respectively, to approximately $\{1.7\%, 0.7\%, 0.05\%\}$ of the Ws (after screening) yielding noncontracting behavior.

Our results on N-dependence provide support for the conjecture that, given a random set of Ws that satisfy $\rho(W) < 1 < \mu_0 < \mu(W)$, the fraction of those Ws that violate the ESP decreases with increasing N. This in turn is consistent with Jaeger's conjecture (see Introduction), in which, given arbitrarily small positive δ and ϵ, an N is conjectured to exist.

4 Conclusions

In this paper, we have presented analytic and numerical results for the learning of sine waves by linear and nonlinear ESNs. For the linear case, we demonstrated analytically the existence of a set of values of the adjustable weights W^L that enable an ESN with N nodes, and arbitrary reservoir connection strengths, to generate a linear superposition of K sine waves having arbitrary prescribed periods, provided $N \geq 2K$ (and except for singular cases that span a set of measure zero). The correct periods are stored in the form of complex-conjugate pairs of eigenvalues of the matrix $M = W + W^F W^L$. When W satisfies the ESP, such a W^L is learned by the usual training process, with an accuracy that depends on the training time, noise (if any), and the number (and range of the set of periods) to be learned.

Having learned the desired weights and hence the teacher signal, the linear ESN retains memory of the K sine-wave periods even after it is perturbed by noise that is then turned off. The linear ESN has no way to retain memory of the correct amplitudes or phases (even the relative phases) of the sine waves

after being perturbed, however. This neutral stability with respect to changes in amplitude and phase is of course a general feature of linear systems.

For nonlinear ESNs, we found numerically that a single sine wave is readily learned provided its period is not very long compared with the dynamical time scale. A sum of two sine waves having periods whose LCM is moderate is often (but not always) learned and then recovered (along with the correct amplitudes and relative phase) following a noisy interval when in free-running mode. Memory of the overall phase of the pattern (relative to the teacher signal) is of course destroyed by the noise perturbation, since all systems with spontaneously broken time-translation invariance are neutrally stable with respect to changes in overall phase. Analytic results on the learning of sine waves by nonlinear ESNs would have obvious value, both for advancing the general state of understanding of ESNs and as a guide for further numerical experiments and applications.

Finally, although empirical results have suggested (Jaeger, 2002; Jaeger & Haas, 2004) that $\rho(W) < 1$ might be sufficient in practice for nonlinear ESNs to satisfy the ESP, we have demonstrated the existence of nonlinear ESNs having $\rho(W) < 1$ but lacking the echo state property. The increasing scarcity of such counterexamples with increasing N is consistent with a conjecture of Jaeger.

5 References

Buehner, M. & Young, P. (2006). A tighter bound for the echo state property. *IEEE Transactions on Neural Networks, 17*, 820-26.

Jaeger, H. (2001). The "echo state" approach to analyzing and training recurrent neural networks. GMD Technical Report 148, German National Research Center for Information Technology.

Jaeger, H. (2002). Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the echo state network approach. GMD Technical Report 159, Fraunhofer Institute AIS,
http://www.faculty.iubremen.de/hjaeger/pubs/ESNTutorial.pdf

Jaeger, H. & Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science, 304*, 78-80.

Maass, W., Natschläger, T. & Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation, 14*(11), 2531-60.