Analysis and Design of Echo State Networks for Function Approximation

Mustafa C. Ozturk, Dongming Xu, and José C. Principe
Computational NeuroEngineering Laboratory
Department of Electrical and Computer Engineering,
University of Florida, Gainesville, FL 32611
{can, dmxu, principe}@cnel.ufl.edu

Abstract

The present design of echo state network (ESN) parameters relies on the selection of the maximum eigenvalue of the linearized system around zero. However, this has been found far from optimal for function approximation. This letter presents a function approximation perspective to better understand the operation of ESNs and proposes an information-theoretic measure, the average entropy of echo states, to assess the “richness” of the ESN dynamics. Furthermore, it provides a new interpretation of the ESN dynamics rooted in system theory as a combination of linearized systems where their poles move according to the input signal dynamics. With this interpretation, we will be able to a priori design ESNs with uniform pole distributions covering the frequency spectrum optimally. With adaptive read-outs, the designed ESN can be used as a general infrastructure to represent information in time.

1Correspondence address: Mustafa C. Ozturk, Computational NeuroEngineering Laboratory, Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, Tel: (352)392-2682, Fax: (352)392-0044
1 Introduction

Dynamic computational models require the ability to store and access the time history of their inputs and/or outputs. The most common neural architecture to achieve this goal is the time delay neural network that couples delay lines with a nonlinear static architecture. The conventional delay line utilizes ideal delay operators, but delay lines with local first order recursive filters have been proposed by Werbos (Werbos, 1992), from which the most studied example is the gamma model (de Vries, 1991; Principe et al., 1993). They are interesting because they effectively decrease the number of delays necessary to create time embeddings, while they can still be studied in the same general framework of Laurent series expansions of system transfer functions (Principe, 2001). Dynamic computational models require the ability to store and access the time history of their inputs and or outputs. The most common neural architecture to achieve this goal is the time delay neural network that couples delay lines with a nonlinear static architecture. From the point of view of embeddings, recurrent neural networks implement a different type of embedding yet largely unexplored. Theoretical studies of recurrent neural networks are rare (Siegelmann and Sontag, 1991; Siegelmann, 1993; Kremer, 1995), but they have been widely used in many applications such as system identification and control of dynamical systems (Kechriotis et al., 1994; Puskorius and Feldkamp, 1994; Delgado et al., 1995). The main problem with the RNNs is the difficulty to adapt the system weights. Various algorithms, such as back propagation through time (BPTT) (Werbos, 1990) and real time recurrent learning (RTRL) (Williams and Zipser, 1989), have been proposed to train recurrent systems. However, these algorithms suffer from a variety of problems: computational complexity resulting in slow training, complex performances surfaces, the possibility of instability, and the decay of gradients through the topology and time (Haykin, 1998). The problem of
decaying gradients has been addressed with special PEs (Hochreiter and Schmidhuber, 1997). Alternative, second order training methods based on extended Kalman filtering (Singhal and Wu, 1989; Puskorius and Feldkamp, 1994; Feldkamp et al., 1998) and the multi-streaming training approach (Feldkamp et al., 1998) provide more reliable performance and have enabled practical applications.

Recently, a new recurrent network topology has been proposed by Jaeger under the name of echo state networks (ESN) (Jaeger, 2001a,b; Jaeger and Hass, 2004). ESNs possess a highly interconnected and recurrent topology of nonlinear PEs which constitutes a “reservoir of rich dynamics” (Jaeger, 2001a) and contain information about the history of input or/and output patterns when properly dimensioned (the echo state). The outputs of these internal PEs are fed to a memoryless readout network that reads the reservoir and produces the network output. The interesting property of ESN is that only the memoryless readout (generally linear) is trained, whereas the recurrent topology has fixed connection weights. This reduces the complexity of RNN training to simple linear regression while preserving the power of recurrent topology.

Let us look at the architecture and recursive update equation of a typical ESN more closely. Consider the recurrent discrete-time neural network given in Figure 1 with $M$ input units, $N$ PEs and $L$ output units. The value of the input unit at time $n$ is $u(n) = [u_1(n), u_2(n), \ldots, u_M(n)]^T$, of internal units are $x(n) = [x_1(n), x_2(n), \ldots, x_N(n)]^T$, and of output units $y(n) = [y_1(n), y_2(n), \ldots, y_L(n)]^T$. The connection weights are given in a $N \times M$ weight matrix $W^{in} = (w^{in \ i}_i)$ for connections between the input and the internal states, in an $N \times N$ matrix $W = (w_{ij})$ for connections between the internal PEs, in an $L \times M$ matrix $W^{inout} = (w^{inout \ i}_i)$ for connections from input units to the output units, in an $L \times N$ matrix $W^{out} = (w^{out \ i}_i)$ for connections from internal units to the output units, in an $L \times L$ matrix $W^{outout} = (w^{outout \ i}_i)$ for connections between the output units, and in an $N \times L$ matrix $W^{back} = (w^{back \ i}_i)$ for the connections that project back from the
output to the internal units (Jaeger, 2001a). The activation of the internal PEs is updated according to

\[ x(n+1) = f(W^{in}u(n+1) + Wx(n) + W^{back}y(n)), \]  

(1.1)

where \( f = (f_1, f_2, \ldots, f_N) \) are the internal unit’s output functions. The output is computed according to

\[ y(n+1) = f^{out}(W^{out}x(n+1) + W^{inout}u(n+1) + W^{outout}y(n)), \]  

(1.2)

where \( f^{out} = (f^{out}_1, f^{out}_2, \ldots, f^{out}_L) \) are the output unit’s output functions (Jaeger, 2001a,b).

The echo state condition is defined in terms of the spectral radius of the reservoir’s weight matrix (\( \|W\| < 1 \)), which relates to the condition under which the recurrent network states are strongly coupled with the input history. Although this condition is very useful in terms of defining the region of parameters resulting in echo states, it does not specify a sufficiently accurate design principle to construct ESNs for function approximation.

This letter first proposes a signal processing interpretation of basis functions and projections to describe and understand the ESN architecture. According to this interpretation, the reservoir states function as a set of basis (representation space) constructed dynamically by the input, while the readout simply projects in this space the best representation of the desired response. An information theoretic quantity, average state entropy (ASE) is further proposed to assess the richness of ESN dynamics. Entropy measures the amount of information contained in a given random variable (Shannon, 1948). Here, the random variable is the state amplitude from which the instantaneous entropy for the (vector) state is estimated whereas the ASE is the state entropy averaged over time. The average state entropy estimates the volume of the state manifold.

Thirdly, the ESN dynamics are interpreted as a combination of time varying linear systems obtained from the linearization of the system in a small local neighborhood of
the current state on the nonlinear function. With this interpretation, we propose to design ESNs with uniform pole distributions covering the frequency spectrum optimally. We will demonstrate that ESNs with uniform pole distribution generate echo states with highest ASE since the system includes different time constants corresponding to different pole locations. This interpretation also shines light on the power of nonlinear PEs since the poles of the linearized ESN move in the frequency domain in response to the input signal amplitude.

2 A signal processing interpretation of Echo State Networks

ESNs are not very similar to the RMLP architecture proposed in (Puskorius and Feldkamp, 1996) and also used by (Sanchez, 2004) in brain machine interfaces. The critical difference is the difference in the dimensionality of the hidden recurrent PE layer and the adaptation of the recurrent weights. In a sense ESNs are trading the adaptive connections in the RMLP hidden layer by a brute force approach of creating diversified dynamics in the hidden layer. For this reason we believe that the ideas of function approximation in terms of basis and projections, so useful in adaptive signal processing (Principe, 2001), should be utilized to understand this architecture. A similar perspective for information processing in biological networks has been investigated by Sejnowski (Pouget and Sejnowski, 1997). They explored the possibility that the response of single neurons in parietal cortex serve as a basis function for the transformations from the sensory input to the motor responses. They proposed that “the role of spatial representations is to code the sensory inputs and posture signals in a format that simplifies subsequent computation, particularly in the generation of motor commands” (Pouget and Sejnowski, 1997). In the case of ESN, the spatial transformations are done by the reservoir which simplifies the computation for the simple instantaneous read-out.
Let $f(u)$ be a real-valued function of a real-valued vector $\mathbf{u} = [u_1, u_2, \ldots, u_d]^T$. In function approximation, the goal is to estimate the behavior of $f(u)$ as a combination of simpler functions $\varphi_i$, called the basis functions, such that its approximant $\hat{f}(u)$, given by

$$\hat{f}(u) = \sum_{i=1}^N w_i \varphi_i,$$

is obtained, where $N$ is the number of basis. Here, the $w_i$'s are the projections of $f(u)$ onto each basis function. One of the central questions in function approximation is how to choose the set of basis functions to approximate a given desired signal. In signal processing, the choice normally goes for complete set of orthogonal basis, independent of the input such as complex exponentials, etc. In neural computing (which includes adaptive signal processing) the basic idea is to use the input signal to derive the set of basis. For instance, in the FIR filter, the bases are the delayed versions of the input (a delay embedding). In neural networks, consider a single hidden layer multilayer perceptron (MLP) with $N$ PEs and a linear output. Its hidden layer outputs $\varphi_i$ can be considered a set of basis functions dependent upon the input

$$\hat{\varphi}_i(u) = g(\sum_j a_{ij} u_j),$$

where $a_{ij}$'s are the input layer weights, and $g$ is the PE nonlinearity. The approximation produced by the MLP is then

$$\hat{f}(u) = \sum_{i=1}^N w_i \varphi_i(u),$$

(2.1)

where $w_i$'s are the weights of the output layer. Assuming a linear read-out network without output feedback connections ($\mathbf{W}_\text{outout} = 0$ and $\mathbf{W}_\text{back} = 0$), equation 1.2 has the same form of equation 2.1, where the $\varphi_i$'s and $w_i$'s are replaced by the echo states and the readout weights, respectively.

The output of the MLP is a linear combination of its internal representations, but we still need to enforce linear independence among the $\varphi_i$'s to achieve a basis. The effect
of nonlinear MLP transformation on the input signal has been investigated by many authors. Oh has shown that correlation among the weighted sums decrease after they pass through the sigmoid nonlinear function which can be approximated by piecewise linear functions (Oh and Lee, 1994). Ito has shown the linear independence of plane waves for commonly used activation functions (Ito, 1996). The following theorem, proved in (Shah and Poon, 1999), provides a nice theoretical framework for function approximation with MLPs and also ESNs.

**Theorem** (Shah and Poon, 1999): For every nonzero vector, \( u \in \mathbb{R}^T \) of distinct elements, there exists ordered pairs \((a_i, b_i), a_i \in \mathbb{R}, b_i \in \mathbb{R}, i = 1, \ldots, T \) such that \( \{\tanh(a_1 u + b_1 1), \ldots, \tanh(a_T u + b_T 1)\} \) forms a complete set of basis for \( \mathbb{R}^T \). Here \( 1 \) is the \( N \)-dimensional vector of ones representing the hidden unit biases.

Now, let us consider the ESN topology as an architecture for approximation of time functionals (Sandberg and Xu, 1997). The MLP is a special case of the ESN where the internal connection weight matrix is chosen to be the \( 0 \) matrix. Hence, according to the above theorem, there exist weight values for the ESN such that the echo states form a set of basis functions for the \( T \) dimensional vector space where \( u \in \mathbb{R}^T \). In fact, the recurrent weights in the ESN make this choice even easier since they are able to create phase shifts among the states, which is impossible with the MLP (a static network).

For parameter sets that provide independent states, the span of an \( N \) state ESN still depends both on the selection of the architecture and the value of its parameters. Parameter selection is presently done with the help of the echo state condition defined in terms of the spectral radius of the reservoir weight matrix (\( \|W\| < 1 \)). In fact, the echo state guarantees stability of the linearized recurrent network. The span of the space created by the ESN states is however not controlled, i.e. there are many possible weight matrices that obey the echo state property but provide a large diversity of spans.

A simple experiment was designed to demonstrate that this selection of the fixed
echo state parameters is not the most suitable for function approximation. Consider an echo state network of 100-units where the input signal is $\sin(2\pi n/T)$. Mimicking (Jaeger, 2001a), the goal is to let the ESN generate the seventh power of the input signal, $\sin^7(2\pi n/T)$, where $T$ is chosen to be $10\pi$. Different realizations of a randomly connected 100 unit ESN were constructed where the weight values are set to 0.4, -0.4 and 0 with probabilities of 0.025, 0.025 and 0.95. In other words, we obtain different random $W$ matrices with the given constraint. Moreover, $W$ is normalized for each realization such that a spectral radius of 0.88 is obtained (Jaeger, 2001a). $W_{\text{inout}}$ and $W_{\text{outout}}$ are both zero, and the input weight values are set to +1 or -1 with equal probabilities. For each realization of the $W$ matrix with the given specification, the input was applied for 300 time steps and the echo states calculated using equation 1.1. The next step is to construct the desired signal from the echo states using a memoryless linear readout network. One method to determine the optimal output weight matrix, $W_{\text{out}}$ in the mean square error (MSE) sense, is to use the Wiener solution given by (Haykin, 2001).

$$W_{\text{out}} = E[xx^T]^{-1}E[xd] = (\frac{1}{N} \sum_n x(n)x(n)^T)^{-1}(\frac{1}{N} \sum_n x(n)d(n)) \quad (2.2)$$

Here, $E[.]$ denotes the expected value operator and $d$ denotes the desired signal which is $\sin^7(2\pi n/T)$. Figure 2 depicts the MSE values for 50 different realizations of an ESN with the given distribution of weights and spectral radius. As observed from the figure, even though each ESN has the same distribution of weights and spectral radius, the MSE values obtained vary greatly among different realizations. The minimum MSE value obtained among the 50 realizations is $5.9 \times 10^{-9}$ whereas the maximum MSE is $8.9 \times 10^{-5}$. This experiment demonstrates that a design strategy that is based solely on the spectral radius is not sufficient to specify the system architecture for function approximation.
3 Average State Entropy as a Measure of the Richness of ESN Reservoir

The concept of “rich dynamical reservoir” (Jaeger, 2001a) has not been quantified with a well-defined metric. Here, we propose to use the instantaneous state entropy to quantify the distribution of instantaneous amplitudes across the ESN reservoir states. The average state entropy (ASE), defined as the state entropy averaged over time, will be the parameter used to quantify the diversity in the dynamical reservoir of the ESN. Entropy is appropriate for this quantification because if the echo state’s amplitudes are concentrated on only a few values across the ESN state dynamic range, the ability to construct the desired response is limited and performance will suffer. On the other hand, if the ESN states provide a diversity of amplitudes, then it is much easier to achieve the desired mapping. Hence, the entropy of the amplitude states seems a good measure to quantify “richness” of the dynamics for function approximation with instantaneous mappers. Moreover, from nonlinear dynamics, there is evidence that entropy is an appropriate measure of the volume of the signal manifold spanned by the reservoir states.

Renyi’s quadratic entropy is employed here because of the existence of an efficient nonparametric estimator that avoids explicit pdf estimation (Principe et al., 2000). Renyi’s entropy with parameter $\alpha$ for a random variable $Y$ with a pdf $f_Y(y)$ is given by (Renyi, 1970)

$$H_\alpha(y) = \frac{1}{1-\alpha} \log E[f^{\alpha-1}(y)].$$

The Renyi’s quadratic entropy is obtained for $\alpha = 2$ (for $\alpha = 1$ Shannon’s entropy is obtained). Given $N$ samples $\{y_1, y_2, \ldots, y_N\}$ drawn from the unknown pdf to be
estimated, Parzen windowing approximates the underlying pdf by

\[ f(y) = \frac{1}{N} \sum_{i=1}^{N} K_\sigma(y - y_i), \]

where \( K \) is the kernel function with the kernel size \( \sigma \). Then the Renyi’s quadratic entropy can be estimated by (Principe et al., 2001)

\[ H_2(y) = -\log \left[ \frac{1}{N^2} \sum_j \left( \sum_i K_\sigma(y_j - y_i) \right) \right]. \tag{3.1} \]

The instantaneous state entropy is estimated using equation 3.1 where the samples are the entries of the state vector \( \mathbf{x}(n) = [x_1(n), x_2(n), \ldots, x_N(n)]^T \), of an ESN with \( N \) internal PEs. Results will be shown with a Gaussian kernel with a variable kernel size chosen to be 0.3 of the standard deviation of the current state. As a demonstration of the state entropy computation and interpretation, let us consider the same 100 unit ESN that we used in the previous section built with three different spectral radii 0.2, 0.5, 0.8 by multiplying the internal connection matrix, \( \mathbf{W} \), with suitable constants. The input to the ESN is the signal, \( \sin(2\pi n/T) \), where \( T \) is 10. Figure 3A depicts the echo states over 200 time ticks. The instantaneous state entropy is also calculated at each time step using equation 3.1 and plotted in Figure 3B. First, note that the instantaneous state entropy changes over time with the distribution of the echo states as we could expect. Since entropy is not scale invariant, when the input signal takes values around zero, all echo states are also close to zero resulting in a sharp decrease in the state entropy. Second, as the spectral radius increases, the diversity in the echo states increases. This is an important observation that will be very useful when we are designing ESNs for function approximation in the next sections. In practice, to quantify the overall representation ability over time, we will use ASE, which has values of -0.735, -0.007 and 0.335, for the spectral radii of 0.2, 0.5 and 0.8 respectively. Although we have presented experiments including only sinusoidal signals, similar results are obtained for other inputs as long as the input dynamic range is properly selected.
Now let us address the question of the usefulness of ASE to predict better MSE in approximation for the same spectral radius. Figure 4 depicts the MSE and the corresponding ASE values for 50 different realizations of 30-unit ESNs with the spectral radius of 0.9. The figure shows that the high values of ASE correlates negatively with low MSE values in function approximation (the correlation coefficient between the ASE and MSE is -0.53). However, there are cases where a low value of MSE is not correlated with a high value of ASE (simulation 5). Likewise, a high MSE is normally related to a low value of ASE (except simulation 28). As a conclusion, ASE is a finer descriptor of the MSE in function approximation, but it may not be the only variable controlling the approximation ability of the ESN.

4 Interpreting ESN Dynamics as a Combination of Linear Systems

We know that the dynamics of a nonlinear system can be approximated by that of a linear system in a small local neighborhood of a given equilibrium point (Kuznetsov et al., 1998). Here, we will perform the analysis with hyperbolic tangent PEs and approximate the ESN dynamics by the dynamics of the linearized system in the local neighborhood of the current value of the system state. Hence, when the system operating point varies over time, the linear system approximating the ESN dynamics will change. We are particularly interested in the movement of the poles of the linearized ESN. Consider the update equation for the ESN without output feedback given by

\[ x(n + 1) = f(W^{in}u(n + 1) + Wx(n)). \]
Linearizing the system around the current state, \( \mathbf{x}(n) \), one obtains the Jacobian matrix, \( \mathbf{J} \), defined by
\[
\mathbf{J} = \begin{bmatrix}
\dot{f}(x_1)w_{11} & \dot{f}(x_1)w_{12} & \cdots & \dot{f}(x_1)w_{1N} \\
\dot{f}(x_2)w_{21} & \dot{f}(x_2)w_{22} & \cdots & \dot{f}(x_2)w_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
\dot{f}(x_N)w_{N1} & \dot{f}(x_N)w_{N2} & \cdots & \dot{f}(x_N)w_{NN}
\end{bmatrix} = \begin{bmatrix}
\dot{f}(x_1) & 0 & \cdots & 0 \\
0 & \dot{f}(x_2) & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \dot{f}(x_N)
\end{bmatrix} \cdot \mathbf{W}.
\]

In equation 4.1, \( x_i \) denotes the \( i^{th} \) entry of the current state vector and \( w_{ij} \) denotes the \((i, j)^{th}\) entry of \( \mathbf{W} \). The poles of the linearized system are given by the eigenvalues of the Jacobian matrix \( \mathbf{J} \). When the amplitude of each PE input changes, the local slope changes, and so the poles of the linearized system will be time varying, although no parameter is changing in the ESN. In order to visualize the movement of the poles, consider an ESN with 100 states. The entries of the internal weight matrix are chosen to be 0, 0.4 and -0.4 with probabilities 0.9, 0.05 and 0.05. The weights are scaled such that a spectral radius of 0.95 is obtained for the linearized system at the origin. The entries of the input weight matrix, \( \mathbf{W}^{in} \) are chosen to be 1 or -1 with equal probability. A sinusoidal signal with the period of 100 is fed to the system and the echo states are computed according to equation 1.1. Then, the Jacobian matrix and the eigenvalues are calculated using equation 4.1. Figure 5 shows the pole tracks of the linearized ESN when the input goes through a cycle. A single ESN with fixed parameters implements a combination of many linear systems with varying pole locations, hence many different time constants that modulates the richness of the reservoir of dynamics as a function of input amplitude. Higher amplitude portions of the signal tend to saturate the nonlinear function and cause the poles to shrink toward the origin of the z-plane (decreases the spectral radius), which results in a system with large stability margin. When the input is close to zero, the poles of the linearized ESN are close to the maximal spectral radius chosen, decreasing the stability margin. An ESN with more states results in a detailed
coverage of the z plane dynamics (ε-coverage), which illustrates the power of nonlinear
systems, when compared to their linear counterpart.

5 Designing Echo State Networks

Our aim is to design ESN computational reservoirs for a variety of tasks that will lead
to small approximation errors, even when we do not have any prior knowledge of the
input and output signals. Our proposal is to design the ESN such that the linearized
system around zero state has uniform pole distribution inside the unit circle of the z-
plane. With this design scenario, the system dynamics will include uniform coverage
of time constants arising from the uniform distribution of the poles. This principle was
chosen by analogy to the identification of linear systems using Kautz filters (Kautz,
1954) which shows that the best approximation of a given transfer function by a linear
system with finite order is achieved when poles are placed in the neighborhood of the
spectral resonances.

In order to obtain the pole locations that result in a uniform coverage of the unit
circle, we will use again a maximum entropy principle to distribute a number of points
in a specified area uniformly. The constraints of a circle for discrete linear systems and
complex conjugate locations are easy to include as boundary conditions for the pole
distribution (Thogula, 2003). The poles are first initialized at random locations, the
information potential is calculated (Erdogmus and Principe, 2002) and they are moved
such that the entropy of the new distribution is increased over iterations. In linear
system theory, the poles that are close to the unit circle correspond to many sharp band
pass filters specialized in different frequency regions whereas the inner poles realize
filters of larger frequency support. Moreover, different orientations (angles) of the poles
creates filters of different center frequency. There are some minor deviations from
uniformity due to the details in the entropy estimation (Thogula, 2003), but this method is efficient to find a uniform coverage of the unit circle with an arbitrary number of poles.

Now, our problem is to construct an internal weight matrix from the pole locations. This is equivalent to the problem of designing $W$ when its eigenvalues are known. In principle we would like to create a sparse matrix $W$, so we started with the sparsest matrix (with an inverse) which is the direct canonical structure given by (Kailath, 1980)

$$
W = \begin{bmatrix}
-a_1 & -a_2 & \cdots & -a_{N-1} & -a_N \\
1 & 0 & \cdots & 0 & 0 \\
0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 1 & 0
\end{bmatrix}
$$

which has the characteristic polynomial

$$
c(s) = \det(sI - W) = s^N + a_1s^{N-1} + a_2s^{N-2} + \cdots + a_N = (s - p_1)(s - p_2)\ldots(s - p_N),
$$

where $p_i$’s are the eigenvalues of the matrix $W$ and $a_i$ are the coefficients of the characteristic polynomial of $W$. Here, we know the pole locations of the linear system obtained from the linearization of the ESN, so using equation 5.2, we can obtain the characteristic polynomial for $W$ and construct a $W$ matrix in the canonical form using equation 5.1. We will call the ESN constructed based on the uniform pole principle as ASE-ESN. All other possible solutions with the same eigenvalues can be obtained by $T^{-1}WT$, where $T$ is any nonsingular matrix. However, performances of systems for different $T$ matrices are yet to be analyzed.

To corroborate our hypothesis, we would like to show that the linearized ESN designed with the recurrent weight matrix having the eigenvalues uniformly distributed inside the unit circle creates the highest ASE for a given spectral radius compared to any other ESNs with random internal connection weight matrices. We will consider a
30 state ESN, and use our procedure to create the uniformly distributed linearized ASE-ESN matrix for different spectral radius between [0.1, 0.95]. Similarly, we constructed ASE-ESNs with sparse random $W$ matrices with different sparseness constraints and the same spectral radii as above. This corresponds to a weight distribution having the values 0, $c$ and $-c$ with probabilities $p_1$, $(1 - p_1)/2$ and $(1 - p_1)/2$, where $p_1$ defines the sparseness of $W$ and $c$ is a constant that takes a specific value depending on the spectral radius. Then, for different $W^{in}$ matrices, we run the ASE-ESNs with the sinusoidal input given in section 3 and calculated ASE. These values are obtained as an average over 1000 realizations of the ESNs for each spectral radius. Figure 6 compares the ASE values obtained for ASE-ESN having $W$ with uniform eigenvalue distribution and the ESN with random $W$ matrix. As observed from the figure, the ASE-ESN with uniform pole distribution generates higher ASE on the average for all spectral radii compared to ESNs with random sparse connections.

6 Experiments

In the previous section, we proposed a design scheme for the internal connection weight matrix, $W$, of ESNs and showed that ASE-ESNs designed with this method generates highest ASE on the average. Function approximation with basis generated by the input is a problem that requires information from both the input and desired responses, i.e. it is an optimal projection in the joint input output space. Therefore, we are aware that a condition on the input space alone, and indirectly quantifying the richness of the bases, as ASE, is ill prepared to become optimal for all problems. However, we want to show that the ASE-ESN with uniform pole distribution is a reasonable choice, better than just constraining the spectral radius as initially proposed by Jaeger. Therefore, we will use a family of functions to evaluate the performance of our ASE-ESN.
The first set of simulations constructs a family of powers of a signal input to the ESN. In particular, the input to the ESN is chosen to be $\sin(2\pi n/T)$ and the desired output is $\sin^\alpha(2\pi n/T)$, where $\alpha$ is varied over the values 1, 3, 5, 7, 9, and 11. Here, $T$ is chosen to be 10. For each task, we compared the performance of ASE-ESNs with 30 echo states formed using our method with the randomly created ones having different sparseness constraints. ESNs used in this experiment do not have output feedback connections. Three different sets of random connection weights are generated with the values 0, $c$ and $-c$ with probabilities $p_1$, $(1-p_1)/2$ and $(1-p_1)/2$, where $p_1$ is one of 0.8, 0.9 and 0.94 and $c$ is a constant that takes a suitable value corresponding to a spectral radius of 0.95. The ESN internal connection weight matrix generated using our method also has the spectral radius of 0.95. For each set with different weight distribution, 1000 different realizations of ESN are created with different $W$ and $W^{\text{in}}$ matrices. For a particular value of $\alpha$, after generating $W$ and $W^{\text{in}}$ of an ESN, we run the input signal to calculate the echo states using equation 1.1 and eliminated the first 100 steps corresponding to the initial transient. Then, the optimal weights of the linear readout are calculated using equation 2.2 and the corresponding MSE value is calculated. Figure 7 depicts the MSE values averaged over 1000 different realizations of the ESN as a function of $\alpha$ for each set of weight distributions. We see that the ASE-ESNs having uniform pole distribution generate the lowest MSE for all powers compared to randomly connected ESNs. In the second set of simulations, we tackle the problem of identification of 2 different systems.

**System 1** is $y(n+1) = 0.3y(n) + 0.6y(n-1) + f[u(n)]$, where $f(u) = 0.6\sin(\pi u) + 0.3\sin(3\pi u) + 0.1\sin(5\pi u)$. The input signal is chosen to be $\sin(2\pi n/25)$.

**System 2** is $y(n+1) = \frac{y(n)y(n-1)[y(n)+2.5]}{1+y(n-2)^2+y(n-1)^2}$, with the same input as above.

For each task, we compared the performance of 30 echo state ESNs formed using our method labeled as ASE-ESN, with the randomly created ones having different
sparseness constraints labeled as ESN2, ESN3, and ESN4. The internal connection matrix of ASE-ESN is constructed according to uniform pole distribution principle and a spectral radius of 0.95. ESN2, ESN3, ESN4 are formed by choosing the W matrix (entries can only be one of three distinct values) randomly with the same spectral radius of 0.95 and the sparseness of 0.93, 0.8, 0.6, respectively. For each ESN \( i (i=2,3,4) \) in each of the system identification tasks, 1000 different realizations of ESN are created with different realizations of W and \( W^{in} \) matrices. The entries of the input weight matrix, \( W^{in} \) are chosen to be 1 or -1 with equal probability. For each problem, after generating W and \( W^{in} \) of an ESN, we run the input signal to calculate the echo states using equation 1.1 and eliminated the samples corresponding to the initial transient. For identification of system1 and system2, the input to the ESN is the same as the input to the models. Then, the optimal weights of the linear read-out are calculated using equation 2.2 and the corresponding MSE value is calculated. Table 1 shows the mean and variance of the MSE values obtained in addition to the variance of the MSE over 1000 realizations.

The same ASE-ESNs constructed are used for a well-known benchmark task of predicting the chaotic Mackey-Glass (MG) time series. M-G time series are generated by a nonlinear differential equation and are well-known for the evaluation of prediction methods. Here, we generated MG data of 1500 samples for a time constant of 30 and a sampling period of 6. The first 1000 values are employed for training the ASE-ESN read-out and the remaining 500 values for the testing set. Here, we report the MSE values obtained for the testing set in Table 1.

As observed from the table, the MSE values obtained for the ASE/ESN gives lower average MSE values. In all cases, the variance of the MSE values are lower for ASE/ESN since W matrix is fixed and only \( W^{in} \) is random. This shows that the selection of W matrix using the uniform pole distribution principle is a reasonable choice for the selection
of ESN that can be used as a general infrastructure for function approximation.

7 Conclusions

The great appeal of echo state networks (ESNs) is their ability to construct arbitrary mappings of signals with rich and time varying temporal structures without requiring adaptation of the free parameters of the recurrent layer. The echo state condition allows the recurrent connections to be fixed and only a linear output layer needs to be trained. However, the literature did not elucidate on how to properly choose the recurrent parameters in system identification applications. In this letter, in order to evaluate the potential of any specific ESN, we propose to interpret echo states as a set of functional bases formed by fixed nonlinear combinations of the input. The linear readout at the output stage simply computes the projection of desired output space onto this representation space. We further introduce an information-theoretic criterion, ASE, for better understanding and evaluating the capability of a given echo state network to construct such a representation layer. The entropy of the distribution of the echo states at each time step is calculated and averaged over time to obtain the ASE measure. Simulations show that higher ASE with respect to the recurrent connections effectively brings much richer reservoir of dynamics among the echo states as measured by the ASE, as well as smaller MSE values in function approximation. However, ASE is a surrogate for MSE and it does not guarantee a smaller MSE in all conditions. Further research is therefore required to investigate under which conditions smaller MSE can be expected.

The results support the view that function approximation performance improves as the spectral radius of the recurrent weight matrix approaches one. It is interesting to relate this observation with the computational properties found in dynamical systems “at the edge of chaos” (Packard, 1988; Langton, 1990; Bertschinger and Natschlger,
2004). Here however, we provide a system theoretic view and explain the behavior with the diversity of dynamics achieved with linearizations that have poles close to the unit circle, exactly at the "edge of chaos".

In this letter, we propose an effective interpretation of ESN dynamics as being produced by the linearized system at the current nonlinear PEs operating point. In response to the input signal amplitude, the fixed-coefficient ASE-ESN can generate instantaneously a family of representations that may cover the full frequency spectrum. ESN design can then be determined by optimizing the location of poles of the linearized system. Experiments have indeed shown that ASE-ESNs with uniformly distributed poles tend to have the best performance in terms of MSE among all the ESNs with the same spectral radius.

There are many interesting issues to be researched in this exciting new area. Besides an evaluation tool, ASE may also be utilized to train the ESN’s representation layer in an unsupervised fashion. In fact we can easily adapt with the SIG (stochastic information gradient) described in (Erdogmus et al., 2003) the recurrent layer weights to maximize output entropy. This may fine tune continuously in an unsupervised manner the parameters of the ESN among different inputs. Effectively the reservoir of recurrent PEs can be thought as a new form of a time to space mapping. Unlike the delay line that forms an embedding (Takens, 1981) this mapping may have the advantage of filtering noise and produces representations with better SNRs to the peaks of the input, which is very appealing for signal processing and seems to be used in biology. However, further theoretical work is necessary in order to understand the embedding capabilities of ESNs more comprehensively. One of the disadvantages of the correlated basis is in the design of the readout. Gradient based algorithms will be very slow to converge (due to the large eigenvalue spread of modes) and even if recursive methods are used their stability may be compromised by the condition number of the matrix. However our
recent results incorporating a L1 norm penalty in the LMS (Rao et al., 2005) shows
great promise of solving this problem.

Finally we would like to briefly comment on the implications of these models to
neurobiology and computational neuroscience. The work by Sejnowski (Pouget and
Sejnowski, 1997) has shown that the available physiological data is consistent with
the hypothesis that the response of a single neuron in parietal cortex serves as a basis
function that is generated by the sensory input in a nonlinear fashion. In other words,
the neurons transform the sensory input into a format (representation space) such that
the subsequent computation is simplified. Then, whenever a motor command (output
of the biological system) needs to be generated, this simple computation to read out the
neuronal activity is done. There is an intriguing similarity between the interpretation
of the neuronal activity by Pouget and Sejnowski and our interpretation of echo states
in ESN. We believe that similar ideas can be adopted for liquid state machines (LSMs),
since LSMs are the counterparts of ESNs in the microscopic (neuronal) level.

Acknowledgements

This work was partially supported by ONR N00014-1-1-0405.
References


Table Captions

Table 1 The comparison of the MSE values of 4 ESNs formed with different methods evaluated for 2 different system identification and M-G time series prediction tasks. W matrix for ASE-ESN is formed based on distributing the poles uniformly inside the unit circle. The spectral radius is set to 0.95. ESN 2, ESN 3, ESN 4 are formed by choosing the W matrix randomly with the same spectral radius of 0.95 and the sparseness of 0.93, 0.8, 0.6, respectively. All experiments are run for 1000 trials with different realization of W^{in} matrix. The values of W^{in} are chosen randomly from +1 or -1 with equal probability. In all experiments, the MSE values obtained with our method result in lower mean, smaller variance and minimum MSE values.
Figure Captions

Figure 1  Diagram of an echo state network (ESN). ESN is composed of two parts: a fixed-weight (W) recurrent network and a linear readout. The recurrent network is a reservoir of highly interconnected dynamical components, states of which are called echo states. The memoryless linear readout is trained to produce the output.

Figure 2  Performances of ESNs for different realizations of W with the same weight distribution. The weight values are set to 0.4, -0.4 and 0 with probabilities of 0.025, 0.025 and 0.95. All realizations have the same spectral radius of 0.88. In the 50 realizations, MSEs vary from $5.9 \times 10^{-9}$ to $8.9 \times 10^{-5}$. Results show that spectral radius of W is not the unique parameter that determines the performance of an ESN.

Figure 3  Examples of echo states and instantaneous state entropy. (A) Outputs of echo states (100 PEs) produced by systems with spectral radius of 0.2, 0.5 and 0.8, from up to bottom, respectively. The diversity of echo states increases when spectral radius increases. Within the dynamic range of the output states, systems with smaller spectral radius can only generate uneven representations. While $\|W\| = 0.8$, outputs of echo states almost uniformly distributes within their dynamic range. (B) Instantaneous state entropy is calculated using equation 3.1. Information contained in the output states is changing over time according to the input amplitude. Therefore, richness of representation is controlled by the input amplitude.

Figure 4  MSE and the corresponding ASE values for 50 different realizations of 30-unit ESNs with the spectral radius of 0.9. The figures show that the high values of
ASE correlate with low MSE values in function approximation (the correlation coefficient between the ASE and MSE is -0.53).

**Figure 5** The pole tracks of the linearized ESN with 100 PEs when the input goes through a cycle. A single ESN with fixed parameters implements a combination of many linear systems with varying pole locations. (A) One cycle of sinusoidal signal with a period of 100. (B)-(E) show the positions of poles of the linearized systems when the input values are at A, B, C, and D in Figure 5(A). (F) Trajectory of pole locations shows the movement of the poles as the input changes. Due to the varying pole locations, different time constants modulates the richness of the reservoir of dynamics as a function of input amplitude. Higher amplitude signal tends to saturate the nonlinear function and causes the poles to shrink toward the origin of the z-plane (decreases the spectral radius), which results in a system with large stability margin. When the input is close to zero, the poles of the linearized ESN are close to the maximal spectral radius chosen, decreasing the stability margin. An ESN with more states results in a detailed coverage of the z plane dynamics (ε-coverage), which illustrates the power of nonlinear systems, when compared to their linear counterpart.

**Figure 6** Comparison of ASE values obtained for ASE-ESN having W with uniform eigenvalue distribution and the ESN with random W matrix. Randomly generated weights have sparseness of 0.07, 0.1 and 0.2. ASE values are calculated from the networks with spectral radius from 0.1 to 0.95. As observed from the figure, the ASE-ESN with uniform pole distribution generates higher ASE on the average for all spectral radii compared to ESNs with random sparse connections.

**Figure 7** We compare performances of ESNs with different structures in terms of MSE values. Three networks are generated with randomly connected weights.
Sparseness of connections are 0.06, 0.1 and 0.2, respectively. The 4\textsuperscript{th} network is constructed with uniformly distributed poles. All of them have a spectral radius of 0.95. Different set of desired signals are generated by varying the $\alpha$ in $\sin^\alpha(2\pi n T)$. ESN with uniformly distributed poles outperforms the others in all cases.
<table>
<thead>
<tr>
<th></th>
<th>ESN Type</th>
<th>Average MSE</th>
<th>Variance of MSE</th>
<th>Minimum MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System 1</strong></td>
<td>ASE/ESN</td>
<td>1.83x10^{-6}</td>
<td>9.66x10^{-11}</td>
<td>1.70x10^{-11}</td>
</tr>
<tr>
<td></td>
<td>ESN 2</td>
<td>4.93x10^{-5}</td>
<td>4.86x10^{-9}</td>
<td>2.57x10^{-9}</td>
</tr>
<tr>
<td></td>
<td>ESN 3</td>
<td>1.23x10^{-5}</td>
<td>4.75x10^{-10}</td>
<td>1.02x10^{-9}</td>
</tr>
<tr>
<td></td>
<td>ESN 4</td>
<td>1.09x10^{-5}</td>
<td>4.51x10^{-10}</td>
<td>1.02x10^{-9}</td>
</tr>
<tr>
<td><strong>System 2</strong></td>
<td>ASE/ESN</td>
<td>2.59x10^{-3}</td>
<td>2.97x10^{-5}</td>
<td>3.68x10^{-12}</td>
</tr>
<tr>
<td></td>
<td>ESN 2</td>
<td>3.03x10^{-2}</td>
<td>2.52x10^{-4}</td>
<td>7.14x10^{-7}</td>
</tr>
<tr>
<td></td>
<td>ESN 3</td>
<td>2.36x10^{-2}</td>
<td>2.18x10^{-4}</td>
<td>6.37x10^{-7}</td>
</tr>
<tr>
<td></td>
<td>ESN 4</td>
<td>2.46x10^{-2}</td>
<td>2.23x10^{-4}</td>
<td>1.16x10^{-7}</td>
</tr>
<tr>
<td><strong>Mackey Glass</strong></td>
<td>ASE/ESN</td>
<td>2.22x10^{-3}</td>
<td>3.47x10^{-7}</td>
<td>1.02x10^{-3}</td>
</tr>
<tr>
<td></td>
<td>ESN 2</td>
<td>3.24x10^{-3}</td>
<td>5.22x10^{-7}</td>
<td>1.66x10^{-3}</td>
</tr>
<tr>
<td></td>
<td>ESN 3</td>
<td>2.66x10^{-3}</td>
<td>3.98x10^{-7}</td>
<td>1.29x10^{-3}</td>
</tr>
<tr>
<td></td>
<td>ESN 4</td>
<td>2.66x10^{-3}</td>
<td>3.50x10^{-7}</td>
<td>1.54x10^{-3}</td>
</tr>
</tbody>
</table>
Figure 1:
Different Realizations

Figure 2:
Figure 3:

(A) Echo States

(B) State entropy

- SR = 0.2
- SR = 0.5
- SR = 0.8
Figure 4:
Figure 5:
Comparison of ASE values for ESNs formed with different methods

- ASE–ESN
- sparseness = 0.07
- sparseness = 0.1
- sparseness = 0.2

Figure 6:
Figure 7: MSE values as a function of alpha for ESNs formed with different methods.