Published in final edited form as:
IEEE Trans Neural Netw Learn Syst. 2012 January ; 23(1): 175–182. doi:10.1109/TNNLS.2011.2178562.

Nonlinear System Modeling with Random Matrices: Echo State Networks Revisited

Bai Zhang,
Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203 USA

David J. Miller, and
Department of Electrical Engineering, Pennsylvania State University, University Park, PA 16802 USA

Yue Wang
Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203 USA

Bai Zhang: baizhang@vt.edu; David J. Miller: djmiller@engr.psu.edu; Yue Wang: yuewang@vt.edu

Abstract

Echo state networks (ESNs) are a novel form of recurrent neural networks (RNNs) that provide an efficient and powerful computational model approximating nonlinear dynamical systems. A unique feature of an ESN is that a large number of neurons (the “reservoir”) are used, whose synaptic connections are generated randomly, with only the connections from the reservoir to the output modified by learning. Why a large randomly generated fixed RNN gives such excellent performance in approximating nonlinear systems is still not well understood. In this brief, we apply random matrix theory to examine the properties of random reservoirs in ESNs under different topologies (sparse or fully connected) and connection weights (Bernoulli or Gaussian). We quantify the asymptotic gap between the scaling factor bounds for the necessary and sufficient conditions previously proposed for the echo state property. We then show that the state transition mapping is contractive with high probability when only the necessary condition is satisfied, which corroborates and thus analytically explains the observation that in practice one obtains echo states when the spectral radius of the reservoir weight matrix is smaller than 1.

Index Terms

Circular law; concentration of measure; echo state networks; echo state property; random matrix theory; recurrent neural networks

I. Introduction

Recurrent neural networks (RNNs) are widely used to model nonlinear dynamical systems. Recently, a new framework for RNNs, namely echo state networks (ESNs), was proposed by Jaeger *et al.* [1], [2]. ESNs (and closely related liquid state machines, independently proposed by Maass *et al.* [3]) share some features characteristic of models for learning in biological brains and they exhibit superior performance when used as “black-box” time-

series models. In an ESN, neurons in a fixed (non-trainable) recurrent layer, known as “the reservoir,” are driven by the input signals, and the trainable output neurons combine the output of the excited reservoir state to generate task-specific temporal patterns. This new RNN paradigm is also known as “reservoir computing.”

ESNs have drawn great interest from the research community and have been successfully applied to various tasks, e.g., chaotic time-series prediction [4], communication channel equalization [1], dynamical pattern recognition [5], [6], and gene regulatory network modeling [7]. Various ESN schemes have been explored, including a small-world recurrent neural system with scale-free distribution [8], decoupled ESNs with lateral inhibition [9], ESNs with uniformly distributed poles and adaptive bias [10], augmented complex ESNs [11], and echo state Gaussian process [12]. Rodan and Piňo investigated the minimal complexity of reservoir construction required to achieve good representation power for ESNs, and proposed three simple deterministically constructed reservoir topologies [13]. Lukoševičius and Jaeger presented a comprehensive review on the theoretical results and applications of ESNs in [14].

The salient difference from traditional RNNs [15], [16] is that an ESN employs a large number of *randomly* connected neurons (usually on the order of 50 to 1000), namely the “reservoir,” i.e., unlike traditional RNNs, the connection weights between neurons in the recurrent (reservoir) layer do not require any supervised training—only connection weights to output neurons are optimized. Thus, training is greatly simplified compared to traditional RNNs and well-known RNN training problems of slow convergence, even lack of convergence, and local minima are avoided. In fact, if the ESN employs a linear activation function in the output layer, ESN training reduces to a simple linear regression problem.

The working principle of an ESN derives from an important algebraic property of the reservoir, namely the *echo state property* (ESP). A recurrent reservoir driven by an external input signal has the ESP if the reservoir states are systematic variations of the input driving signal. Essentially, satisfying the ESP means that the effect of both previous states and previous inputs on a future state will gradually vanish (i.e., neither persist nor become amplified) as time passes [2]. If the ESP holds, the reservoir network state will asymptotically (in time) depend only on the input history and the nonlinear system will be well-approximated through a linear combination of the reservoir’s “echo state” signals. Metaphorically, under the ESP, the reservoir state signal can be thought of as an “echo” of the input history.

Jaeger presented both a necessary condition (under the assumption that the input space includes the zero sequence) and a sufficient condition for the ESP [2]. Buehner and Young proposed a less restrictive sufficient condition based on minimizing the matrix operator **D**-norm over the set of diagonal matrices [17]. However, these papers did not consider the unique characteristic of the reservoir, i.e., that it is *randomly* generated. Here, by exploiting this fact and applying results from random matrix theory, we will show that the sufficient conditions in [2] and [17] are rather conservative.

The topology of the reservoir in ESNs has been of great research interest, with the classical form a randomly generated and sparsely connected network [1], [2]. Several attempts have been made to search for a better topology—the small-world, scale-free, and biologically inspired reservoir topologies. However, quoting [14], "none of the investigated network topologies was able to perform significantly better than simple random networks, both in terms of eigenvalue spread as well as testing error." Also, we again note [13] which, while not giving designs that outperform simple random reservoirs, sought the minimum complexity topology needed to achieve good modeling power.

The novel contributions of this brief are threefold. First, motivated by the above quotation, we analytically examine the essential characteristics of random reservoirs. We apply recent results from random matrix theory to demonstrate the asymptotic distributions of eigenvalues and singular values of reservoir weight matrices. We then show that randomly generated reservoirs, either sparsely or fully connected, either with Bernoulli or Gaussian connection weights (or, in fact, with weights distributed according to other density families), are all expected to behave similarly. These results thus explain the above-quoted observation from [14]. Second, we quantify the gap between the scaling factor bounds used to define the ESP necessary and sufficient conditions proposed in previous works. We show that, asymptotic in the size of the reservoir, this gap becomes quite large, with the necessary condition bound twice as large as the sufficient condition bound. Finally, we show that, when the spectral radius of the reservoir weight matrix is smaller than 1 (the *necessary* condition for the ESP when the input space contains the zero sequence), the state transition mapping is in fact contractive with high probability, given a sufficiently large reservoir. This result corroborates the observation in [2] that the necessary condition for the ESP is often good enough in practice, such that violations of the ESP are not practically observed. This result, together with the factor of two asymptotic gap between the scaling factor bounds, indicates the conservativeness of the sufficient conditions from [2] and [17]. The practical implication of these results is that standard ESN design approaches, based on use of the sufficient conditions, are suboptimal—use of a conservative scaling factor compromises the amount of memory in the RNN, and thus the ability to accurately model a given target dynamical system.

The remainder of this brief is organized as follows. In Section II, we revisit the ESN model, random reservoirs, and the ESP. This is followed by detailed discussion in Section III on relevant results from random matrix theory, the properties of random reservoirs, and the gap between the sufficient and necessary conditions previously proposed for the ESP. In Section IV, we prove that the necessary condition for the ESP ensures the state transition mapping is contractive with high probability. We briefly conclude our work in Section V.

II. ESN Formulation

A. Basic ESN Formulation

A typical ESN is shown in Fig. 1. It can be represented by state update and output equations. While enhanced representation power for an RNN may be achieved by the use of output feedback, this can also introduce instability problems [14], [18]. To avoid these issues and also to simplify the mathematical analysis, we will focus in this brief on ESNs without

output feedback, as also adopted by others [2], [17]. Thus, the activation of internal units is updated according to

$$\mathbf{x}(n+1)=f(\mathbf{W}\mathbf{x}(n)+\mathbf{W}_{\text{in}}\mathbf{u}(n+1)) \quad (1)$$

where $\mathbf{x}$ is a $N \times 1$ vector of the reservoir state, $\mathbf{W}$ is a $N \times N$ reservoir weight matrix, $\mathbf{W}_{\text{in}}$ is an $N \times N_{\text{in}}$ input weight matrix, $\mathbf{u}$ is a $N_{\text{in}} \times 1$ vector of system inputs, $\mathbf{y}$ is a $N_{\text{out}} \times 1$ vector of system outputs, and f is the neuron activation function (usually a tanh sigmoid function), applied component-wise.

For notational convenience, we denote the state transition equation by

$$\begin{aligned}\mathbf{x}(n+1)&=T(\mathbf{x}(n),\mathbf{u}(n+1))\\&=f(\mathbf{W}\mathbf{x}(n)+\mathbf{W}_{\text{in}}\mathbf{u}(n+1))\end{aligned} \quad (2)$$

and the output equation by

$$\mathbf{y}(n)=g\left(\mathbf{W}_{\text{out}}\begin{bmatrix}\mathbf{x}(n)\\\mathbf{u}(n)\end{bmatrix}\right) \quad (3)$$

where $\mathbf{W}_{\text{out}}$ is the $N_{\text{out}} \times (N + N_{\text{in}})$ output weight matrix, and g is usually a tanh sigmoid or an identity function, applied component-wise.

B. Random Reservoirs in ESNs

A salient feature that distinguishes ESNs from conventional RNNs is the use of large fixed random reservoirs. The classical ESN reservoir topology is a randomly generated and sparsely connected network [1]. It was thought that "this condition lets the reservoir decompose into many loosely coupled subsystems, establishing a richly structured reservoir of excitable dynamics" [1]. Nevertheless, this is not generally true and it has in fact been reported that fully connected reservoirs work just as well as sparsely connected ones [18]. Such observation leads to inquiry of the essential characteristics of random reservoirs and their role in approximating nonlinear dynamical systems.

The types of random reservoirs are characterized by the structure of the reservoir weight matrix. Assume the matrix $\mathbf{W} = \alpha\mathbf{W}_N$, where α is a properly chosen global scaling factor (whose utility will be discussed later), and where the elements of the matrix $\mathbf{W}_N$ are random variables that are independent and identically distributed (i.i.d.). Here we consider the following three types of reservoir weight matrices.

Sparse random reservoir—This is the most common type of random reservoir in ESNs [1], [2]. The random variable w (which characterizes each element of $\mathbf{W}_N$) follows the modified Bernoulli probability mass function (PMF)

$$\begin{cases} \Pr(w=0)=1-c \\ \Pr(w=\pm 1)=\frac{c}{2} \end{cases} \quad (4)$$

where Pr(·) denotes probability of an event and $c \in [1/3, 1)$ is "the connectivity" of the reservoir. Note that if $\mathbf{W}_N[i, j] = 0$, there is no connection from reservoir neuron i to reservoir neuron j. Thus, using the modified Bernoulli PMF leads to a realization of $\mathbf{W}$ that is sparsely connected.

Fully connected Gaussian random reservoir—*w* follows a standard normal distribution

$$w \sim N(0, 1). \quad (5)$$

Fully connected Bernoulli random reservoir—*w* follows the Bernoulli distribution

$$\Pr(w=\pm 1)=\frac{1}{2}. \quad (6)$$

These three types of reservoir weight matrices exhibit different network topologies, i.e., either sparsely connected or fully connected neurons in the reservoir, and different types of weights, i.e., either continuous-valued or discrete-valued. All three types have been used as random reservoirs in ESNs and have been successfully applied.

C. Definition of ESP

In order to work properly, an ESN should possess the ESP, as defined in [2].

Definition 1 (Jaeger [2])—Assume standard compactness conditions, i.e., inputs drawn from a compact input space U and network states restricted to a compact set A. Assume that the network has no output feedback connections. Then, the network has echo states if the network state $\mathbf{x}(n)$ is uniquely determined by any left-infinite input sequence $\bar{\mathbf{u}}^{-\infty}$. More precisely, this means that for every input sequence, …, $\mathbf{u}(n-1)$, $\mathbf{u}(n) \in U^{-\mathbb{N}}$, for all state sequence pairs …, $\mathbf{x}(n-1)$, $\mathbf{x}(n) \in A^{-\mathbb{N}}$ and …, $\mathbf{x}'(n-1)$, $\mathbf{x}'(n) \in A^{-\mathbb{N}}$, where $\mathbf{x}(k) = T(\mathbf{x}(k-1), \mathbf{u}(k))$, $\mathbf{x}'(k) = T(\mathbf{x}'(k-1), \mathbf{u}(k))$, and $\mathbb{N}$ is the set of natural numbers, it holds that $\mathbf{x}(n) = \mathbf{x}'(n)$.

The definition of the ESP implies that similar echo state sequences must represent similar input histories. In [2], Jaeger also provided several equivalent characterizations of echo states, e.g., the properties of being state contracting, state forgetting, and input forgetting. However, the ESP definition is hard to check in practice. A known *sufficient* algebraic condition for the ESP is that the largest singular value of $\mathbf{W}$ (defined as the square root of the largest eigenvalue of $\mathbf{W}\mathbf{W}^T$) is smaller than 1. On the other hand, the ESP is violated (for input space containing the zero sequence) when the spectral radius of $\mathbf{W}$ (defined as its largest magnitude eigenvalue) is greater than 1. Therefore, the spectral radius of $\mathbf{W}$

restricted to being less than or equal to 1 serves as a *necessary* condition for the ESP. The following theorem formally presents these two conditions for the network to possess the ESP.

Theorem 1 (Jaeger [2])—Assume a sigmoid network, i.e., with $f = \tanh$, applied component-wise: 1) let the weight matrix $\mathbf{W}$ satisfy $\sigma_{\max} < 1$, where $\sigma_{\max}$ is its largest singular value. Then $d(T(\mathbf{x}, \mathbf{u}), T(\mathbf{x}', \mathbf{u})) < d(\mathbf{x}, \mathbf{x}')$ for all inputs $\mathbf{u} \in U$, for all states $\mathbf{x}, \mathbf{x}' \in [-1, 1]^N$, where $d(\cdot, \cdot)$ is any distance metric. This implies the ESP holds, and 2) let the weight matrix have spectral radius $|\lambda_{\max}| > 1$, where $\lambda_{\max}$ is the eigenvalue of $\mathbf{W}$ with the largest absolute value. Then the network has an asymptotically unstable null state. This implies that it does not satisfy the ESP for input space U containing $\mathbf{0}$ and admissible state space $A = [-1, 1]^N$.

As suggested in [2], a convenient strategy to obtain ESNs is to start with some weight matrix $\mathbf{W}_N$ and then select a global scaling factor α to suitably define $\mathbf{W} = \alpha\mathbf{W}_N$. Let $\sigma_{\max}(\mathbf{W}_N)$ and $|\lambda_{\max}(\mathbf{W}_N)|$ denote the largest singular value and the spectral radius of $\mathbf{W}_N$, respectively. Then, according to [2], for the ESP to hold, the sufficient condition is $\alpha < \sigma_{\max}^{-1}(\mathbf{W}_N)$ and the necessary condition is $\alpha < |\lambda_{\max}^{-1}(\mathbf{W}_N)|$.

Furthermore, although the existence of the ESP for $\alpha \in [\sigma_{\max}^{-1}(\mathbf{W}_N), |\lambda_{\max}^{-1}(\mathbf{W}_N)|]$ has not been theoretically proved, it has been observed, albeit without analytical justification, that "one obtains echo states even when α is only marginally smaller than $|\lambda_{\max}^{-1}(\mathbf{W}_N)|$" and "the sufficient condition is very restrictive" [2].

Buehner and Young proposed a tighter sufficient condition for the ESP. The main idea is to minimize the matrix operator $\mathbf{D}$-norm over the set of diagonal matrices [17]. The $\mathbf{D}$-norm of a vector $\mathbf{x} \in \mathbb{R}^N$ is defined to be $\|\mathbf{x}\|_{\mathbf{D}} = \|\mathbf{D}\mathbf{x}\|$, where $\mathbf{D} \in \mathbb{R}^{N \times N}$ is nonsingular. Then, the matrix operator $\mathbf{D}$-norm (the induced $\mathbf{D}$-norm) of a matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ is given by

$$\|\mathbf{W}\|_{\mathbf{D}} = \sigma_{\max}\left(\mathbf{D}\mathbf{W}\mathbf{D}^{-1}\right).$$

However, because the matrix $\mathbf{D}$ does not have full structure (and in fact was restricted to being diagonal), the sufficient condition derived in [17] is still in general conservative. Pertinent to the sequel, we observe that the derivations of the existing results on the sufficient condition [2], [17] have not taken into account the primary unique characteristic of an ESN, i.e., that the reservoir matrix is a *random* matrix.

III. Random Matrix Theory and Random Reservoirs

In this section, we first introduce some recent results in random matrix theory, and then apply them in developing some relevant properties of random reservoirs in ESNs.

A. Empirical Spectral Distribution (ESD) of Random Matrices

Let

$$\mu_{\mathbf{W}_N}(s,t) := \frac{1}{N}|\{i | 1 \le i \le N, \mathrm{Re}(\lambda_i) \le s, \mathrm{Im}(\lambda_i) \le t\}| \quad (7)$$

be the ESD of $\mathbf{W}_N$'s eigenvalues $\lambda_i \in \mathbb{C}$, $i = 1, \ldots, N$, where $|\cdot|$ denotes the cardinality of the set and Re(·) and Im(·) are the real and imaginary parts of the complex number, respectively. A well-known conjecture is *the circular law of random matrices*, which states that asymptotically, as N gets large, the eigenvalues of a properly normalized random matrix $\mathbf{W}_N$ are uniformly distributed on the unit disk in the complex plane. After many pioneering efforts in proving the circular law for various scenarios, including sparse random matrices [19]–[23], it was proved in full generality, in both weak and strong forms, quite recently [24].

Theorem 2 (Circular Law [24])—Let $\mathbf{W}_N$ be the $N \times N$ random matrix whose entries are i.i.d. complex random variables with mean 0 and variance 1. Define $\mathbf{W} = (1/\sqrt{N})\mathbf{W}_N$. Then the ESD of $\mathbf{W}$ converges (in both the strong and weak senses) to the uniform distribution on the unit disk, as $N \to \infty$.

Corollary 1—The ESDs of reservoir weight matrices $\mathbf{W}$ as defined in (4) with the scaling factor $\alpha = (1/\sqrt{cN})$, (5) with the scaling factor $\alpha = (1/\sqrt{N})$, and (6) with the scaling factor $\alpha = (1/\sqrt{N})$ all have the same limit distribution and, more specifically, converge (in both the strong and weak senses) to the uniform distribution on the unit disk.

The circular law implies that, when N is sufficiently large (as is typical for ESNs), the eigenvalues of $\mathbf{W}$ spread out evenly over the unit disk in the complex plane, independent of the specific distribution of w, as illustrated in Fig. 2. It is also important to note that, for the circular law to hold for sparse matrices, the connectivity c of the sparse matrix must satisfy the inequality $c > N^{-1+\varepsilon_1}$, where $\varepsilon_1 > 0$ is a small positive constant, because otherwise, with non-negligible probability, the sparse matrix would lose its rank-efficiency as N gets large ([23], Th. 1.3).

B. Singular Values of Random Matrices

Similarly, let $\sigma_1, \sigma_2, \ldots, \sigma_N$ be the singular values of $\mathbf{W}$. The empirical distribution of the squares of the singular values of $\mathbf{W}$ is defined by

$$\nu_{\mathbf{W}}(t) := \frac{1}{N}\left|\left\{i | 1 \le i \le N, \sigma_i^2 \le t\right\}\right|. \quad (8)$$

It has been shown that $\nu_{\mathbf{W}}$ is governed by the Marchenko–Pastur law [25]–[27].

Theorem 3 (Marchenko–Pastur Law)—Let $\mathbf{W}_N$ be the $N \times N$ random matrix whose entries are i.i.d. complex random variables with mean 0 and variance 1. Define $\mathbf{W} = (1/\sqrt{N})\mathbf{W}_N$. Then the empirical distribution of the squares of the singular values of $\mathbf{W}$, $\nu_{\mathbf{W}}(t)$, converges (both in the sense of probability and in the almost sure sense) to $(1/2\pi)\int_0^{\min(t,4)} \sqrt{((4/x)-1)}dx$, as $N \to +\infty$.

Remark—Supported by rigorous mathematical proofs, the circular and Marchenko–Pastur laws reveal an important fundamental property of random matrices, i.e., that both the eigenvalues and the singular values of random reservoir weight matrices have unique limit distributions, independent of the distribution and connectivity of w, as $N \to \infty$.

C. Gap Between the Sufficient and Necessary Conditions in [2]

As discussed in [2] and as aforementioned in Section II-C, the global rescaling factor a must be properly chosen to ensure the ESP for $\mathbf{W} = a\mathbf{W}_N$. Specifically, when $a<|\lambda_{\max}^{-1}(\mathbf{W}_N)|$, the system is stable, which serves as the necessary condition (assuming the input space contains the zero sequence), when $a<\sigma_{\max}^{-1}(\mathbf{W}_N)$, the ESP is guaranteed, i.e., this serves as the sufficient condition. However, the sufficient condition $a<\sigma_{\max}^{-1}(\mathbf{W}_N)$ is considered conservative, with the practical implication being that the associated ESN design will be suboptimal, with the amount of memory in the dynamical system compromised (the smaller a, the shorter the system memory). In fact, it has been observed that one obtains echo states even when a is only marginally smaller than $|\lambda_{\max}^{-1}(\mathbf{W}_N)|$ [2].

The discrepancy between the theoretical sufficient condition for the ESP and the empirical observation that the necessary condition often works well in practice raises a natural question: how big is the gap between $\sigma_{\max}^{-1}(\mathbf{W}_N)$ and $|\lambda_{\max}^{-1}(\mathbf{W}_N)|$]? Let the ratio $r = (\sigma_{\max}(\mathbf{W}_N)/|\lambda_{\max}(\mathbf{W}_N)|)$ quantify the gap between the sufficient and necessary condition bounds. It turns out that this gap is quite large: the asymptotic value of r is 2 as $N \to \infty$.

Before we give the proof of this result, we first introduce two theorems from the random matrix theory literature.

Theorem 4 (Bai [28])—Let $\{w_{ij} : i = 1, 2, \ldots, N, j = 1, 2, \ldots, N\}$ be i.i.d. random variables, and $\mathbf{W}_N$ be the $N \times N$ matrix $(w_{ij})_{N\times N}$, $i, j = 1, 2, \ldots, N$. Suppose: 1) $E[w_{11}] = 0$; 2) $E[w_{11}^2]=\sigma^2$; and 3) $E[|w_{11}|^4] < \infty$. Then $\limsup_{N\to\infty}\max_{1\le i\le N}|\lambda_i(\mathbf{W}_N/\sqrt{N})| \le \sigma$ a.s. where $\lambda_i(\mathbf{W}_N/\sqrt{N})$, $i = 1, 2, \ldots, N$, are eigenvalues of $\mathbf{W}_N/\sqrt{N}$.

Theorem 5 (Yin [29])—Let $\{w_{ij} : i = 1, 2, \ldots, N, j = 1, 2, \ldots, N\}$ be i.i.d. random variables, and $\mathbf{W}_N$ be the $N \times N$ matrix $(w_{ij})_{N\times N}$, $i, j = 1, 2, \ldots, N$. Suppose: 1) $E[w_{11}] = 0$; 2) $E[w_{11}^2]=\sigma^2$; and 3) $E[|w_{11}|^4] < \infty$. Let $\overline{\sigma}_N^2$ be the largest singular value of $\mathbf{W}_N/\sqrt{N}$. Then $\lim_{N\to\infty}\overline{\sigma}_N^2 = 4\sigma^2$ a.s.

Theorem 6 (Gap Between the Sufficient and Necessary Conditions)—If the random reservoir weight matrix is generated according to (4), (5), or (6), then $r \xrightarrow{a.s.} 2$, as $N \to +\infty$.

Proof: First, it is straightforward to verify that the three distributions specified by (4)–(6) all have zero mean and finite fourth-moment, and their variances are c, 1, and 1, respectively.

We consider random reservoir weight matrices generated according to (5) and (6). From Theorem 4, we have

$$\left|\lambda_{\max}\left(\frac{1}{\sqrt{N}}\mathbf{W}_N\right)\right| \leq 1, \quad \text{almost surely, as } N \to +\infty. \quad (9)$$

Then, combining (9) with the conclusion of the circular law, we have

$$\left|\lambda_{\max}\left(\frac{1}{\sqrt{N}}\mathbf{W}_N\right)\right| \xrightarrow{a.s.} 1, \quad \text{as } N \to +\infty. \quad (10)$$

Next, from Theorem 5, we have

$$\sigma_{\max}\left(\frac{1}{\sqrt{N}}\mathbf{W}_N\right) \xrightarrow{a.s.} 2, \quad \text{as } N \to +\infty. \quad (11)$$

Therefore, we have

$$r = \frac{\sigma_{\max}(\mathbf{W}_N)}{|\lambda_{\max}(\mathbf{W}_N)|} = \frac{\sigma_{\max}(\frac{1}{\sqrt{N}}\mathbf{W}_N)}{|\lambda_{\max}(\frac{1}{\sqrt{N}}\mathbf{W}_N)|} \xrightarrow{a.s.} 2, \quad \text{as } N \to +\infty. \quad (12)$$

For the case of random reservoir weight matrices generated according to (4), if we replace $1/\sqrt{N}$ by $1/\sqrt{cN}$ in the above equations, it is straightforward to show the same conclusion stated in (12).

Fig. 3 illustrates the asymptotic trend of $\sigma_{\max}(\mathbf{W})$, $\lambda_{\max}(\mathbf{W})$, and $\|\mathbf{W}\|_{\mathbf{D}}$ for Gaussian, Bernoulli, and sparse reservoir weight matrices as N increases. Each point in Fig. 3 is the average of 20 independent simulations, and $\|\mathbf{W}\|_{\mathbf{D}}$ is calculated using MATLAB μ-analysis Toolbox as suggested in [17]. First, we can see in Fig. 3 that, when N is large, Gaussian, Bernoulli, and sparse reservoirs all have similar respective values for $\sigma_{\max}(\mathbf{W})$, $\lambda_{\max}(\mathbf{W})$, and $\|\mathbf{W}\|_{\mathbf{D}}$. Second, as N increases, $\sigma_{\max}(\mathbf{W})$ tends to 2, and $\lambda_{\max}(\mathbf{W})$ tends to 1. Thus, consistent with Theorem 4, the bound for the necessary condition is about twice the bound for the sufficient condition for an ESN to possess the ESP as N gets large. Also, although we do not have theoretical results suggesting this, we observe in Fig. 3 that $\lambda_{\max}(\mathbf{W})$ is approaching its asymptote from above, while $\sigma_{\max}(\mathbf{W})$ approaches its asymptote from below. That is, the gap, and thus the level of conservativeness (and the associated degree of potential suboptimality in using the sufficient condition in ESN design, relative to a design based on the necessary condition), is empirically observed to increase with N. Third, for the sufficient bound proposed in [17], $\|\mathbf{W}\|_{\mathbf{D}}$ is indeed tighter than $\sigma_{\max}(\mathbf{W})$ when N is small, for example, for $N = 20$, but $\|\mathbf{W}\|_{\mathbf{D}}$ approaches very close to $\sigma_{\max}(\mathbf{W})$ as N gets large. Thus, empirically from Fig. 3, there appears to be little to gain in using the sufficient condition from [17], rather than the sufficient condition from [2], as N gets large.

IV. Why the Necessary Condition for Echo States is Often "Sufficient in Practice"

To establish the sufficient condition for the ESP, Jaeger [2] and Buehner and Young [17] showed that, with **W** scaled to have its largest singular value less than 1, the distance between two states $\mathbf{x}(n)$ and $\mathbf{x}\tilde{(n)}$ shrinks at every time step, i.e., $d(T(\mathbf{x}(n), \mathbf{u}(n+1)), T(\mathbf{x}\tilde{(n)}, \mathbf{u}(n+1))) < d(\mathbf{x}(n), \mathbf{x}\tilde{(n)})$, regardless of the input. This Lipschitz condition results in echo states.

In this section, alternatively, we will show that, asymptotically, as the size of the reservoir grows, for a much *less* conservative scaling of **W** that is *essentially* equivalent to scaling **W** just enough so that the necessary condition for the ESP is satisfied, the state transition mapping $T(\cdot, \cdot)$ is contractive with high probability, regardless of the input. In essence, we will thus show that the necessary condition is "sufficient in practice." In order to make our mathematical analysis tractable and, thus, to establish our results, we consider a slightly unorthodox (albeit a still reasonable) procedure for scaling of the matrix **W**. Normally, and as considered in [2], one first randomly generates the matrix $\mathbf{W}_N$ and then sets $\mathbf{W} = \alpha\mathbf{W}_N$, where α is specifically chosen to satisfy an ESP condition—choosing $\alpha \leq |\lambda_{\max}^{-1}(\mathbf{W}_N)|$ meets the necessary condition, while setting $\alpha \leq \sigma_{\max}^{-1}(\mathbf{W}_N)$ ensures sufficiency. While choosing α in this way strictly ensures one (or both) of these ESP conditions, it also makes α a function of the random matrix, i.e., α is itself a random variable, with, moreover, a distribution that is dependent on *N*. Choosing α in this way will complicate our analysis. Alternatively, from (10), we know that, if we choose $\mathbf{W} = (\rho/\sqrt{N})\mathbf{W}_N$, the spectral radius of **W** converges to ρ as $N \to \infty$. That is, picking a *constant* scaling factor $\rho < 1$, *independent* of both the dimension *N* and the particular realization of the random matrix $\mathbf{W}_N/\sqrt{N}$, satisfies the necessary condition for the ESP almost surely as *N* gets large. From this standpoint, choosing **W** in this "unconventional" way—one that is more amenable to analysis—is reasonable. More significantly, in the following, we will show that, by choosing **W** in this unconventional way, the state transition mapping $T(\cdot, \cdot)$ is contractive with high probability, regardless of the input. More specifically, for $\mathbf{x}(n), \mathbf{x}\tilde{(n)} \in [-1, 1]^N$ and a random reservoir weight matrix $\mathbf{W} = (\rho/\sqrt{N})\mathbf{W}_N$, $\rho < 1$, the inequality $d(T(\mathbf{x}(n), \mathbf{u}(n+1)), T(\mathbf{x}\tilde{(n)}, \mathbf{u}(n+1))) < d(\mathbf{x}(n), \mathbf{x}\tilde{(n)})$ holds with probability $1 - O(e^{-C_\rho N})$, where the constant C_ρ depends on ρ. In this sense, we show that asymptotically, for large *N*, the necessary condition is "*sufficient in practice*." Finally, although our theoretical results will assume an unconventional procedure for scaling **W**, we will subsequently demonstrate at least *empirically* that "sufficiency of the necessary condition in practice" also applies if one uses the more standard procedure for scaling **W**.

A key ingredient for establishing our results is the *concentration of measure phenomenon* [30], i.e., the fact that, when projecting a state vector **x** onto the properly normalized random reservoir weight matrix **W**, the ℓ_2 norm of **Wx** is approximately equal to the ℓ_2 norm of **x**, when *N* is sufficiently large.

Let $\mathbf{W}_N = (w_{ij})_{N\times N}$, $\mathbf{W} = a\mathbf{W}_N$, and $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$. Suppose $\mathbf{W}_N$ follows (4)–(6), with a set to $(1/\sqrt{cN})$ under (4) or $(1/\sqrt{N})$ under (5) and (6). We have

$$\mathbf{W}\mathbf{x}=\left[\sum_j \alpha w_{1j}x_j, \sum_j \alpha w_{2j}x_j, \ldots, \sum_j \alpha w_{Nj}x_j\right]^T. \quad (13)$$

For the i^{th}-element, we have

$$E\left[\sum_j \alpha w_{ij}x_j\right]=\sum_j \alpha E[w_{ij}]x_j=0 \quad (14)$$

$$Var\left[\sum_j \alpha w_{ij}x_j\right]=E\left[\sum_j\sum_k \alpha^2 w_{ij}x_j w_{ik}x_k\right] =\frac{1}{N}\sum_j x_j^2 \quad (15)$$

where $E[\cdot]$ denotes expectation and $Var[\cdot]$ denotes variance.

Thus, using (14) and (15), we have

$$E\left[\|\mathbf{W}\mathbf{x}\|^2\right]=E\left[\sum_i\left(\sum_j \alpha w_{ij}x_j\right)^2\right] =\sum_i \frac{1}{N}\sum_j x_j^2=\|\mathbf{x}\|^2 \quad (16)$$

where $\|\cdot\|$ denotes the ℓ_2 norm, i.e., the expected squared length of $\mathbf{W}\mathbf{x}$ is the same as the squared length of $\mathbf{x}$. Now we need to investigate how the distribution of $\|\mathbf{W}\mathbf{x}\|$ concentrates around $\|\mathbf{x}\|$. We first develop the following lemma.

Lemma 1

Assume the random matrix $\mathbf{W}$ follows (4), (5), or (6), with the scaling factor set to $(1/\sqrt{cN})$ or $(1/\sqrt{N})$, as appropriate. Let $\mathbf{x}\hat{}\in\mathbb{R}^N$ be a unit vector, then, $\|\mathbf{W}\mathbf{x}\|\hat{}$ converges to 1 in probability, as $N\to\infty$.

Proof—[31, Lemmas 4 and 5] state that for $\mathbf{W}$ as in (5) (Lemma 4) and for $\mathbf{W}$ as in (4) and (6) (Lemma 5), the following two inequalities hold for all N and for all $0 < \varepsilon_2 < 1$

$$\Pr\left(\|\mathbf{W}\hat{\mathbf{x}}\|^2 \geq (1+\varepsilon_2)\right)<\exp\left(-\frac{N}{2}\left(\frac{\varepsilon_2^2}{2}-\frac{\varepsilon_2^3}{2}\right)\right) \quad (17)$$

$$\Pr\left(\|\mathbf{W}\hat{\mathbf{x}}\|^2 \leq (1-\varepsilon_2)\right) < \exp\left(-\frac{N}{2}\left(\frac{\varepsilon_2^2}{2}-\frac{\varepsilon_2^3}{2}\right)\right). \quad (18)$$

Then, for $0 < \varepsilon_3 < 1$

$$\begin{aligned}\Pr(|\|\mathbf{W}\hat{\mathbf{x}}\|-1| \geq \varepsilon_3) &= \Pr(\|\mathbf{W}\hat{\mathbf{x}}\| \leq 1-\varepsilon_3)+\Pr(\|\mathbf{W}\hat{\mathbf{x}}\| \geq 1+\varepsilon_3)\\ &= \Pr(\|\mathbf{W}\hat{\mathbf{x}}\|^2 \leq (1-\varepsilon_3)^2)+\Pr(\|\mathbf{W}\hat{\mathbf{x}}\|^2 \geq (1+\varepsilon_3)^2)\\ &< \Pr(\|\mathbf{W}\hat{\mathbf{x}}\|^2 \leq 1-\varepsilon_3)+\Pr(\|\mathbf{W}\hat{\mathbf{x}}\|^2 \geq 1+\varepsilon_3)\\ &< 2\exp\left(-\tfrac{N}{2}\left(\tfrac{\varepsilon_3^2}{2}-\tfrac{\varepsilon_3^3}{2}\right)\right).\end{aligned} \quad (19)$$

Therefore, as $N \to +\infty$, $\Pr(|\|\mathbf{W}\hat{\mathbf{x}}\| - 1| \geq \varepsilon_3) \to 0$.

Given the Lemma, we can now state and prove our contraction mapping main result.

Theorem 7

Assume the network defined in (2) and (3) with neuron activation function f = tanh, applied componentwise. Suppose that $\mathbf{x}(n), \tilde{\mathbf{x}}(n) \in [-1, 1]^N$, and $\mathbf{W}$ is a random reservoir weight matrix defined by $\mathbf{W} = a\mathbf{W}_N$, according to (4), (5), or (6), with $\alpha = \rho/\sqrt{N}$ under (5), (6), and $\alpha = \rho/\sqrt{cN}$ under (4), where $0 < \rho < 1$. Then

$$\Pr(\|\mathbf{x}(n+1)-\tilde{\mathbf{x}}(n+1)\| \leq \|\mathbf{x}(n)-\tilde{\mathbf{x}}(n)\|) > 1-\exp\left(-\frac{N}{2}\left(\frac{(1-\rho)^2}{2}-\frac{(1-\rho)^3}{2}\right)\right) \quad (20)$$

where $\mathbf{x}(n+1) = T(\mathbf{x}(n), \mathbf{u}(n+1))$ and $\tilde{\mathbf{x}}(n+1) = T(\tilde{\mathbf{x}}(n), \mathbf{u}(n+1))$.

Proof—Let $\mathbf{z}(n) = \mathbf{x}(n) - \tilde{\mathbf{x}}(n)$. We start by writing

$$\begin{aligned}\|\mathbf{z}(n+1)\| &= \|T(\mathbf{x}(n), \mathbf{u}(n+1)) - T(\tilde{\mathbf{x}}(n), \mathbf{u}(n+1))\|\\ &= \|f(\mathbf{W}\mathbf{x}(n)+\mathbf{W}_{\text{in}}\mathbf{u}(n+1)) - f(\mathbf{W}\tilde{\mathbf{x}}(n)+\mathbf{W}_{\text{in}}\mathbf{u}(n+1))\|\\ &\leq \|(\mathbf{W}\mathbf{x}(n)+\mathbf{W}_{\text{in}}\mathbf{u}(n+1)) - (\mathbf{W}\tilde{\mathbf{x}}(n)+\mathbf{W}_{\text{in}}\mathbf{u}(n+1))\|\\ &= \|\mathbf{W}\mathbf{x}(n) - \mathbf{W}\tilde{\mathbf{x}}(n)\|\\ &= \|\mathbf{W}(\mathbf{x}(n) - \tilde{\mathbf{x}}(n))\|\\ &= \|\mathbf{W}\mathbf{z}(n)\|\end{aligned} \quad (21)$$

where the inequality four lines above follows because the tanh(·) function satisfies the (element-wise) Lipschitz condition $|\tanh(\upsilon) - \tanh(z)| \leq |\upsilon - z|, \forall \upsilon, z \in \mathbb{R}$. Let $\hat{\mathbf{z}}(n) = \mathbf{z}(n)/\|\mathbf{z}(n)\|$. Then rewrite (21) as

$$\|\mathbf{z}(n+1)\| \leq \|\mathbf{W}\mathbf{z}(n)\| = \|\mathbf{W}\hat{\mathbf{z}}(n)\| \cdot \|\mathbf{z}(n)\|.$$

We have $\mathbf{W} = \alpha\mathbf{W}_N$, and $\mathbf{W}_N$ generated according to (4), (5), or (6), with α equaling ($\rho/\sqrt{cN}$), ($\rho/\sqrt{N}$), or ($\rho/\sqrt{N}$), respectively. From the circular law and Theorem 4, we know that the spectral radius of $\mathbf{W}$ converges to ρ as $N \to \infty$.

Applying Lemma 1, we thus have

$$\begin{aligned}\|\mathbf{z}(n+1)\| &\le \|\mathbf{W}\hat{\mathbf{z}}(n)\| \cdot \|\mathbf{z}(n)\| \\ &= \rho\|\tfrac{1}{\rho}\mathbf{W}\hat{\mathbf{z}}(n)\| \cdot \|\mathbf{z}(n)\| \xrightarrow{p} \rho\|\mathbf{z}(n)\| \quad \text{as } N \to \infty.\end{aligned}$$

Further, let us characterize the probability that $\|\mathbf{z}(n + 1)\| \ge \|\mathbf{z}(n)\|$, i.e., that the contractive property is *not* satisfied, when N is finite. First, define $\varepsilon = 1 - \rho$. Then, we have

$$\begin{aligned}\Pr(\|\mathbf{z}(n+1)\| \ge \|\mathbf{z}(n)\|) &\le \Pr(\|\mathbf{W}\mathbf{z}(n)\| \ge \|\mathbf{z}(n)\|) \\ &= \Pr(\|\mathbf{W}\hat{\mathbf{z}}(n)\| \ge 1) \\ &= \Pr\left(\|\tfrac{1}{\rho}\mathbf{W}\hat{\mathbf{z}}(n)\| \ge \tfrac{1}{\rho}\right) \\ &= \Pr\left(\|\tfrac{1}{\rho}\mathbf{W}\hat{\mathbf{z}}(n)\| \ge \tfrac{1}{1-\varepsilon}\right) \\ &< \Pr\left(\|\tfrac{1}{\rho}\mathbf{W}\hat{\mathbf{z}}(n)\| \ge 1+\varepsilon\right) \quad \left(\because 1+\varepsilon < \tfrac{1}{1-\varepsilon}\right) \\ &= \Pr\left(\|\tfrac{1}{\rho}\mathbf{W}\hat{\mathbf{z}}(n)\|^2 \ge (1+\varepsilon)^2\right) \\ &\le \Pr\left(\|\tfrac{1}{\rho}\mathbf{W}\hat{\mathbf{z}}(n)\|^2 \ge 1+\varepsilon\right) \\ &< \exp\left(-\tfrac{N}{2}\left(\tfrac{(1-\rho)^2}{2} - \tfrac{(1-\rho)^3}{2}\right)\right)\end{aligned}$$

where the first inequality above follows from (21) and the final inequality follows from [31, Lemmas 4 and 5], specified earlier.

We thus see that the probability that $\|\mathbf{z}(n + 1)\| > \|\mathbf{z}(n)\|$ is exponentially decreasing with N. Moreover, $\|\mathbf{z}(n + 1)\| \le \|\mathbf{z}(n)\|$ with probability $1-O(e^{-C_\rho N})$, where $C_\rho = (1/2)((1 - \rho)^2/2 - (1 - \rho)^3/2)$.

Theorem 7 shows that, when $\rho < 1$, for $\mathbf{x}(n)$, $\mathbf{x}(\tilde{n}) \in [-1, 1]^N$ and a random reservoir weight matrix $\mathbf{W}$, $T(\cdot, \cdot)$ is contractive with probability $1-O(e^{-C_\rho N})$. This result supports and provides theoretical grounding for previous observations in ESN research: "extensive experience with this scaling game indicates that one obtains echo states when α is only marginally smaller than $\alpha_{\max}$" [2] ($\alpha_{\max} = |\lambda_{\max}^{-1}(\mathbf{W}_N)|$). To give a caveat on this result, we also note that, while we have shown that there is a contractive property with high probability for large N, Theorem 7 is not definitive on whether the *strict* ESP given in Definition 1 holds with high probability for large N. This remains an open question.

Finally, let us address the reader's possible concern that, in Theorem 7, we have assumed an unorthodox way of selecting the scaling factor on the weight matrix, in order to achieve our proof result, i.e., the reader may think our result is not relevant to the more conventional matrix scaling procedure. To address this concern, we next show that, from a practical standpoint, the result and insights obtained from Theorem 7 *also* apply if one considers the

more conventional scheme for scaling **W**. The logical argument goes as follows. There are two choices for the scaling factor—the conventional choice $\alpha=\rho|\lambda_{\max}^{-1}(\mathbf{W}_N)|$, and our unorthodox choice $\alpha=\rho/\sqrt{N}$ (for simplicity of discussion, we only consider the Gaussian and Bernoulli cases). Suppose that we could show that the spectral radius resulting from conventional scaling (which is the constant value, ρ) is *always* (for every realization of $\mathbf{W}_N$) less than or equal to the spectral radius resulting from our unorthodox scaling procedure. If this were true, one can see (from inspection of the proof of Theorem 7) that the Theorem 7 statement would directly apply not *only* for our unorthodox scaling procedure, but also for the more conventional scaling scheme. Likewise, if ρ is larger than the unorthodox scheme's spectral radius with vanishing probability as *N* gets large, we could say that Theorem 7 "practically applies" to conventional scaling, for large *N*. Let us consider two cases: 1) asymptotically large *N*, and 2) relatively large finite (but increasing) *N*. For the asymptotic case, we simply note that, from the proof steps of Theorem 6, we know that the spectral radius obtained using these two different scaling methods converges to the *same* value (ρ) as $N \to \infty$. Thus, Theorem 7 is certainly relevant to the conventional scaling procedure in the limit of large *N*. Second, let us consider the case of finite (but increasing) *N*. There are two choices for the scaling factor—the conventional choice $\alpha=\rho|\lambda_{\max}^{-1}(\mathbf{W}_N)|$, and our unorthodox choice $\alpha=\rho/\sqrt{N}$. Now, it is not in fact true that ρ is *strictly* less (for all realizations $\mathbf{W}_N$) than the spectral radius obtained based on our unorthodox scaling. However, empirically, we will next demonstrate the following results: 1) For large but finite *N*, the frequency with which conventional scaling leads to a larger spectral radius than unorthodox scaling is quite small, moreover, the "spread" of the unorthodox scaling's spectral radius distribution (around ρ) is small, and 2) This frequency is observed to *decrease* with increasing *N*.

We simulated 10 000 trials for each of the three types of reservoirs, for $N = 500$, 1000, and 1500. We set $\rho = 0.91$ and observed that, using unorthodox scaling, for $N = 1000$ and $N = 1500$, the necessary ESP condition was met in every trial (with a small number of violations for $N = 500$). Our results, shown in Table I, demonstrate that, very infrequently, ρ is greater than the spectral radius of the unconventional procedure. Furthermore, this frequency decreases for increasing *N*. Fig. 4 shows the distribution of the unconventional procedure's spectral radius which, though skewed, is seen to have small spread about ρ. These experimental results suggest that Theorem 7 "practically applies" to conventional scaling as *N* gets large. The results also further corroborate our previous observation that, for finite *N*, the mean of the spectral radius seems to converge from above to 1.

V. Conclusion

In this brief, we applied random matrix theory to examine the properties of the random reservoirs used by ESNs, including different reservoir topologies (sparse or fully connected) and different connection weights (Bernoulli or Gaussian). The asymptotic uniform distribution of the eigenvalues of the reservoir weight matrix ensures diverse dynamical patterns of the reservoir states. Moreover, this phenomenon does not depend on the topology of the reservoir or on the distribution of the weights of the connections. We showed that, asymptotic in the reservoir size, the bound for the necessary condition in [2] is about twice

the bound for the sufficient condition in [2] for an ESN to possess the ESP. Finally, we showed that, when the spectral radius $\rho < 1$, the state transition mapping $T(\cdot, \cdot)$ is contractive with high probability, which explains why the necessary condition has been found to be "sufficient in practice."

Acknowledgments

The authors would like to thank W. T. Baumann for his valuable suggestions.

This work was supported in part by the National Institutes of Health, under Grant CA149147 and Grant NS029525.

References

1. Jaeger H, Haas H. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. Science. Apr; 2004 304(5667):78–80. [PubMed: 15064413]
2. Jaeger, H. Tech Rep. German National Research Center Information Technology; St. Augustin, Germany: 2001. The 'echo state' approach to analysing and training recurrent neural networks; p. 148
3. Maass W, Natschläger T, Markram H. Real-time computing without stable states: A new framework for neural computation based on perturbations. Neural Comput. Nov; 2002 14(11):2531–2560. [PubMed: 12433288]
4. Shi Z, Han M. Support vector echo-state machine for chaotic time-series prediction. IEEE Trans Neural Netw. Mar; 2007 18(2):359–372. [PubMed: 17385625]
5. Ozturk MC, Príncipe JC. An associative memory readout for ESNs with applications to dynamical pattern recognition. Neural Netw. Apr; 2007 20(3):377–390. [PubMed: 17513087]
6. Skowronski MD, Harris JG. Automatic speech recognition using a predictive echo state network classifier. Neural Netw. Apr; 2007 20(3):414–423. [PubMed: 17556115]
7. Zhang, B.; Wang, Y. Echo state networks with decoupled reservoir states. Proc. IEEE Workshop Mach. Learn. Signal Process; Oct. 2008; p. 444-449.
8. Deng Z, Zhang Y. Collective behavior of a small-world recurrent neural system with scale-free distribution. IEEE Trans Neural Netw. Sep; 2007 18(5):1364–1375. [PubMed: 18220186]
9. Xue Y, Yang L, Haykin S. Decoupled echo state networks with lateral inhibition. Neural Netw. 2007; 20(3):365–376. [PubMed: 17517490]
10. Ozturk MC, Xu D, Príncipe JC. Analysis and design of echo state networks. Neural Comput. Jan; 2007 19(1):111–138. [PubMed: 17134319]
11. Xia Y, Jelfs B, Van Hulle M, Príncipe J, Mandic D. An augmented echo state network for nonlinear adaptive filtering of complex noncircular signals. IEEE Trans Neural Netw. Jan; 2011 22(1):74–83. [PubMed: 21075724]
12. Chatzis S, Demiris Y. Echo state Gaussian process. IEEE Trans Neural Netw. Sep; 2011 22(9): 1435–1445. [PubMed: 21803684]
13. Rodan A, Tiňo P. Minimum complexity echo state network. IEEE Trans Neural Netw. Jan; 2011 22(1):131–144. [PubMed: 21075721]
14. Lukoševicĭus M, Jaeger H. Reservoir computing approaches to recurrent neural network training. Comput Sci Rev. Aug; 2009 3(3):127–149.
15. Elman JL. Finding structure in time. Cognit Sci. 1990; 14(2):179–211.
16. Jordan, MI. Attractor dynamics and parallelism in a connectionist sequential machine. In: Diederich, J., editor. Artificial Neural Networks: Concept Learning. IEEE Computer Society Neural Networks Technology; Los Alamitos, CA: IEEE Computer Society Press; Jan. 1990 p. 112-127.
17. Buehner M, Young P. A tighter bound for the echo state property. IEEE Trans Neural Netw. May; 2006 17(3):820–824. [PubMed: 16722187]
18. Jaeger H. Echo state network. Scholarpedia. 2007; 2(9):2330.

19. Mehta, M. Random Matrices and the Statistical Theory of Energy Levels. New York: Academic; 1967.
20. Edelman A. The probability that a random real Gaussian matrix has *k* real eigenvalues, related distributions, and the circular law. J Multivar Anal. Feb; 1997 60(2):203–232.
21. Girko VL. Circular law. Theory Probab Appl. 1984; 29(4):694–706.
22. Bai ZD. Circular law. Ann Probab. 1997; 25(1):494–529.
23. Tao T, Vu V. Random matrices: The circular law. Commun Contemp Math. 2008; 10(2):261–307.
24. Tao T, Vu V, Krishnapur M. Random matrices: Universality of ESDs and the circular law. Ann Prob. 2010; 38(5):2023–2065.
25. Tao T, Vu V. Random matrices: The distribution of the smallest singular values. Geometric Funct Anal. Mar; 2010 20(1):1–43.
26. Marcenko VA, Pastur LA. Distribution of eigenvalues for some sets of random matrices. Math USSR-Sbornik. Apr; 1967 1(4):457–483.
27. Yin YQ. Limiting spectral distribution for a class of random matrices. J Multivar Anal. Oct; 1986 20(1):50–68.
28. Bai ZD, Yin YQ. Limiting behavior of the norm of products of random matrices and two problems of Geman-Hwang. Probab Theory Rel Fields. 1986; 73(4):555–569.
29. Yin YQ, Bai ZD, Krishnaiah PR. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. Prob Theory Rel Fields. 1988; 78(4):509–521.
30. Ledoux, M. The Concentration of Measure Phenomenon. Providence, RI: AMS; 2001.
31. Achlioptas, D. Database-friendly random projections. Proc. 20th ACM SIGMOD-SIGACT-SIGART Symp. Principles Database Syst.; 2001. p. 274-281.

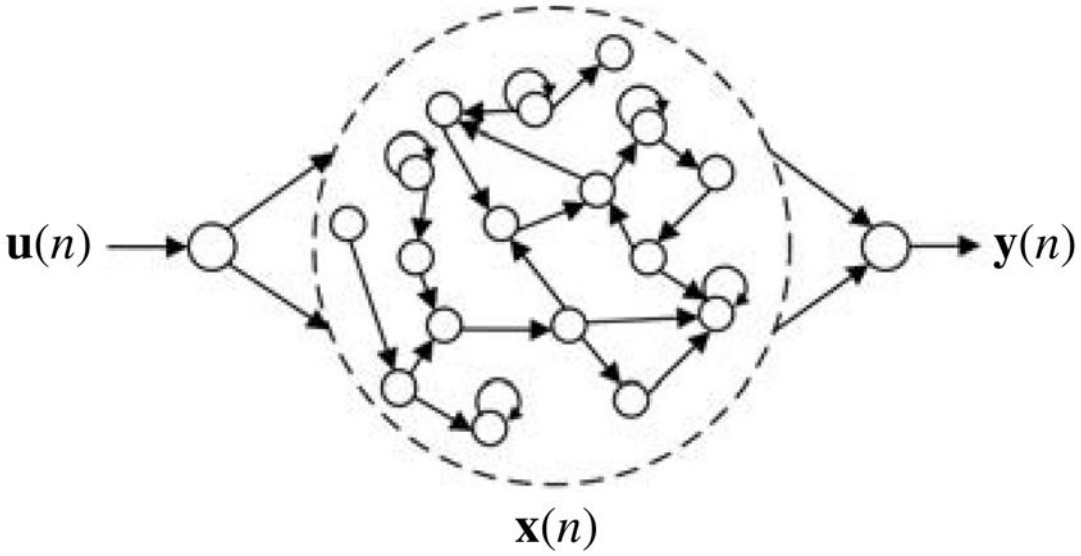

Fig. 1.
Illustration of an ESN.

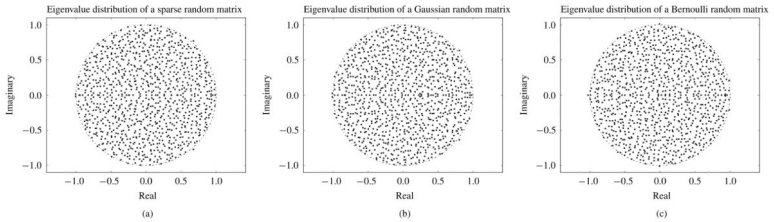

Fig. 2.

Empirical eigenvalue distributions of three types of random matrices ($N = 1000$). (a) Sparse random matrix. (b) Gaussian random matrix. (c) Bernoulli random matrix.

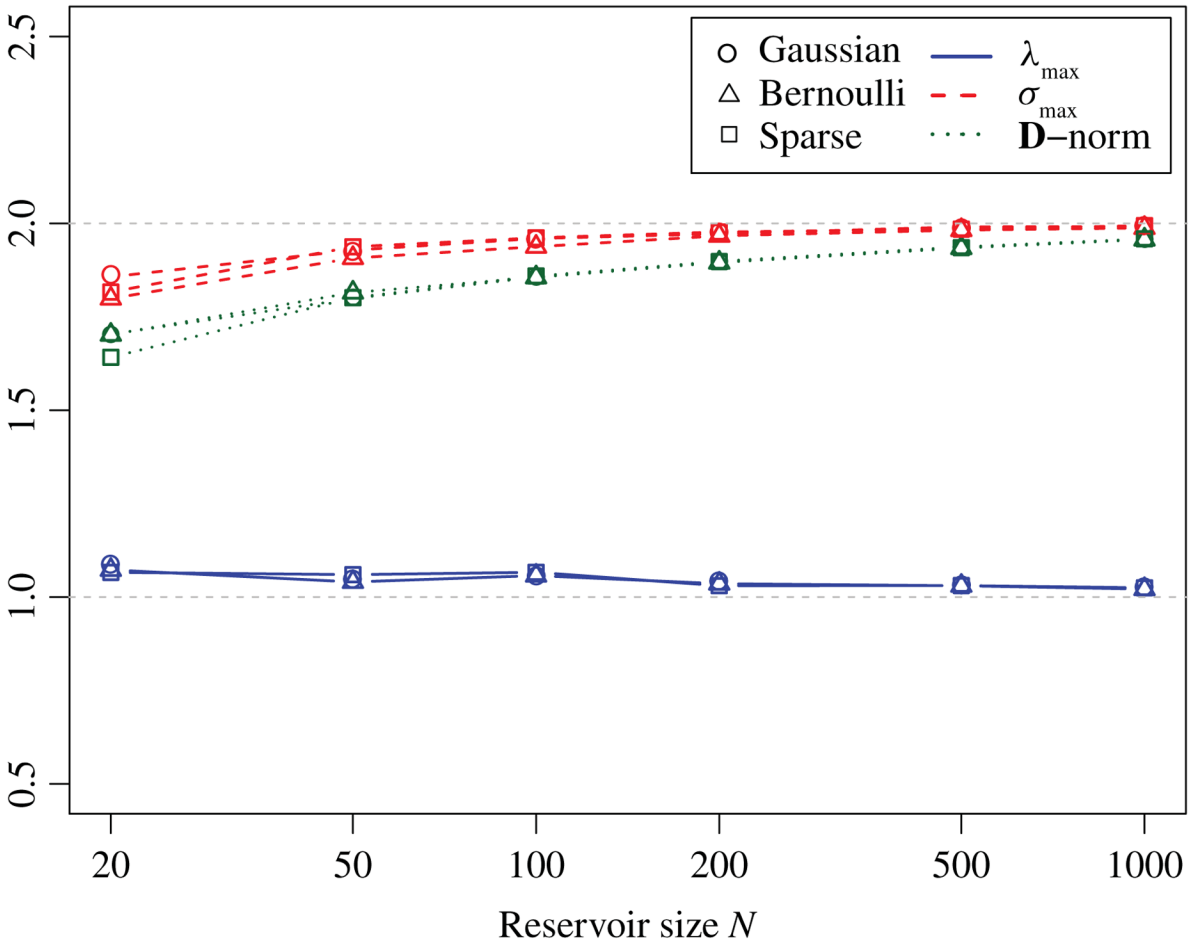

Fig. 3.

Simulation study on $\sigma_{\max}(\mathbf{W})$, $\lambda\max(\mathbf{W})$, and $\|\mathbf{W}\|_D$ for Gaussian, Bernoulli, and sparse reservoirs, respectively, as N increases.

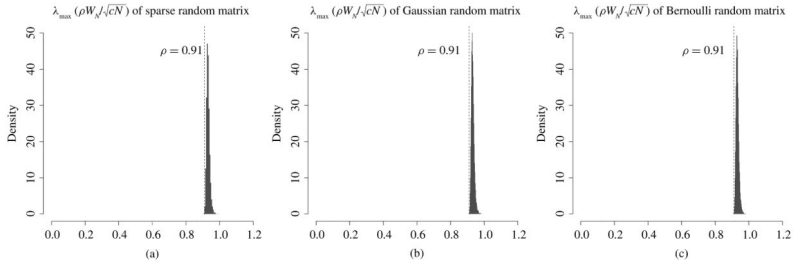

Fig. 4.

Histograms of the spectral radius of random matrices using the scaling factor in Theorem 7 with $\rho = 0.91$ and $N = 1000$. (a) Sparse random matrices. (b) Gaussian random matrices. (c) Bernoulli random matrices.

TABLE I

Spectral Radius ($\hat{\rho}$) of Random Matrices Using the Scaling Factor in Theorem 7 with $\rho = 0.91$ (10 000 Trials)

	Sparse random reservoir			Gaussian random reservoir			Bernoulli random reservoir		
N	$\Pr(\hat{\rho} \geq 1)$	$\Pr(\hat{\rho} < \rho)$	Mean($\hat{\rho}$) (std)	$\Pr(\hat{\rho} \geq 1)$	$\Pr(\hat{\rho} < \rho)$	Mean($\hat{\rho}$) (std)	$\Pr(\hat{\rho} \geq 1)$	$\Pr(\hat{\rho} < \rho)$	Mean($\hat{\rho}$) (std)
500	0.09%	0.38%	0.938 (±0.014)	0.13%	0.46%	0.938 (±0.014)	0.08%	0.29%	0.938 (±0.013)
1000	0.00%	0.07%	0.932 (±0.009)	0.00%	0.03%	0.932 (±0.009)	0.00%	0.01%	0.932 (±0.009)
1500	0.00%	0.00%	0.929 (±0.007)	0.00%	0.01%	0.928 (±0.007)	0.00%	0.01%	0.928 (±0.007)