[5] J. Cao and J. Liang, "Boundedness and stability for Cohen–Grossberg neural network with time-varying delays," *J. Math. Anal. Appl.*, vol. 296, no. 2, pp. 665–685, 2004.
[6] T. Chen and L. Rong, "Delay-independent stability analysis of Cohen–Grossberg neural networks," *Phys. Lett. A*, vol. 317, no. 5–6, pp. 436–449, 2003.
[7] M. A. Cohen and S. Grossberg, "Absolute stability and global pattern formation and parallel memory storage by competitive neural networks," *IEEE Trans. Syst., Man, Cybern.*, vol. 13, no. 5, pp. 815–821, 1983.
[8] A. Friedman, *Stochastic Differential Equations and Their Applications*. New York: Academic, 1976.
[9] P. Gahinet, A. Nemirovsky, A. J. Laub, and M. Chilali, *LMI Control Toolbox: For Use With Matlab*. Natick, MA: The Math Works, Inc., 1995.
[10] F. Gu, "An integral inequality in the stability problem of time-delay systems," in *Proc. IEEE 39th Conf. Decision Contr.*, Sydney, Australia, Dec. 2000, pp. 2805–2810.
[11] J. K. Hale, *Theory of Functional Differential Equations*. New York: Springer-Verlag, 1977.
[12] H. Huang, D. W. C. Ho, and J. Lam, "Stochastic stability analysis of fuzzy Hopfield neural networks with time-varying delays," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 52, no. 5, pp. 251–255, May 2005.
[13] C.-C. Hwang, C.-J. Cheng, and T.-L. Liao, "Globally exponential stability of generalized Cohen–Grossberg neural networks with delays," *Phys. Lett. A*, vol. 319, no. 1–2, pp. 157–166, 2003.
[14] Y. Li, "Existence and stability of periodic solutions for Cohen–Grossberg neural networks with multiple delays," *Chaos Sol. Fract.*, vol. 20, no. 3, pp. 459–466, 2004.
[15] Y. Liu, Z. Wang, and X. Liu, "Global exponential stability of generalized recurrent neural networks with discrete and distributed delays," *Neural Netw.*, 2006, to be published.
[16] W. Lu and T. Chen, "Dynamical behaviors of Cohen–Grossberg neural networks with discontinuous activation functions," *Neural Netw.*, vol. 18, no. 3, pp. 231–242, 2005.
[17] C. M. Marcus and R. M. Westervelt, "Stability of analog neural networks with delay," *Phys. Rev. A*, vol. 39, pp. 347–359, 1989.
[18] L. Rong, "LMI-based criteria for robust stability of Cohen–Grossberg neural networks with delay," *Phys. Lett. A*, vol. 339, no. 1–2, pp. 63–73, 2005.
[19] S. Ruan and R. S. Filfil, "Dynamics of a two-neuron system with discrete and distributed delays," *Physica D*, vol. 191, pp. 323–342, 2004.
[20] J. Sun and L. Wan, "Global exponential stability and periodic solutions of Cohen–Grossberg neural networks with continuously distributed delays," *Physica D: Nonlin. Phenom.*, vol. 208, no. 1–2, pp. 1–20, 2005.
[21] D. W. Tank and J. J. Hopfield, "Neural computation by concentrating information in time," *Proc. Nat. Acad. Sci.*, vol. 84, pp. 1896–1991, 1987.
[22] L. Wan and J. Sun, "Mean square exponential stability of stochastic delayed Hopfield neural networks," *Phys. Lett. A*, vol. 343, no. 4, pp. 306–318, 2005.
[23] L. Wang, "Stability of Cohen–Grossberg neural networks with distributed delays," *Appl. Math. Comput.*, vol. 160, no. 1, pp. 93–110.
[24] Z. Wang, Y. Liu, and X. Liu, "On global asymptotic stability of neural networks with discrete and distributed delays," *Phys. Lett. A*, vol. 345, no. 4–6, pp. 299–308, 2005.
[25] Z. Wang, H. Shu, Y. Liu, D. W. C. Ho, and X. Liu, "Robust stability analysis of generalized neural networks with discrete and distributed time delays," *Chaos Solitons Fractals*, 2006, to be published.
[26] W. Xiong and J. Cao, "Global exponential stability of discrete-time Cohen–Grossberg neural networks," *Neurocomputing*, vol. 64, pp. 433–446, 2005.

# A Tighter Bound for the Echo State Property

Michael Buehner and Peter Young

***Abstract*—This letter provides a brief explanation of echo state networks (ESNs) and provides a rigorous bound for guaranteeing asymptotic stability of these networks. The stability bounds presented here could aid in the design of echo state networks that would be applicable to control applications where stability is required.**

***Index Terms*—Echo state networks (ESNs), Lyapunov stability, nonlinear systems, recurrent neural networks (RNN), robust controls, weighted operator norms.**

## I. Introduction

Artificial neural networks may be used in control system applications such as modeling nonlinear system dynamics and control of nonlinear systems. Two types of networks are used for these applications, namely feed-forward neural networks (FFNN) and recurrent neural networks (RNN). FFNN are attractive since they are easy to train in a stable manner (e.g., using back propagation), but are limited in the sense that they are only capable of providing a static map between inputs and outputs (i.e., they have no way of internally representing the dynamics of a nonlinear system). In contrast, RNN may be very difficult (and take a long time) to train stably; however, the recurrent connections of a RNN form a dynamical system. It is this dynamical nature of RNN that allows them to capture the dynamics of a nonlinear system, which makes them more applicable to nonlinear system modeling and control. Examples of RNN are Hopfield networks, Elman networks, liquid state machines, and echo state networks (ESNs), the last-mentioned being the focus of this letter. For a review of these RNN and the problems associated with training them, see either [1], [2], or [3].

The recent development of echo state networks [4] ESNs provides a class of RNN that alleviates the problem of training, but the design methodology of ESNs is still not fully understood. ESNs are characterized by their ability to uniquely map a temporal input history to an "echo state." An ESN that has this characteristic is said to have the *echo state property*. Currently, the echo state property may be verified from two sufficient conditions, namely one for the existence of echo states for all inputs and one for the nonexistence of echo states for certain inputs. This letter reformulates and further develops these conditions into separate necessary and sufficient condition for the existence of echo states for all inputs. As mentioned in [4], the current sufficient condition for the existence of echo states appears, in practice, to be rather restrictive. This problem is addressed by deriving a new sufficient condition that is less conservative. Specifically, a result that is well known in the *Robust Controls* community is used to reduce the conservatism and in some cases make the bounds tight (i.e., provide a single bound that is both necessary and sufficient). This letter concludes with some simple simulations to demonstrate the improvement that can be achieved by using the new sufficient condition.

## II. Notation

Let $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$ be either the field of complex or real numbers, respectively. For any square matrix $W \in \mathbb{F}^{n \times n}$, let $\overline{\sigma}(W)$ denote

Manuscript received August 2, 2005; revised November 6, 2005. This work was supported by the National Science Foundation under Award 0 245 291.

The authors are with the Electrical and Computer Engineering Department, Colorado State University (CSU), Fort Collins, CO 80523 USA (e-mail: mbuehner@engr.colostate.edu; pmy@engr.colostate.edu).

Digital Object Identifier 10.1109/TNN.2006.872357

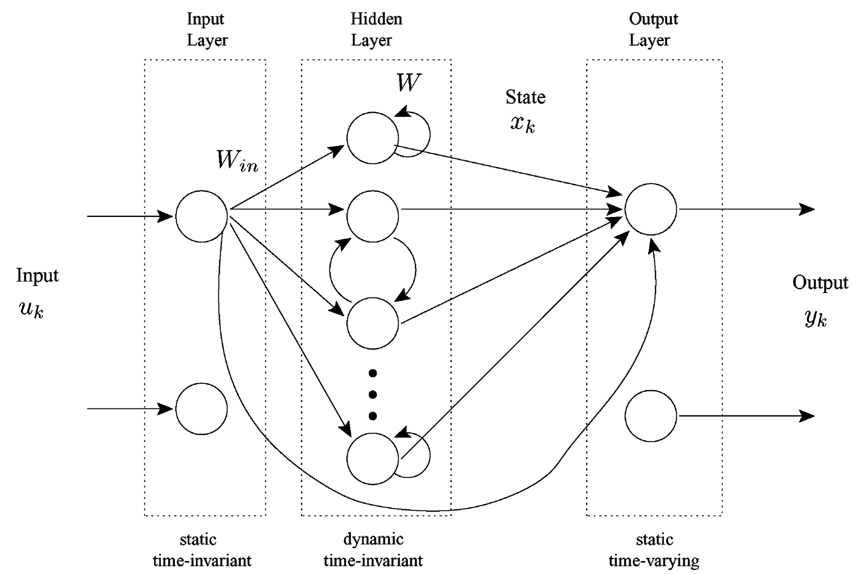


Fig. 1. ESN architecture.

the largest singular value and let $\rho(W) \doteq \max\{|\lambda| : \lambda$ is an eigenvalue of $W\}$ denote the spectral radius of $W$. For any square matrix $D \in \mathbb{F}^{n\times n}$, let $\mathcal{N}(D)$ denote the null space of $D$. A column vector $x_k \in \mathbb{F}^n$ is a vector of length $n$ at discrete-time instance $k$ (i.e., $x_k \equiv [x_{1k}, x_{2k}, \ldots, x_{nk}]^T$, where $[\cdot]^T$ is the standard matrix transpose). The set $U^{-N}$ contains all inputs of $N$ previous time steps $\{u_{k-N}, \ldots, u_{k-1}, u_k\}$ that are bounded by some uniform constant; it is therefore compact. Similarly, the sets $U^{-\infty}$ and $U^{+\infty}$ contains all left and right infinite input sequences $\{\ldots, u_{k-1}, u_k\}$ and $\{u_k, u_{k+1}, \ldots\}$, respectively, which are bound by some uniform constant; they are compact, too. Certain scaling matrices will be used. The classes of these matrices will be denoted as $\mathcal{D} \equiv \mathbb{F}^{n\times n}$ and $\overline{\mathcal{D}} = \{\mathrm{diag}(\delta_1, \ldots, \delta_n), \delta_i \in \mathbb{F}$ for $i = 1, \ldots, n\}$. To avoid confusion, $D_\delta$ will be used to represent diagonal matrices in $\overline{\mathcal{D}}$ and $D$ will be used to represent full matrices in $\mathcal{D}$. $I_n$ is an $n \times n$ identity matrix. The subscript for this will be dropped when the dimensions are obvious from context. The function $f(\cdot)$ applied to a vector is an element-by-element operation (i.e., $f(x) \equiv [f(x_1), f(x_2) \ldots f(x_n)]^T$). As a final note, all of the results are valid for $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$; however, in the context of an ESN, it is typically assumed that $\mathbb{F} = \mathbb{R}$.

## III. Prior Work

An ESN is a multilayered network that consists of a feed-forward input layer, a recurrently connected hidden layer, and a feed-forward output layer. This is shown in Fig. 1.

The input layer weights and the hidden layer weights are fixed (i.e., time-invariant) and only the output layer weights are trained (i.e., time-varying). This allows for a neural network that has the dynamic modeling capabilities from the (stable) recurrent connections in the hidden layer and the stable adaptive capabilities from the feed-forward network in the output layer. As will be shown, the echo state property holds for ESNs with asymptotically stable recurrent connections.

The input layer (with weights $W_{\mathrm{in}} \in \mathbb{R}^{n\times m}$) is used to map the lower dimensional inputs $u_k \in \mathbb{R}^m$ to a larger dimension $W_{\mathrm{in}} u_k \in \mathbb{R}^n$ (i.e., $n > m$). In practice, the weights $W_{\mathrm{in}}$ do not appear to have an impact on the performance [4]. The echo states $x_k$ are generated from the input layer and the recurrently and interconnected neurons in the hidden layer. Specifically, for a given an input vector $u_k$, the transition from one (echo) state vector $x_{k-1}$ to the next (echo) state vector $x_k$ is defined to be

$$x_k = T(x_{k-1}, u_k) = f(W_{\mathrm{in}} u_k + W x_{k-1}) \tag{1}$$

where $W \in \mathbb{R}^{n\times n}$ is a square matrix (containing the hidden layer weights) that is applied to the state vector $x_{k-1}$, and $f(\cdot)$ is a nonlinear "squashing" function that is applied to every element of the vector $W_{\mathrm{in}} u_k + W x_k$. The state vectors $x_k$ are fed into the final output layer. In the general statement of an ESN, feedback connections from the output layer to the hidden layer are allowed; however, in most applications, these feedback connections are not used. The theory for dealing with these connections will not be explored here.

The hidden layer weight matrix $(W)$ is a randomly generated sparse matrix (i.e., usually only 5% to 20% of the entries are nonzero) that is used to generate a (random) basis for the echo states $x_k$. The properties of this matrix may be used to determine if an ESN has the echo state property. The definition of an ESN (i.e., an ESN that has the echo state property) is given next.

*Definition (Jaeger [4]):* Assume standard compactness conditions (i.e., the inputs $u_k$ and states $x_k$ come from the compact sets $U^{-\infty}$ and $X^{-\infty}$, respectively). Assume that the network has no output feedback connections. Then, the ESN has echo states (i.e., the echo state property) if every echo state vector $x_k$ is uniquely determined for every left infinite input sequence $u^{-\infty} \in U^{-\infty}$.

This definition implies that nearby echo states must represent similar input histories. In turn, this means that the echo states should depend more heavily on the most recent inputs and states. Intuitively, this means that the state-space is not disjoint, and the echo states represent the current dynamics, or state, of the system the ESN is modeling. In [4], this property is shown to be satisfied by requiring the echo states to have a certain convergence property. This is stated more formally in the following discussion.

In Jaeger's paper [4], the existence of echo states may be verified in terms of separate necessary and sufficient conditions on the (square) hidden layer weight matrix $W \in \mathbb{R}^{n\times n}$. The necessary condition is $\rho(W) < 1$ and the sufficient condition is $\bar{\sigma}(W) < 1$. In [4], the necessary condition is stated as a sufficient condition for the nonexistence of echo states when $\rho(W) > 1$. The reason for this constraint is that if the underlying linear system is unstable, then the nonlinear system (resulting from the application of the squashing function) will also exhibit instability. From this point of view, the sufficient condition for the nonexistence of echo states is really for $\rho(W) \geq 1$. The necessary condition for the existence of echo states results from this. From a systems point of view, this requires that the nonlinear recurrent system be locally asymptotically stable at the origin. For global asymptotic stability, a more restrictive (sufficient) condition is required.[1] The original proof of the sufficient condition (taken from [4]) is outlined next.

Let $x_k$ and $\tilde{x}_k$ be two distinct state vectors and $y_k = x_k - \tilde{x}_k$. From an equivalent definition, echo states exist if the states $x_k$ and $\tilde{x}_k$ satisfy the convergence property $\|y_k\| \to 0$ as $k \to \infty$ for all right infinite input sequences $u^{+\infty} \in U^{+\infty}$. Note that in this definition, the vector norm is not specified; however, in Jaeger's proof, the standard Euclidean norm (i.e., the 2-norm) is used. Since all finite dimensional norms are equivalent, proving convergence in the 2-norm guarantees convergence in every other (finite dimensional) norm. Jaegers's original sufficient condition is now (re)stated as Theorem 1.

*Theorem 1 (Jaeger [4]):* Let an ESN have a fixed internal weight matrix $W \in \mathbb{R}$ and let $f(x) = \tanh(x)$. If $\overline{\sigma}(W) < 1$, then the network has the echo state property, i.e., $\lim_{k\to\infty} \|y_k\|_2 = 0$ for all right infinite input sequences $u^{+\infty} \in U^{+\infty}$.

*Proof:*

$$\begin{aligned}
\|y_{k+1}\|_2 &= \|x_{k+1} - \tilde{x}_{k+1}\|_2 \\
&= \|T(x_k, u_k) - T(\tilde{x_k}, u_k)\|_2 \\
&= \left\| f(W^{\text{in}} u_k + W x_k) - f(W^{\text{in}} u_k + W\tilde{x}_k) \right\|_2 \\
&\leq \left\| (W^{\text{in}} u_k + W x_k) - (W^{\text{in}} u_k + W\tilde{x}_k) \right\|_2 \qquad (2) \\
&= \|W x_k - W\tilde{x}_k\|_2 \\
&= \|W(x_k - \tilde{x}_k)\|_2 \\
&\leq \|W\|_2 \|y_k\|_2 \\
&= \bar{\sigma}(W)\|y_k\|_2 \qquad (3)
\end{aligned}$$

where $W$ satisfies the contraction property $\bar{\sigma}(W) < 1$. Note that this clearly implies the required convergence property, namely $\lim_{k\to\infty} \|y_k\|_2 = 0$. ■

*Remark:* Note that (2) assumes that the squashing function will shrink *every* element of the vectors $W^{\text{in}} u_k + W x_k$ and $W^{\text{in}} u_k + W\widetilde{x}_k$ toward zero. Therefore, the difference of the "un-squashed" version will have a larger norm than the difference of the "squashed" version. In the proof given in [4], the squashing function is assumed to be $f(x) = \tanh(x)$; however, it is mentioned that any function satisfying the (element-wise) Lipshitz condition $|f(v) - f(z)| \leq |v - z| \forall v$, $z \in \mathbb{R}$ will do. As a side note, if $f(\cdot)$ is differentiable, then the Lipshitz condition is equivalent to $|f'(v)| \leq 1 \forall v \in \mathbb{R}$.

This proof yields a rather conservative sufficient condition [4]. In this letter, a less restrictive sufficient condition is derived by considering a different norm.

## IV. The Weighted Operator Norm

In linear algebra, it is a well-known fact that there exists an operator norm (sometimes referred to as an induced norm) for a matrix that is arbitrarily close to the spectral radius of the matrix. This is summarized in Lemma 2.

*Lemma 2:* For every matrix $W \in \mathbb{F}^{n\times n}$ and for every $\epsilon > 0$, there exists an operator norm $\|\cdot\|_D$ such that

$$\rho(W) \leq \|W\|_D \leq \rho(W) + \epsilon. \qquad (4)$$

*Proof:* See [5]. ■

[1]For an ESN, bounded-input–bounded-out (BIBO) stability is satisfied trivially, since outputs from the squashing functions are bounded for all inputs. Here, the echo state property is defined in terms of an global asymptotic stability requirement on the echo state vectors

*Remark:* This operator norm is achieved by choosing an appropriate weighted operator norm, namely $\|W\|_D = \|DWD^{-1}\|$ with $D \in \mathbb{F}$ nonsingular, that is specific to the matrix $W$. This weighted operator norm does not depend on the underlying norm used (e.g., any of the $p$-norms such as $p = 1$, 2, or $\infty$), but rather on the weighting matrix $D \in \mathbb{F}$ that is selected based on the matrix $W$. Note that all finite-dimensional norms are equivalent. Therefore, choosing a different norm may require a different $D$; however, the property will hold for any chosen norm. For computational reasons, the 2-norm will be used.

### A. The Vector D-Norm

The $D$-norm of a vector $x \in \mathbb{F}^n$ is defined to be $\|x\|_D = \|Dx\|$, where $D \in \mathbb{F}^{n\times n}$ is nonsingular and $\|\cdot\|$ is a vector norm (e.g., one of the $p$-norms). It is easy to show that $\|\cdot\|_D$ is in fact a vector norm provided $D$ is nonsingular [6]. In this letter, the $D$-norm will be defined in terms of the weighted 2-norm as $\|x\|_D = \|Dx\|_2$, where $D$ is an arbitrary nonsingular matrix to be chosen later.

### B. The Matrix Operator $D$-Norm

Let $x = D^{-1}y$, where $y \in \mathbb{F}^n$. Then, the induced $D$-norm of a matrix $W \in \mathbb{F}^{n\times n}$ is given as

$$\begin{aligned}
\|W\|_D &= \sup_{x\neq 0} \frac{\|Wx\|_D}{\|x\|_D} \\
&= \sup_{x\neq 0} \frac{\|DWx\|_2}{\|Dx\|_2} \\
&= \sup_{y\neq 0} \frac{\|DWD^{-1}y\|_2}{\|y\|_2} \\
&= \bar{\sigma}(DWD^{-1}) \qquad (5)
\end{aligned}$$

where $\bar{\sigma}(DWD^{-1})$ is the largest singular value of the matrix $DWD^{-1}$. Since $D$ is nonsingular, $\mathcal{N}(D) = \mathcal{N}(D^{-1}) = \{0\}$, so $y = 0$ if and only if $x = 0$. Therefore the constraint $y \neq 0$ is equivalent to $x \neq 0$. The last equality in (5) follows from the definition of the induced 2-norm of a matrix.

### C. Minimizing the Matrix Operator $D$-Norm

Since $D$ is arbitrary, it may be chosen such that $\|W\|_D = \bar{\sigma}(DWD^{-1})$ satisfies Lemma 2 for a given $\epsilon$. If $D$ is allowed to have full structure, then

$$\inf_{D\in\mathcal{D}} \bar{\sigma}(DWD^{-1}) = \rho(W) \qquad (6)$$

where infimum is used instead of minimum since $D$ (or $D^{-1}$), in many cases, will be approaching a singular matrix.

If $\mathcal{D}$ is a set of matrices that has some structure imposed upon it, say the set $\overline{\mathcal{D}}$, then $\|W\|_{D_\delta} = \bar{\sigma}(D_\delta W D_\delta^{-1})$ will not necessarily approach the spectral radius of $W$. Instead, the following relationship holds:

$$\rho(W) \leq \inf_{D_\delta\in\overline{\mathcal{D}}} \bar{\sigma}\left(D_\delta W D_\delta^{-1}\right) \leq \bar{\sigma}(W). \qquad (7)$$

In (7), the upper bound is obvious since $D_\delta = I$ is always an option. For general $W$, taking the infimum over all possible $D_\delta \in \overline{\mathcal{D}}$ will result in a measure that is strictly less than the $\overline{\sigma}(W)$ and strictly greater than $\rho(W)$. However, there are classes of matrices for which the lower bound of (7) is exact. This leads to the following theorem.

*Theorem 3:* Let $W \in \mathbb{F}^{n\times n}$ be in one of the following two classes:

1) normal matrices;
2) upper and lower triangular matrices.

Then, there exists a $D_\delta \in \overline{\mathcal{D}}$ such that $\|W\|_{D_\delta} = \rho(W) + \epsilon$ for all $\epsilon > 0$.

*Proof:*

Class 1) The first class is the easiest to prove since the singular values of a normal matrix are equal to the absolute values of its eigenvalues. Therefore, the maximum singular value and spectral radius are equal.

Class 2) In contrast, the gap between the spectral radius and maximum singular value of a triangular matrix may be arbitrarily large, but the operator $D_\delta$-norm can always be made arbitrarily close to the spectral radius. To see this, let $\delta > 0$, $W \in \mathbb{F}^{n\times n}$ be upper triangular, $D_\delta = \operatorname{diag}(1, \delta, \delta^2, \ldots, \delta^{n-1})$, and $D_\delta^{-1} = \operatorname{diag}(1, (1/\delta), (1/\delta^2), \ldots, (1/\delta^{n-1}))$. Then

$$D_\delta W D_\delta^{-1} = \begin{pmatrix} w_{1,1} & \frac{1}{\delta} w_{1,2} & \cdots & \frac{1}{\delta^{n-1}} w_{1,n} \\ 0 & w_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{\delta} w_{n-1,n} \\ 0 & \cdots & 0 & w_{n,n} \end{pmatrix}. \tag{8}$$

Using a limiting argument yields

$$\begin{aligned} \lim_{\delta\to\infty} \overline{\sigma}\left(D_\delta W D_\delta^{-1}\right) &= \overline{\sigma} \begin{pmatrix} w_{1,1} & 0 & \cdots & 0 \\ 0 & w_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & w_{n,n} \end{pmatrix} \\ &= \max_{1\le k\le n} |w_{k,k}| \\ &= \rho(W) \end{aligned} \tag{9}$$

which from Lemma 2 is the smallest that any operator norm may approach. ■

*Remark:* For a lower triangular matrix, the same argument would be used with the limit $\delta \to 0$.

*Remark:* In this derivation, $D^{-1}$ is approaching a (rank one) singular matrix, but a finite $D$ may be chosen that is arbitrarily close to the to the infimum as stated in Lemma 2.

*Remark:* Theorem 3 also holds for matrices that may be permuted to triangular matrices by swapping the matching rows and columns (e.g., swapping rows 3 and 5 followed by swapping columns 3 and 5). In the context of an ESN, this amounts to a relabeling of the recurrently connected neurons.

The results from this section will be used to obtain a tighter (and in some cases exact) bound for the echo state property.

## V. A New Sufficient Condition for the Echo State Property

In this section, a new sufficient condition is derived using the results from Section IV.

*Theorem 4:* Let an ESN have a fixed internal weight matrix $W \in \mathbb{R}$ and assume the squashing function $f(\cdot)$ satisfies the (element-wise) Lipshitz condition $|f(v) - f(z)| \le |v - z| \forall v, z \in \mathbb{R}$. If $\inf_{D_\delta \in \overline{\mathcal{D}}} \overline{\sigma}(D_\delta W D_\delta^{-1}) < 1$, then the network has the echo state property, i.e., $\lim_{k\to\infty} \|y_k\|_{D_\delta} = 0$ for all right infinite input sequences $u^{+\infty} \in U^{+\infty}$.

*Proof:*

$$\begin{aligned} \|y_{k+1}\|_{D_\delta} &= \|x_{k+1} - \tilde{x}_{k+1}\|_{D_\delta} \\ &= \|T(x_k, u_k) - T(\tilde{x}_k, u_k)\|_{D_\delta} \\ &= \left\| f(W^{\text{in}} u_k + W x_k) - f(W^{\text{in}} u_k + W \tilde{x}_k) \right\|_{D_\delta} \\ &\le \left\| (W^{\text{in}} u_k + W x_k) - (W^{\text{in}} u_k + W \tilde{x}_k) \right\|_{D_\delta} \qquad (10) \\ &= \|W x_k - W \tilde{x}_k\|_{D_\delta} \\ &= \|W(x_k - \tilde{x}_k)\|_{D_\delta} \\ &\le \|W\|_{D_\delta} \|y_k\|_{D_\delta} \\ &= \bar{\sigma}(D_\delta W D_\delta^{-1}) \|y_k\|_{D_\delta} \qquad (11) \end{aligned}$$

where $\inf_{D_\delta \in \overline{\mathcal{D}}} \overline{\sigma}(D_\delta W D_\delta^{-1}) < 1$ satisfies the contracting property. Note that this clearly implies the required convergence property, namely $\lim_{k\to\infty} \|y_k\|_{D_\delta} = 0$. ■

*Remark:* Equation (10) is the reason for using the set of diagonal scaling matrices $\overline{\mathcal{D}}$, which follows from removing the squashing function. If $D_\delta$ was allowed to have full structure (i.e., if $D \in \mathcal{D}$ was used), this inequality will not hold for all state vectors $x_k$ and $\tilde{x}_k$. However, the diagonally structured $D_\delta$ ensures that every element of the "squashed" version will be less than the "unsquashed" version and hence the $D_\delta$-norm will be less. Therefore, a sufficient condition for the existence of echo states is $\inf_{D_\delta \in \overline{\mathcal{D}}} \overline{\sigma}(D_\delta W D_\delta^{-1}) < 1$.

*Remark:* Since all finite dimensional norms are equivalent, $\lim_{k\to\infty} \|y_k\|_{D_\delta} = 0$ implies that $\lim_{k\to\infty} \|y_k\|_2 = 0$. Therefore, the original echo state property is satisfied by the new constraint.

*Corollary 5:* If $W$ is a normal matrix or a (permuted) triangular matrix, then $\rho(W) < 1$ is both a necessary and sufficient condition for the existence of echo states for all inputs $u_k$.

*Proof:* The proof of this follows from Theorem 3. ■

*Remark:* If $W \in \mathbb{F}$ is triangular, then $\rho(W) < 1$ is identical to $\max |\operatorname{diag}(W)| < 1$.

*Remark:* The stability requirement $\inf_{D_\delta \in \overline{\mathcal{D}}} \overline{\sigma}(D_\delta W D_\delta^{-1}) < 1$ is equivalent to a strong form of Lyapunov stability [7].

Since $D$ must be diagonal, this new bound is not tight in the sense that it is not necessarily equivalent to $\rho(W) < 1$ for any arbitrary $W$, but it is considerably less conservative than the bound $\bar{\sigma}(W) < 1$.

The reason for using the operator 2-norm (as the underlying norm) is that there exists commercial software for minimizing $\|DWD^{-1}\|_2$ when $D$ must be a structured matrix. In fact, this is a well-studied problem in the robust control community where $\|DWD^{-1}\|_2$ is an upper bound for the structured singular value of $W$ (usually denoted as $\mu(W)$). For more information on robust controls and the structured singular value, see [8]. Using (for example) MATLAB's *$\mu$-Analysis and Synthesis Toolbox* [9], the infimum of $\|W\|_{D_\delta}$ may be calculated using the command

$$[\mathtt{muUB}, \mathtt{muLB}] = \mathtt{mu(W)}.$$

Here

$$\mathtt{muUB} = \inf_{D_\delta \in \overline{\mathcal{D}}} \overline{\sigma}(D_\delta W D_\delta^{-1}).$$

Therefore, if $\mathtt{muUB} < 1$, the ESN has the echo state property. Note that if $\mathtt{muUB} \ge 1$, a new ESN may be defined with internal matrix $\tilde{W} = (0.99/\mathtt{muUB})W$, which will satisfy the new sufficient condition, and, hence, have the echo state property.

## VI. Triangular Versus Unstructured Weight Matrices

In order to demonstrate the usefulness of the new sufficient condition, an ESN was used to model a mass-spring-damper system. The details of this experiment may be found in [10]. The basic ESN architecture was two inputs, namely position and velocity, and one output, namely acceleration. The hidden layer contained 20 neurons (i.e., $W_{\text{in}} \in \mathbb{R}^{20\times 20}$), and the output layer was trained to estimate the acceleration based on the echo states generated from the previous position and velocity inputs. After the output layer was trained, a test set (which was different than the training set) of (position and velocity) inputs were used to assess the various ESNs. Performance for each ESN was assessed in terms of the mean-squared error (mse) between the predicted acceleration and the theoretical acceleration. The next two sections summarize the results from simulating ESNs with full and triangular hidden layer weight matrices.

TABLE I
RESULTS OF ESN SIMULATIONS FOR UNSTRUCTURED $W$

| | $\overline{\sigma}(W)$ | $\|W\|_{D_\delta}$ | $\rho(W)$ | mse |
|---|---|---|---|---|
| $\overline{\sigma}(W) < 1$ | 0.99 | 0.7390 | 0.5303 | 0.0136 |
| $\|W\|_{D_\delta} < 1$ | 1.3099 | 0.99 | 0.7056 | 0.0127 |
| $\rho(W) < 1$ | 1.9111 | 1.4913 | 0.99 | 0.0148 |

TABLE II
RESULTS OF ESN SIMULATIONS FOR STRUCTURED $W$

| | $\overline{\sigma}(W)$ | $\|W\|_{D_\delta}$ | $\rho(W)$ | mse |
|---|---|---|---|---|
| symmetric | 0.99 | 0.99 | 0.99 | 0.0143 |
| triangular | 1.6698 | 0.99 | 0.99 | 0.0155 |

### *A. ESN With Unstructured Hidden Weight Matrices*

For these experiments, the hidden layer weight matrix $W$ had a density of 20%, which meant that is had $0.2 \times 20^2 = 80$ nonzero entries. The ESNs were simulated (with $W$ scaled tight against the bound) using the old sufficient condition, new sufficient condition, and the necessary condition only. Recall that satisfying the sufficient conditions guarantees satisfying the necessary condition but not vice versa. In these experiments, $W$ was created from sampling 80 entries of a precalculated random matrix $W_{\mathrm{rand}} \in \mathbb{R}^{20\times 20}$, which was fixed for all experiments. The resulting $W$ matrices were then scaled to be tight against the bound being used. In each of these experiments, ten random samplings were used and their results averaged. The averaged results are summarized in Table I.

For this particular experiment, the ESNs that used the new sufficient condition worked the best; however, the performance of a given ESN depends strongly on the topology used. The topic of topology is beyond the scope of this letter; however, a more detailed explanation of the dependence of an ESN on topology may be found in [10]. While the bound $\rho(W) < 1$ is generally used in practice [4], this scaling is "dangerous" in the sense that it does not guarantee the echo state (global stability) property for all allowable inputs. The interest here stems from the fact that a larger scaling parameter may be used and still guarantee global stability of an ESN. In most cases tested, the general topology (usually nonoptimal) performed the best when the hidden layer weight matrix was scaled to be tight against the new sufficient condition bound.

### *B. ESNs With Structured Hidden Weight Matrices*

The same experiment (as aforementioned) was used to test the usefulness of both normal and triangular hidden layer weight matrices $W$. In these experiments, the normal matrix was a symmetric matrix with 80 nonzero entries, and the triangular matrix was an upper triangular matrix with 40 nonzero entries. Since the rank of $W$ appears to have an impact on the performance of an ESN [10], the upper triangular $W$ was forced to have full rank by assigning nonzero values along its main diagonal. The results of these experiments are presented in Table II.

These structured weight matrices produced similar results to the unstructured matrices that satisfied the necessary condition only. However, these matrices, when scaled so that $\rho(W) < 1$, are guaranteed to provide stable ESNs where the unstructured matrices are not. For a more detailed study of how to design and train ESNs, see [10] and [4].

## VII. CONCLUDING REMARKS

The echo state property is directly related to the asymptotic stability of the dynamical system formed in the hidden layer of an ESN. Even though ESNs appear to perform well in practice with only local stability (i.e., only the necessary condition is satisfied) [3], [4], [10]–[12], there is no guarantee that they will always perform well on a physical system. In order to make a more concrete statement about the performance of an ESN for all inputs (i.e., to guarantee that echo states exist for all inputs), a more rigorous global asymptotic stability requirement is needed. While a (rather restrictive) requirement existed for guaranteeing global asymptotic stability, this letter presented a new bound that is less restrictive than the original bound and for some topologies provides exact analysis. Note that this letter does not address the design methodology of an ESN (for that, see [10]). Instead, it provides a test for global asymptotic stability (and hence the existence of echo states for all inputs) which utilizes linear algebra optimization problems with readily available commercial software.

## ACKNOWLEDGMENT

The authors would like to thank K. Bush for his helpful discussions and for providing the code used to test the ESNs.

## REFERENCES

[1] J. J. Steil, *Input-Output Stability of Recurrent Neural Networks*. Göttingen, Germany: Cuvillier Verlag, 1999.

[2] A. F. Atiya and A. G. Parlos, "New results on recurrent network training: unifying the algorithms and accelerating convergence," *IEEE Neural Netw.*, vol. 11, no. 3, p. 697, May 2000.

[3] P.-G. Plöger, A. Arghir, T. Günther, and R. Hosseiny, "Echo state networks for mobile robot modeling and control," in *Proc. RoboCup*, 2003, pp. 157–168.

[4] H. Jaeger, "The echo state approach to analyzing and training recurrent neural networks," German National Research Center for Information Technology, Tech. Rep. 148, 2001.

[5] R. Horn and C. R. Johnson, *Matrix Analysis*: Cambridge University Press, 1985.

[6] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra*. Philadelphia, PA: SIAM, 1997.

[7] A. Packard and J. C. Doyle, "The complex structured singular value," *Automatica*, vol. 29, no. 1, pp. 71–109, 1993.

[8] S. Skogestad and I. Postlethwaite, *Multivariable Feedback Control*. West Sussex, U.K.: Wiley, 1996.

[9] G. J. Balas, J. C. Doyle, K. Glover, A. Packard, and R. Smith, *$\mu$-Analysis and Synthesis TOOLBOX User's Guide*. Natick, MA: The Math Works, Inc., 1996.

[10] K. Bush and B. Tsendjav, "Improving the richness of echo state features using next ascent local search," in *Proc. Artificial Neural Networks In Engineering Conf.*, St. Louis, MO, Nov. 2005, pp. 227–232.

[11] K. Bush and C. Anderson, "Modeling reward functions for incomplete state representations via echo state networks," in *Proc. Int. Joint Conf. Neural Networks*, Montreal, Canada, Jul. 2005, pp. 2995–3000.

[12] K. Ishii, T. Van Der Zant, V. Becanovic, and P. G. Plöger, "Identification of motion with echo state network," in *Oceans'04 (MTS/IEEE)*, Kobe, Japan, Nov. 2004, pp. 1205–1210.

[13] J. W. Demmel, *Applied Numerical Linear Algebra*. Philadelphia, PA: SIAM, 1997.