Proof about *Weakly Universal State Contracting Networks*

**Norbert Michael Mayer**

Dept. of Electrical Engineering & AIM-HI, Nat’l Chung Cheng University, Taiwan

**Keywords:** Reservoir computing, Universal State Contracting Networks, power law

**Abstract**

Recurrent networks that have transfer functions that fulfill the Lipschitz continuity with \( L = 1 \), may be echo state networks if certain limitations on the recurrent connectivity are applied. Initially it has been shown that it is sufficient if the largest singular value of the recurrent connectivity \( \Lambda \) is smaller than 1. Here it is investigated under which it still can be shown that the network is an echo state network if \( \Lambda = 1 \). It turns out that for certain transfer functions indeed the network is still an echo state network. This work is closely related to experimental results outlined in Mayer [2014].

1 Introduction

Jäger [Jäger, 2001, 2010, 2003] has shown under which circumstances recurrent connectivity results (1) in internal states independent from initial states for sufficient long input histories, thus the state of the network finally is completely determined by the input. On the other hand Jäger also determined a set of conditions (2) in which the network behaves chaotic and thus is sensitive to the initial conditions in the internal layer. These conditions are called Echo State (ES) conditions.

However, Jägers mathematical leaves a gap of undetermined configurations for which it is still open if the network belongs to either or the two groups. As outlined in sect. the initial work of Jäger proves the validity of the ES conditions in terms of topology for an open set of conditions. The subject of this work is to check under which conditions the border of the open set (aka. the critical point) belongs to the area where the ES condition is fulfilled, or not.

As it turns out below the behavior of the border depends on the choice of the transfer function; only for certain transfer functions ES conditions are fulfilled. In the following section the definitions for echo state networks are reiterated. Then two different transfer functions are presented that are applied in the proof. Sect. describes the result of the initial proof by Jäger. The following 3 sections outline the extended proof. Finally, the results are discussed.
As a convention the transfer function in the scope of this paper. ESNs are composed of rate-based neurons with real valued transfer functions; the update rule is:

\[
\begin{align*}
x_{\text{lin},t} &= Wx_{t-1} + w^{\text{in}}u_t \\
x_t &= \theta(x_{\text{lin},t}) \\
o_t &= w^{\text{out}}x_t
\end{align*}
\]

where the vectors \(u_t, x_t\), and \(o_t\) are the input, the neurons of the hidden layer, and the neurons of the output layer, respectively, and \(w^{\text{in}}\), \(W\), and \(w^{\text{out}}\) are the matrices of the respective synaptic weight factors.

As a convention the transfer function \(\theta(\cdot)\) is continuous, differentiable and monotonically increasing with the limit \(1 \geq \theta(\cdot) \geq 0\), which is compatible with the requirement that \(\theta(\cdot)\) fulfills the Lipschitz continuity with \(L = 1\).

Jäger’s approach uses random matrices for \(W\) and \(w^{\text{in}}\), learning is restricted to the output layer \(w^{\text{out}}\) (see fig. 1). The learning (i.e. training \(o_t\)) can be performed by linear regression. Since the learning process with regard to some output itself is not of interest in the scope of this paper, this part of the approach is not outlined here.

3 Transfer functions

The following proof considers two different transfer functions

\[
\theta(x) = \tanh(x),
\]
and
\[ \theta(x) = 0.5x - 0.25\sin(2x) \]  
(5)

It has to be noted that in both cases the maximal derivative \( \theta'(x) \) is 1.

4  Echo-state condition

A necessary condition for the performance of an ESN network is that the echo state condition is fulfilled. Consider a time-discrete recursive function \( x_{t+1} = F(x_t, u_t) \), where the \( x_t \)'s are interpreted as internal states and the \( u_t \)'s form some external input sequence, i.e. the stimulus. The definition of the echo-state condition is the following. Assume an infinite stimulus sequence: \( \bar{u}^\infty = u_0, u_1, \ldots \) and two random initial internal states of the system \( x_0 \) and \( y_0 \). To both initial states \( x_0 \) and \( y_0 \), the sequences \( \bar{x}^\infty = x_0, x_1, \ldots \) and \( \bar{y}^\infty = y_0, y_1, \ldots \) can be assigned;

\[
\begin{align*}
x_{t+1} &= F(x_t, u_t) \\
y_{t+1} &= F(y_t, u_t).
\end{align*}
\]  
(6)

Then the system \( F(\cdot) \) is called \textit{universally state contracting} fulfills if \textbf{independent} from the set \( u_t \) and for \textbf{any} \((x_0,y_0)\) and all real values \( \epsilon > 0 \), there exists an iteration \( \tau(\epsilon) \) for which

\[ d(x_t, y_t) \leq \epsilon \]  
(7)

for all \( t \geq \tau \).

Jäger showed ([Jäger, 2001, 2010] pp. 43) that the echo state condition is fulfilled if and only if the network is state contracting. The ESN is designed to be universally state contracting and thus to fulfill the echo-state condition.

5  Echo state condition limit with weak contraction

Jäger’s sufficient echo state condition (see [Jäger, 2001, 2010], App. C, p. 41) has strictly been proven only for non-critical systems (largest singular value \( \Lambda < 1 \)) and with \( \tanh(\cdot) \) as a transfer function. The original proof is based on the fact that \( \tanh \) in combination with \( \Lambda < 1 \) is a contraction. In that case Jäger shows an exponential convergence. Obviously, that proof is not valid for the case that is presented in this present work. However, the initial theorem can be extended under the following circumstances.

\textbf{Theorem:} If hyperbolic tangent or the function of eq. 5 are used as transfer functions \textbf{and} all singular values are larger than zero, the echo state condition (see eq. 7) is fulfilled even if \( \Lambda = 1 \).

\textbf{Summary of the proof:} Basically, as an important precondition the proof requires that both transfer functions fulfill

\[ d(\theta(y_t), \theta(x_t)) \leq d(y_t, x_t) \phi(d(y_t, x_t)), \]  
(8)

where \( \phi(z) \) is defined as

\[
\begin{align*}
z < \gamma & : \phi(z) = (1 - \eta z^\kappa) \\
z \geq \gamma & : \phi(z) = (1 - \eta \gamma^\kappa),
\end{align*}
\]  
(9)
while \( 1 > \gamma, \eta > 0, \kappa \geq 1 \) are parameters that have to be determined for each transfer function separately, and also for each metric norm \( d(.,.) = ||.||_p \).\footnote{Since P-norms are equivalent, once we found a valid set of parameters \( \eta_{p1} \) and \( \gamma_{p1} \) for one P-norm \( p_1 \) we can get a new set for another norm \( p_2 \) by simply rescaling \( \eta_{p2} = a \eta_{p1}, \gamma_{p2} = a \gamma_{p1} \), where \( a > 0 \) is a finite scaling factor.}

In sect. 6 it is shown that indeed both transfer functions fulfill that requirement. It then remains to prove that in the slowest case, we have a convergence in a process with 2 stages. The first stage, if \( d(y_t, x_t) > \gamma \) there is a convergence that is faster or equal to an exponential decay. The second stage is a convergence process that is faster or equal to a power law decay.

**Proof:** Note with regard to the test function \( \phi \):

\[
\phi \leq 1 \\
0 \leq z \leq Z : \phi(z) \geq \phi(Z)
\]

In analogy to Jäger one can check now the contraction between the time step \( t \) and \( t+1 \):

\[
d(y_{t+1}, x_{t+1}) = d(\theta(y_{lin,t+1}), \theta(x_{lin,t+1})) \\
\leq d(y_{lin,t+1}, x_{lin,t+1}) \times \phi(d(y_{lin,t+1} x_{lin,t+1}))
\]

One can rewrite

\[
d(y_{lin,t+1}, x_{lin,t+1}) = ||W y_{lin,t} + I - W x_{lin,t} + I|| = ||W(y_{lin,t} - x_{lin,t})||,
\]

where \( I = w^{in} u_t \). Thus one gets

\[
d(y_{lin,t+1}, x_{lin,t+1}) \times \phi(d(y_{lin,t+1} x_{lin,t+1})) = ||W(y_{lin,t} - x_{lin,t})|| \times \phi(||W(y_{lin,t} - x_{lin,t})||) \\
\leq \Lambda d(y_{lin,t}, x_{lin,t}) \times (\phi(\lambda d(y_{lin,t}, x_{lin,t}))
\]

where \( \Lambda \) and \( \lambda \) are the largest and smallest singular values of \( W \), respectively.

Merging eq. 10 and eq. 12 results in the inequality

\[
d(y_{t+1}, x_{t+1}) \leq \Lambda d(y_{lin,t}, x_{lin,t}) \times (\phi(\lambda d(y_{lin,t}, x_{lin,t}))).
\]

First, assuming \( \Lambda < 1 \) we get an exponential decay

\[
d(y_n, x_n) \leq \Lambda^t d(y_0, x_0).
\]

This case is handled by Jäger’s initial proof. With regard to an upper limit of the contraction speed (cf. eq. 7) one can find

\[
\tau(\epsilon) = \frac{\log(\epsilon) - \log(d(y_0, x_0))}{\log(\Lambda)}.
\]

If the largest singular value \( \Lambda > 1 \), then the largest absolute eigenvalue is also larger than 1 due to the spectral theorem. Thus, the echo state condition is not fulfilled due to Jaeger’s first and second condition (C1 & C2).

What remains is to check the critical case \( \Lambda = 1 \). Here again one considers 2 cases:
If \( d(y_{lin,t}, x_{lin,t}) > \gamma \) we can write the update inequality of eq. \ref{eq:13} as:

\[
d(y_{t+1}, x_{t+1}) \leq (1 - \lambda \eta \gamma^\kappa)d(y_t, x_t).
\]

(16)

Thus, for all \( \epsilon \geq \gamma \), the slowest decay process can be covered by

\[
\tau(\epsilon) = \frac{\log(\epsilon) - \log(d(y_0, x_0))}{\log((1 - \eta \gamma^\kappa))}.
\]

(17)

If \( \epsilon < \gamma \), then eq. \ref{eq:13} becomes:

\[
d(y_{t+1}, x_{t+1}) \leq d(y_t, x_t)(1 - \lambda \eta d(y_t, x_t)^\kappa).
\]

(18)

One can now consider the sequence \( a_t \):

\[
a_{t+1} = a_t(1 - \lambda \eta a_t^\kappa).
\]

(19)

This series fulfills \( a_t \geq d(y_t, x_t) \) for all \( t \) if we set \( a_0 = d(y_0, x_0) \leq \gamma \). Unfortunately, it seems hard to find an explicit form for the sequence of eq. \ref{eq:19}. Fortunately, another sequence,

\[
a^*(t) = \left[ \frac{\lambda \eta t + a_0^{-\kappa}}{\kappa} \right]^{-\frac{1}{\kappa}},
\]

(20)

covers again the sequence \( a_t \) (cf. sect. \ref{sect:7}). From the convergence of the sequence \( a_n \), one can conclude the convergence of \( d(x_t, y_t) \). From the solution one can calculate the upper limit \( \tau(\epsilon) \) if \( \epsilon \leq \gamma \):

\[
\tau(\epsilon) = \frac{\kappa}{\lambda \eta}(\epsilon^{-\kappa} - a_0^{-\kappa}).
\]

(21)

### 6 Weak contraction with the present transfer function

In this section a test function according to eq. \ref{eq:9} i.e. parameters \( \kappa, \gamma \) and \( \eta \) is verified for the function of eq. \ref{eq:5} and hyperbolic tangent. Within this section we test the following
for both transfer functions and Manhatten norm ($||.||_1$). In order to derive these values one can start by considering linear responses $||y_{lin,t} - x_{lin,t}||$ and the final value $||y_t - x_t||$ within one single neuron as

$$
y_t = \theta(y_{lin,t})
$$

$$
x_t = \theta(x_{lin,t}).
$$

(23)

We can define

$$
y_{lin,t} = z + \zeta
$$

$$
x_{lin,t} = \zeta.
$$

(24)

Setting in these definitions in eq. 8 and for a single neuron and for any $\zeta$ we get

$$
z \phi(z) \geq \theta(z + \zeta) - \theta(\zeta).
$$

(25)

Thus, it suffices to consider

$$
\phi(z) \geq \max_{\zeta} \frac{\theta(z + \zeta) - \theta(\zeta)}{z}.
$$

(26)

The $\max_{\zeta}$ can be found by basic analysis. Extremal points can be found at:

$$
\frac{\partial}{\partial \zeta} \left[ \frac{\theta(z + \zeta) - \theta(\zeta)}{z} \right] = \frac{\dot{\theta}(z + \zeta) - \dot{\theta}(\zeta)}{z} = 0.
$$

(27)

Since for both suggested transfer functions $\dot{\theta}$ is an even function one gets $\zeta = -z/2$ as the extremal point. Fundamental analysis shows that this point in both cases is a maximum. Thus eq. 26 becomes

$$
\phi(z) \geq \frac{2\theta(z/2)}{z}.
$$

(28)

The case of $z \leq \gamma$ can be analyzed by Taylor series expansion.

The Taylor series expansion $T^4(f, x_0, x)$ of a function $f$ with a fourth order upper limit estimate is:

$$
f(x) \leq \sum_{0 \leq i \leq 3} T^i(f, x_0, x) + \max_{\xi \in [x_0,x]} \left| \frac{f^{(4)}(\xi)}{24} (x - x_0)^4 \right|,
$$

(29)

where one can call the right hand addend

$$
T^4(f, x_0, x) = \max_{\xi \in [x_0,x]} \left| \frac{f^{(4)}(\xi)}{24} (x - x_0)^4 \right|,
$$

(30)
which is the estimate of the remainder. The other components for \( i < 4 \) are:

\[
T_i^4(f, x_0, x) = \frac{f^{(i)}(x_0)}{i!} (x - x_0)^i.
\]  

(31)

Now, one can get an upper estimate for the right hand side of eq. 28:

\[
2 \frac{T^4(\theta, 0, z/2)}{z} \geq \frac{2\theta(z/2)}{z}
\]  

(32)

In the case of \( \theta = \tanh \) one gets:

\[
T^4(\tanh, 0, z/2) = \frac{z}{2} - \frac{z^3}{48} + \max_{\xi \in [0, z]} \left| \tanh^{(4)}(\xi) \frac{z^4}{24} \right| \leq \frac{1}{96} z^3.
\]  

(33)

For the record one can write the 4th derivative of tangent hyperbolic

\[
\tanh^{(4)}(\xi) = 24\tanh^5(\xi) - 40\tanh^3(\xi) + 16\tanh(\xi).
\]  

(34)

Since for positive \( z \) one can estimate

\[
24\tanh^5(\xi) - 40\tanh^3(\xi) \leq 24\tanh^5(\xi) - 40\tanh^5(\xi) \leq 0,
\]  

(35)

one gets

\[|\tanh^{(4)}(\xi)| \leq 16\tanh(\xi).\]  

(36)

Furthermore, since the range of the approximation is limited to \( 0 \leq z \leq \gamma \), one uses \(|\tanh(\xi)| \leq \gamma = 0.5\) and can estimate

\[
\max_{\xi \in [0, z]} \left| \tanh^{(4)}(\xi) \frac{z^4}{24} \right| \leq \frac{1}{48} z^4 \leq \frac{1}{96} z^3.
\]  

(37)

One gets for left side of eq. 33

\[
T^4(\tanh, 0, z/2) \leq \frac{z}{2} - \frac{z^3}{96}.
\]  

(38)

So one can estimate for small \( z \) the value of the right side of eq. 28

\[
\frac{2\theta(z/2)}{z} \leq 2T^4(\tanh, 0, z/2) \leq 1 - \frac{1}{48} z^2.
\]  

(39)

Thus, one can set the parameters \( \eta = 1/48 \) and \( \kappa = 2 \) of \( \phi(z) \) as in eq. 22.

For \( z > \gamma \) is suffices to check if

\[
\frac{2\tanh(z/2)}{z} \leq 1 - \eta \gamma^\kappa
\]  

(40)

First the inequality is fulfilled for \( z = \gamma \). Since the left side of the equation above is strict monotonously falling for positive \( z \), it is also fulfilled for all values \( z > \gamma \).

Analogous considerations lead to the same parameters to cover up the transfer function from eq. 5. Fig. 2 illustrates how the transfer functions and the \( \phi(z) \) relate to each other.
7 Sequence $a_t^*$ covers up sequence $a_t$

In the following, identical definitions to sect. 5 are used. One can start from the statement
\[ -\frac{\lambda \eta}{\kappa} \geq -\lambda \eta. \]  
(41)

Since $\lambda \geq 0$ and $\eta \geq 0$ and $\kappa \geq 1$, it can easily be seen that the inequality is fulfilled for any combination of $\lambda$, $\eta$ and $\kappa$. One can now extend the numerator and denominator of the right side by $(\frac{\lambda \eta}{\kappa} t + C)$ and add also $(\frac{\lambda \eta}{\kappa} t + C)$ to both sides of the inequality. Here and in the following $C$ is defined as
\[ C = a_0^{-\kappa} \geq 1. \]  
(42)

One obtains
\[ \frac{\lambda \eta}{\kappa} t + C - \frac{\lambda \eta}{\kappa} \geq (\frac{\lambda \eta}{\kappa} t + C) - (\frac{\lambda \eta}{\kappa} t + C) \frac{\lambda \eta}{\lambda \eta / \kappa t + C}. \]  
(43)

Rearranging results in
\[ \frac{\lambda \eta}{\kappa} t + C - \frac{\lambda \eta}{\kappa} \geq (\frac{\lambda \eta}{\kappa} t + C) \times (1 - \frac{\lambda \eta}{\lambda \eta / \kappa t + C}). \]  
(44)

One can add $\frac{\lambda \eta}{\kappa}$ at both sides and has the following
\[ \frac{\lambda \eta}{\kappa} t + C \geq \frac{\lambda \eta}{\kappa} + (\frac{\lambda \eta}{\kappa} t + C) \times (1 - \frac{\lambda \eta}{\lambda \eta / \kappa t + C}). \]  
(45)

Now, one can use the fact that
\[ 1 \geq 1 - \frac{\lambda \eta}{\lambda \eta / \kappa t + C} \geq \left(1 - \frac{\lambda \eta}{\lambda \eta / \kappa t + C}\right)^\kappa \]  
(46)

and
\[ \frac{\lambda \eta}{\kappa} \geq \frac{\lambda \eta}{\kappa} \times \left(1 - \frac{\lambda \eta}{\lambda \eta / \kappa t + C}\right) \geq \frac{\lambda \eta}{\kappa} \times \left(1 - \frac{\lambda \eta}{\lambda \eta / \kappa t + C}\right)^\kappa. \]  
(47)

and thus rewrite inequality eq.45 as
\[ \frac{\lambda \eta}{\kappa} t + C \geq \left(\frac{\lambda \eta}{\kappa} t + C + \frac{\lambda \eta}{\kappa}\right) \times \left(1 - \frac{\lambda \eta}{\lambda \eta / \kappa t + C}\right)^\kappa \]  
(48)

and finally arrive at
\[ \left(\frac{\lambda \eta}{\kappa} t + C\right) \geq \left(\frac{\lambda \eta}{\kappa} t + C + \frac{\lambda \eta}{\kappa}\right) \times \left(1 - \frac{\lambda \eta}{\lambda \eta / \kappa t + C}\right)^\kappa; \]  
(49)
one can multiply both sides with \((\frac{\lambda \eta}{\kappa} t + C + \frac{\lambda \eta}{\kappa})^{-1} \times (\frac{\lambda \eta}{\kappa} t + C)^{-1}\). So,

\[
\left(\frac{\lambda \eta}{\kappa} t + C + \frac{\lambda \eta}{\kappa}\right)^{-1} \geq \left(\frac{\lambda \eta}{\kappa} t + C\right)^{-1} \times \left(1 - \frac{\lambda \eta}{\kappa} t + C\right)^{\kappa}.
\]  

(50)

Taking the \(k\)th square root one gets

\[
\left(\frac{\lambda \eta}{\kappa} t + C + \frac{\lambda \eta}{\kappa}\right)^{-1/k} \geq \left(\frac{\lambda \eta}{\kappa} t + C\right)^{-1/k} \times \left(1 - \frac{\lambda \eta}{\kappa} t + C\right).
\]  

(51)

One can re-arrange the inequality to:

\[
\left(\frac{\lambda \eta}{\kappa} (t + 1) + C\right)^{-1/k} \geq \left(\frac{\lambda \eta}{\kappa} t + C\right)^{-1/k} \times \left(1 - \frac{\lambda \eta}{\kappa} t + C\right)^{\kappa}.
\]  

(52)

One can add the definitions of \(a^*_t\) and \(a^*_{t+1}\), i.e.

\[
a^*(t) = \left[\frac{\lambda \eta}{\kappa} t + a^{-\kappa}_0\right]^{-1/k}
\]  

(53)

one has finally

\[
a^*_{t+1} \geq a^*_t \times (1 - \lambda \eta (a^*_t)^{\kappa}).
\]  

(54)

proven as a true statement. Thus,

\[
a^*_t \geq a_t \geq d(y_t, x_t)
\]  

(55)

for all \(t > 0\) if

\[
a^*_0 = a_0 = d(y_0, x_0).
\]  

(56)

Thus, the positive definite sequence \(a^*_t\) covers \(d(y_t)\).

8 Discussion

The proof shows a network as outlined in sect. 2 is an ESN even if the largest eigenvalue of the recurrent connectivity matrix is equal to 1 and if the transfer function is either eq. 4 or 5. The proof is also extensible to other transfer functions. On the other hand it is obvious that some transfer functions result in networks that are not ESNs. For example a linear transfer function \((\theta(x) = x)\) is not universal state contracting. Even if the network is state contracting it is not necessarily exponential state contracting, but rather a power law in the slowest case. An example of power law forgetting is described in Mayer [2014].
9 Acknowledgements

NSC (Nation Science Council) of Taiwan provided the budget for our projects. Also thanks go to AIM-HI for various ways of support. N. M. M. thanks Chris Shane for his cross reading.

References

Herbert Jäger. The echo state approach to analysing and training recurrent neural networks with an erratum note. In *GMD Report 148, GMD German National Research Institute for Computer Science*, 2001, 2010. http://www.gmd.de/People/Herbert.Jaeger/Publications.html.

Herbert Jäger. Adaptive nonlinear system identification with echo state networks. In *Proc. of NIPS 2002*, 2003. AA14.

N. M. Mayer. Adaptive critical reservoirs with algebraic forgetting of unexpected input events. arXiv preprint arXiv:1404.6334), 2014.