Optimizing individually the variance of the neural activity regulates the echo-state spectral radius

Fabian Schubert and Claudius Gros
Institute for Theoretical Physics, Goethe University, Frankfurt a.M., Germany.

Keywords: variance optimization, echo state network, spectral radius, biological plausibility, self-organization, universality

Abstract

Echo state networks are based on recurrent nets that provide large selections of non-linear filters of the input signal to a perceptron layer. For the echo state network to function optimally as a prediction machine, the spectral radius $R = |\Lambda_{max}|$ of the recurrent synaptic weight matrix needs to be regulated. This implies that the modulus of the largest eigenvalue Λ_{max} of the synaptic weight matrix should be unity, or slightly larger, a condition demanding ostensibly a non-local operation. Here we show that R can be regulated locally if neurons optimize the variance of their own activity. Variance optimization is performed relative to the variance of the input signal by adapting the slope of the transfer function.

The proposed local adaption rule can be carried out online, which implies that it is biologically plausible and that network can react autonomously to changes of the input statistics during normal operations. The effectiveness of the algorithm is supported by numerical simulations and an analytic approximation. The respective optimality condition is universal in the sense that it is independent of the network topology, the link probability and the distribution of synaptic weights.

1 Introduction

Recurrent network activity may process time-dependent input signals in at least two distinct ways. In the first scenario the network activity is present in the absence of a driving input, playing at the same time a key role in processing input signals. The autonomous neural activity may be in this case either structured (Gros, 2009; Berkes et al., 2011; Mitra and Raichle, 2016), or self-organized critical (Petermann et al., 2009; Arviv et al., 2015; Cocchi et al., 2017). Within the second scenario the network becomes active only when stimulated by sequences of input signals. Once externally induced, the recurrent activity provides a reservoir of non-linear transformations (Lukoševičius

and Jaeger, 2009), both in time and in space, that may be utilized by a second-state linear learning unit. One speaks of an echo state network when the constituting neurons are rate-encoding (Jaeger, 2005), and of a liquid-state machine for the case of spiking neurons (Maass et al., 2002).

Training the linear output units transform an echo state network to an highly effective prediction machine (Ozturk et al., 2007; Lukoševičius and Jaeger, 2009), with the performance needing a fine tuning of the network properties, in particular the spectral radius $R = |\Lambda_{max}]$. The activity decays in the absence of external inputs when the largest eigenvalue Λ_{max} of the synaptic weight matrix is smaller than unity, a precondition for the network to be able to encode input signals transiently (Yildiz et al., 2012; Caluwaerts et al., 2013). Information is however lost fast when the spectral radius is too fast, which is detrimental. A spectral radius of about unit may is hence best (Jaeger, 2002b), in the sense that it provides a maximal memory capacity (Boedecker et al., 2012; Farkaš et al., 2016).

A network is said to have the echo-state property when a short-term echo, but not a long-term reverberating activity, is present in response to a given input (Jaeger, 2002b). Echo state networks work however not in alternating on and off modes, but under the influence of a continuous stream of input signals, viz as continuously driven systems. The echo-state property has hence to be generalized (Manjunath and Jaeger, 2013) in terms of pullback attractors (Kloeden, 2000; Caraballo et al., 2006), which implies that two nearby internal states converge under the influence of the same input stream. Optimal performance in terms of a maximal memory capacity is achieved at the edge of chaos, viz at the point when the pullback attractor disappears. Here we show that optimality is attained under the influence of a simple and biological plausible self-organizing principle.

Besides their applications as efficient machine learning algorithms, echo state networks are potentially relevant also for information processing in the brain (Nikolić et al., 2009; Enel et al., 2016). It is hence important to examine whether there exist local and biological plausible principles allowing to tune the properties of the reservoir to the edge of chaos (Livi et al., 2018), in particular when a continuous stream of inputs is present. This rule needs to be independent of the network topology, which is not a locally accessible information, and of the distribution of synaptic weights.

Avalanches in Self-Organized Critical Neural Networks: A Minimal Model for the Neural SOC Universality Class. Matthias Rybarsch, Stefan Bornholdt

Schuecker et al. (2018) Optimal sequence memory in driven random networks. *At the transition point, the variance of the recurrent input to a single unit equals the variance of its own activity.*

(Wainrib and Galtier, 2016) A local Echo State Property through the largest Lyapunov exponent.

(Gallicchio and Micheli, 2017) Echo state property of deep reservoir computing networks.

(Livi et al., 2018) Determination of the edge of criticality in echo state networks through Fisher information maximization.

(Boedecker et al., 2009) Initialization and self-organized optimization of recurrent neural network connectivity.

(Gros, 1990) Criterion for a good variational wave function (variance optimization).

(Cannon and Miller, 2017) Stable Control of Firing Rate Mean and Variance by Dual Homeostatic Mechanisms.

2 Regulating the moments of the neural activity

Rate-encoding neurons come with a set of internal parameters, in particular the threshold b and the gain a. In this study we take $g(z) = \tanh(z)$ as the transfer function, where the argument is $z = a(x - b)$, with x being equivalent to the membrane potential. Neurons adapting on their own dispose of individual gains a_i and threshold b_i. The time evolution of a network characterized by a recurrent weight matrix w_{ij} is then

$$y_i(t+1) = g(a_i(x_i(t) - b_i)), \qquad x_i(t) = \sum_{j=1}^{N} w_{ij} y_j(t) + E_i(t) \tag{1}$$

where N is the number of reservoir neurons. Time is discrete and indexed by $t = 0, 1, \ldots$. For the external input $E_i(t)$ we use normal distributed white noise, with zero mean and standard deviation σ_{ext}.

In (1) we have refrained from introducing a separate set of input units and a matrix linking the inputs to the reservoir neurons, lumping the statistical effect of the external driving into a scalar quantity, $E_i(t)$. For the internal connections w_{ij} we selected a connection probability $p = 0.1$. The magnitude of the non-zero elements of w_{ij} are drawn then from a normal distribution with zero mean and standard deviation $\sigma_{\text{w}}/\sqrt{pN}$, a scaling that ensures that the variance of the two contributions a neuron receives, external and recurrent, are of the same order of magnitude (Sompolinsky et al., 1988). Typically we use $\sigma_{\text{w}} = 1.0$.

The value of internal neural parameters needs to be regulated on the basis of homeostatic principles (Marder and Goaillard, 2006), a mechanism also termed intrinsic adaption (Triesch, 2005). Here we propose that the two basic parameters, the threshold b_i and the gain a_i, have well defined and distinct roles in regulating the activity, with $b_i = b_i(t)$ and $a_i = a_i(t)$ being respectively responsible for setting the mean and the variance of the neural firing rate. The adaption rules $a_i(t+1) = a_i(t) + \Delta a_i(t)$ and $b_i(t+1) = b_i(t) + \Delta b_i(t)$ implementing this principle are

$$\Delta a_i = \epsilon_{\text{a}} \left(\sigma_{\text{t}}^2 - (y_i(t) - \bar{y}_i)^2 \right), \qquad \Delta b_i = \epsilon_{\text{b}} \left(y_i(t) - y_{\text{t}} \right) \tag{2}$$

where we have denoted the target values for the average and for the variance of the neural activity $y_i(t)$ respective by y_{t} and σ_{t}^2. An average over extended time scales are implicitly performed when the update rates ϵ_{a} and ϵ_{b} are small. The trailing average $\bar{y}_i$ showing up in (2) is defined here via

$$\bar{y}_i(t+1) = \bar{y}_i(t) + \bar{\epsilon} \left(y_i(t) - \bar{y}_i(t) \right), \tag{3}$$

where $\bar{\epsilon} \ll 1$.

2.1 Spectral radius renormalization

Echo state networks take their name from the absence of a long-lasting echo, namely that the transient stimulation of the network by an external stimulus evokes an exponentially decaying response. For an isolated system this condition, the echo state condition, implies that the fixpoint corresponding to an inactive network should be stable, namely that the spectral radius $R = |\Lambda_{max}|$ of the rescaled weight matrix should be smaller than unity. Note that this condition is local in time (Wernecke et al., 2019). With the gain a_i multiplying the afferent synaptic weights,

$$z_i = a_i(x_i - b_i) = a_i \left(\sum_{j=1}^{N} w_{ij} y_j(t) + E_i(t) \right) - a_i b_i \,, \tag{4}$$

adapting the gain is equivalent is equivalent to adapting the spectral radius. The key point is, that this can be done using locally available information, the variance of the neural activity and of the distribution of the external input $E_i(t)$. Technically, we evaluate

$$R(\widehat{W}_{\rm a}), \qquad (\widehat{W}_{\rm a})_{ij} = a_i W_{ij}, \tag{5}$$

where $\widehat{W}_a$ it the locally rescaled weight matrix. This procedure reduces to the standard matrix rescaling when the differences between the individual gains vanish, as it will be the case for large network sizes $N \to \infty$.

2.2 Variance optimization vs. entropy maximization

The adaption rules (2) are an abstraction of the principle of polyhomeostatic optimization (Markovic and Gros, 2010), which deals with the optimization of entire distribution functions. However, instead of maximizing the entropy (Triesch, 2005), as for the case of intrinsic adaption, or minimizing the Fisher information (Echeveste and Gros, 2014), when deriving self-limiting Hebbian learning rules, the two leading moments of the firing rate distribution $\rho(y)$ are regulated. In order to clarify this relation we consider the case that one wants to maximize the entropy of the neural activity under the constraint of a given mean μ and variance σ^2. The maximal entropy distributions is in this case a Gaussian (Gros, 2015),

$$N_{\mu,\sigma}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(y-\mu)^2/(2\sigma^2)} \,. \tag{6}$$

Table 1: The parameters used, where N is the network size, p is the connection probability of the recurrent weight matrix and $\sigma_{\rm w}$ the scale of the standard deviation $\sigma_{\rm w}/\sqrt{pN}$, of the w_{ij}, compare (2). Further listed are $\epsilon_{\rm a}$, $\epsilon_{\rm b}$ $\bar{\epsilon}$, the update rates for the gain, the threshold and the trailing average, and the target neural activity, $y_{\rm t}$.

N	p	$\sigma_{\rm w}$	$\epsilon_{\rm b}$	$\epsilon_{\rm a}$	$\bar{\epsilon}$	$y_{\rm t}$
1000	0.1	1.0	$2 \cdot 10^{-4}$	10^{-3}	10^{-4}	0

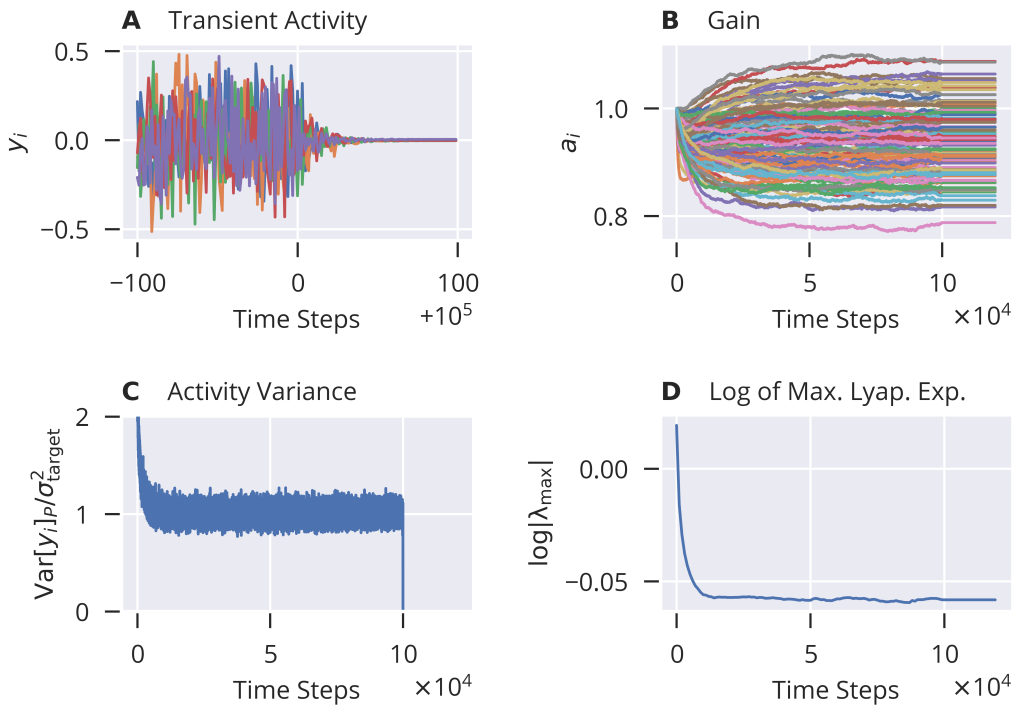

Figure 1: **Time series.** For $N = 1000$ neurons, $\sigma_{\mathrm{t}} = 0.2$, $\sigma_{\mathrm{ext}} = 0.1$ and the parameters given in Table 1, the evolution of selected quantities. At $t^E_{\mathrm{off}} = 10^5$ the external input is turned off. **A:** The activity of a random selection of $N/10$ neurons, with the activity dying out once the external driving is absent. **B:** The evolution of $N/10$ randomly selected gains $a_i = a_i(t)$. The individual gains collapse for large networks, compare Fig. 3. **C:** The population average of the ratio $\sigma^2(y)/\sigma^2_{\mathrm{t}}$ between the actual variance $\sigma^2(y) = \langle (y_i - \bar{y}_i)^2 \rangle / \sigma^2_{\mathrm{t}}$ of the neural activity and the target variance σ^2_{t}. **D:** The largest Lyapunov exponent $\log |\Lambda_{max}|$, where $|\Lambda_{max}|$ is the spectral radius and Λ_{max} the largest eigenvalue, in magnitude, of the locally rescaled weight matrix, see (5).

For the neural activity to be close to (6) one needs to minimize the Kullback-Leibler divergence between the normal distribution $N_{\mu,\sigma}(y)$ and the actual firing rated distribution $\rho(y)$. Doing so, one obtains the adaption rules

$$\begin{aligned} \Delta a &= \epsilon_{\mathrm{a}} \left(1/a - (x-b)\,\Theta\right) \\ \Delta b &= \epsilon_{\mathrm{b}}\, a\, \Theta, \end{aligned} \qquad \Theta = 2y + (1-y^2)(y-\mu)/\sigma^2 \quad . \tag{7}$$

Equivalent equations can be derived when considering $z \to ax - b$ (Schrauwen et al., 2008), instead of $z = a(x-b)$, as done here, or a sigmoidal transfer function (Linkerhand and Gros, 2013). We note that both Δa and Δb are cubic polynomials of y, even though the objectives, the first two moments of $\rho(y)$, involve only linear and quadratic constraints. There are several distinct differences between (7) and (2).

- Despite that σ and μ are the target moments for $\rho(y)$, they are not achieved when the gain and the threshold are adapted under the influence of (7). This is in con-

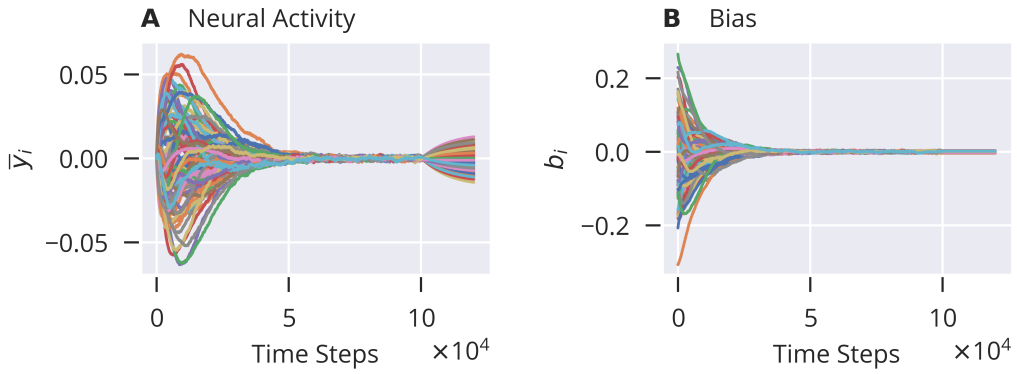

Figure 2: **Threshold adaption.** Simulation parameters as for Fig. 1. At $t_{off} = 10^5$ the external input is turned off together with the adaption of the the gain and the threshold. **A:** The trailing average $\bar{y}_i$ of a random selection of $N/10$ neurons. The average activity approaches the target value $y_t = 0$, recovering slightly once the external input and the adaption of the threshold and the gain is turned off. This happens because the individual b_i are small but non-zero. **B:** The evolution of the individual bias b_i. The stochastic input induces small, barely visible amplitude fluctuations.

trast to (2), which regulates the moments of $\rho(y)$ closely to their target values, y_t and σ_t^2.

- The adaption rule (7) are inherently stable, due to the fact that they result from minimizing an objective function. The variance optimization rules (2) can however not converge when the target values y_t and/or σ_t do not respect that the neural activity $y \in [-1, 1]$ is bounded. A $\mu > 1$ would lead f.i. to an ever decreasing threshold b.

Rescaling y_t and σ_t to realizable values, f.i. via $y_t \to \tanh(\tilde{a} y_t)$, is a route to ensure convergence of (2). The actual target activity would then be $\tilde{y}_t = \tanh(\tilde{a} y_t)$, with $\tilde{a}$ being an appropriate scale. An alternative is to limit unbounded growth dynamically, via

$$\epsilon_a \to (1 - \bar{y}^2)\epsilon_a, \qquad \epsilon_b \to (1 - \bar{y}^2)\epsilon_b \,. \tag{8}$$

In this case the adaption rates become vanishing small when the trailing average $\bar{y}$ of the neural activity approaches the boundary of achievable values, namely ± 1. The rescaling proposed in (8) is a simplification of the dynamical rescaling that results from minimizing appropriate objective functions for (2), as it will be detailed out in the appendix.

3 Simulation results

A typical timeline of network data is presented in Figure 1. We used synchronous updating, as in all simulations. One observes that the activities y_i decay to zero when the external input is turned off, which occurs at $t_{\text{off}} = 10^5$.

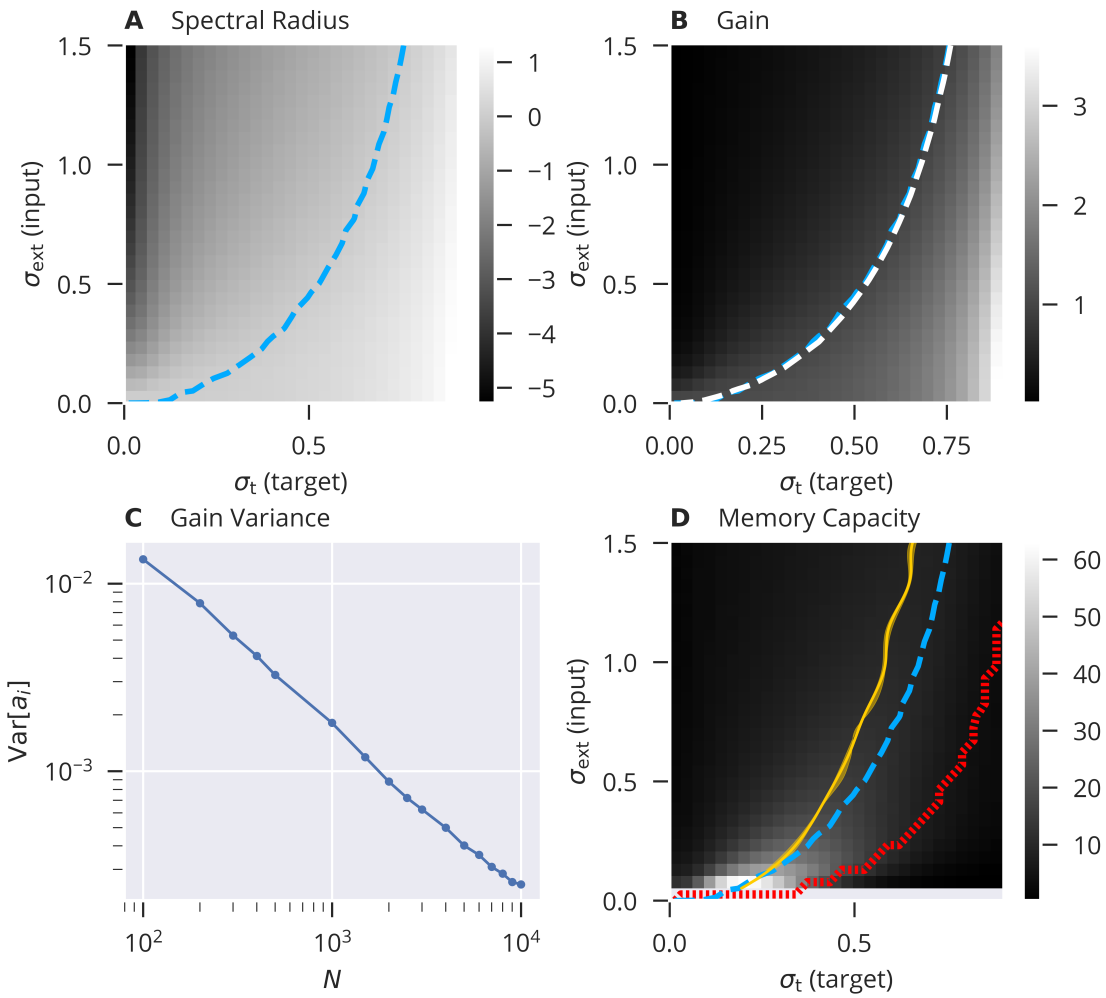

Figure 3: **Parameter sweep.** For a range of $\sigma(\text{target}) = \sigma_\text{t}$, the target for the neuronal activity fluctuations, and $\sigma(\text{input}) = \sigma_\text{ext}$, the standard deviation of the input. **A:** On a log-scale, the spectral radius R of the locally rescaled weight matrix (color coded), as defined by (5). Marked is $\log(R) = 0$ (blue dashed line). **B:** The population average $\langle a_i \rangle_\text{P}$ of the gain a_i (color coded). Marked is $\langle a_i \rangle_\text{P} = 1$ (white dashed line), which is numerically identical to $\log(R) = 0$. **C:** The variance $\text{Var}[a_i] = \langle (a_i - \langle a_i \rangle_\text{P})^2 \rangle_\text{P}$ of the gain as a function of system size N, as log-log plot. The falloff is $\sim 1/N$. **D:** The memory capacity (color coded). Highlighted is the maximal memory capacity for a given σ_ext (orange line, averaged over three trials). Also shown is $\langle a_i \rangle_\text{P} = 1$ (dashed green line) and the onset of the echo state property (striped red line).

The characteristic time scale of the evolution of the gains a_i shown in Figure 1 is $1/\epsilon_\text{b} = 0.5 \cdot 10^4$. For longer times the individual a_i become quasi stationary, with the residual fluctuations being due to the influence of the stochastic input on the network activity. The spread of the gains is a finite-size effect, as shown further below. Also included in Figure 1 is a comparison between the actual variance of the neural activity $\sigma^2(y) = \langle (y_i - \bar{y}_i)^2 \rangle$, with the target variance σ_t^2. One finds that the adaption rule (2) does in indeed lead to $\sigma(y) \to \sigma_\text{t}$. For the parameters selected, $\sigma_\text{t} = 0.2$ and $\sigma_\text{ext} = 0.1$,

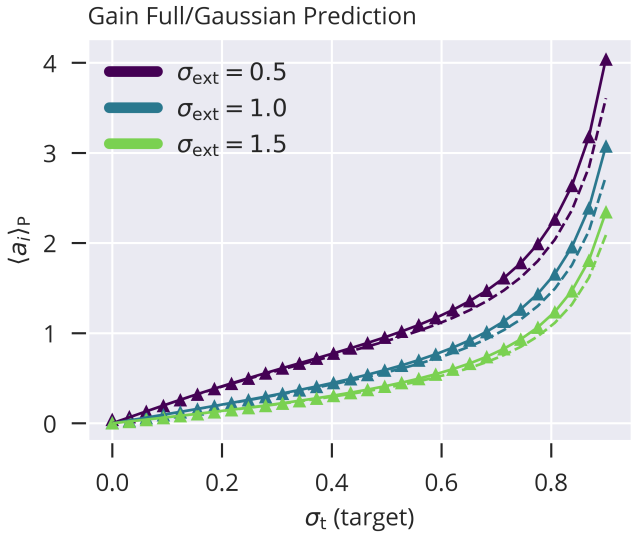

Figure 4: **Theory and simulations compared.** a_i from the simulation (triangles) matches the numerically determined solution (lines) of (16). Dashed lines are approximations given by (20).

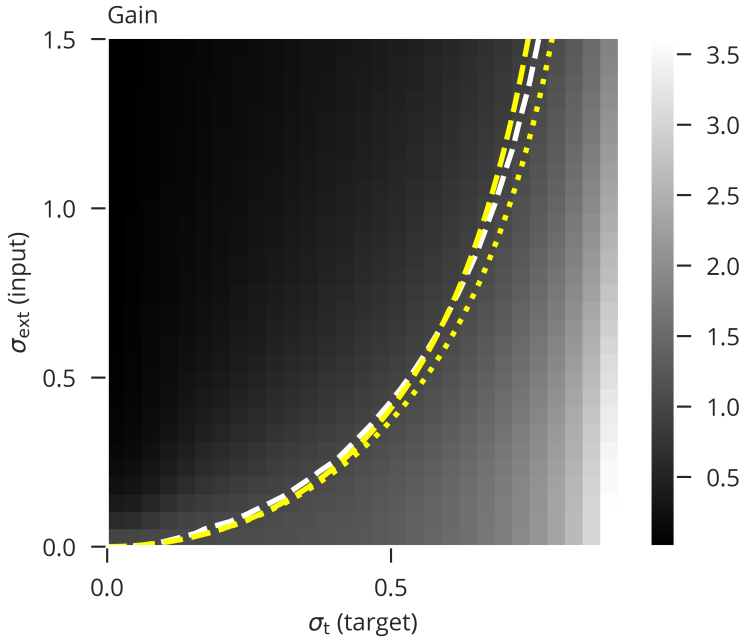

Figure 5: **Spectral Radius Transition Approximation** White dashed lines denotes $\langle a_i \rangle_P = 1/\sigma_w$. Yellow dashed line is given by (25), dotted line by (21).

spectral radius $R = |\Lambda_{max}|$ adapts to a subcritical value, $R < 1$, as shown in Figure 1, which explains the decay of the neural activity once the input is turned off.

For the same parameters as for Figure 1, we present in Fig. 2 the time evolution of the bias b_i and the trailing averages $\bar{y}_i$, as defined by (3). The adaption rule (2) for the

bias ensures that the the average activity approaches the target, namely $y_{\mathrm{t}} = 0$. Note that the stochastic input induces barely-visible amplitude fluctuations for the thresholds b_i.

3.1 Optimal Memory Capacity

The total memory capacity MC was introduced in Jaeger (2002a) as the sum over MC_k and delays $k \in \mathbb{N}$, where MC_k is the squared correlation between input $u(t)$ and optimal readout y_{out} for a delay k:

$$\mathrm{MC} := \sum_{k=1}^{\infty} \mathrm{MC}_k \tag{9}$$

$$\mathrm{MC}_k := \max_{w_{\mathrm{out}}} \frac{\mathrm{Cov}^2\left[u(t-k), y_{\mathrm{out}}(t)\right]_t}{\mathrm{Var}^2\left[u(t)\right]_t \mathrm{Var}^2\left[y_{\mathrm{out}}(t)\right]_t} \tag{10}$$

We measured this quantity for pairs of σ_{t} and σ_{ext}, using a single input sequence, $u(t)$, which was drawn randomly from a Gaussian distribution with zero mean and a variance given by σ^2_{ext}. This signal was fed into a network via a fully connected input weight vector w_{in}, which was drawn from a Gaussian distribution with zero mean and unit variance. This implied that, though individual units received Gaussian input with variance $\sigma^2_{\mathrm{ext}} w^2_{\mathrm{in},i}$, the entire ensemble of inputs would be distributed with a Gaussian of the correct variance σ^2_{ext}.

The resulting performances are shown in Fig. 3D, where we also plotted the curve that traces the σ_{t} values for which the memory capacity is optimal, given σ_{ext}. We see that this curve closely follows the line denoting the stability transition of the zero-activity fixed point (characterized by the spectral radius being one). This is in line with the general "rule of thumb" stating that learning tasks requiring long memory spans perform better with spectral radii close to unity Jaeger (2001). However, we would like to point out that the simple task of memory recall is highly artificial in the sense that it does not require any nonlinear computation on the input. Furthermore, very long memory spans are reached for small values of σ_{t} and σ_{ext}. Since small σ_{t} implies that the network is acting mostly in a linear regime, this is corresponding to the known property that optimal memory capacity is reached in linear reservoirs Jaeger (2002a).

4 Memory Recall with XOR Operation

As stated in the previous section, a simple memory recall task does not require any nonlinear transformation on the input, which means that it performs best for small input and activity variances. We wanted to see how the network performs under different σ_{t} and σ_{ext} if given a task that cannot be solved without nonlinear transformations on the input. For this purpose, we defined a XOR-recall task with delay τ by

$$f(t) = \mathrm{XOR}\left[u(t-\tau), u(t-\tau-1)\right] \tag{11}$$

$$\mathrm{XOR}\left[x, y\right] := \begin{cases} 0 & x = y \\ 1 & \text{else} \end{cases} . \tag{12}$$

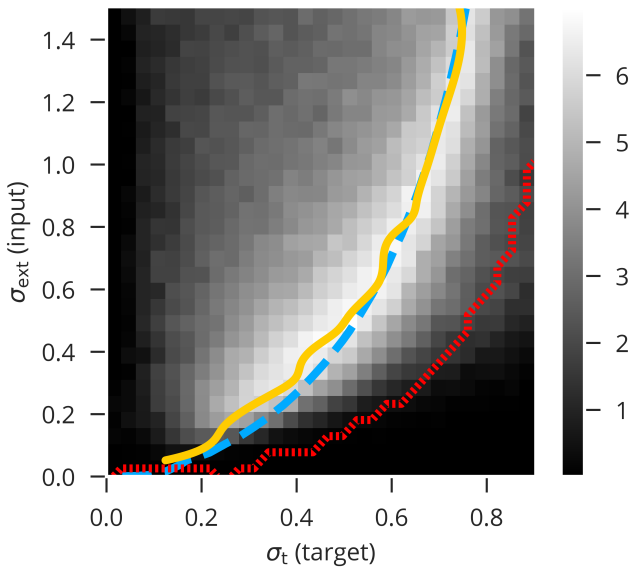

Figure 6: **Memory Capacity for the XOR Task** Colormap encodes $\mathrm{MC}_{\mathrm{XOR}}$ as defined in (13). As in Fig. 3, the blue dashed line marks $\log(R) = 0$ where R is the spectral radius of the rescaled weight matrix. The orange line plots the position of maximal $\mathrm{MC}_{\mathrm{XOR}}$ for a given σ_{ext}. Red striped line marks the loss of the ESP.

In correspondence to (9) and (10), we define the total XOR memory capacity by

$$\mathrm{MC}_{\mathrm{XOR}} := \sum_{k=1}^{\infty} \mathrm{MC}_{\mathrm{XOR},k} \tag{13}$$

$$\mathrm{MC}_{\mathrm{XOR},k} := \max_{w_{\text{out}}} \frac{\mathrm{Cov}^2\left[\mathrm{XOR}\left[u(t-k), u(t-k-1)\right], y_{\text{out}}(t)\right]_t}{\mathrm{Var}^2\left[\mathrm{XOR}\left[u(t), u(t-1)\right]\right]_t \mathrm{Var}^2\left[y_{\text{out}}(t)\right]_t} \,. \tag{14}$$

The results are shown in Fig. 6. In contrast to the simple recall task, the maximally achievable performance remains relatively constant over a wide range of σ_{ext}. This local maximum as a function of σ_{ext} is well described by the condition $R(\widehat{W}_{\text{a}}) = 1$.

5 Theory

In molecular-field theory one assumes that all presynpatice neurons are statistical independent. The optimizing the variance of the neural activity then consists of solving

$$\sigma_{\text{t}}^2 = \int_{-\infty}^{\infty} \mathrm{dx} \tanh^2(ax) N_{\mu,\sigma}(x), \qquad \sigma^2 = \sigma_{\text{w}}^2 \sigma_{\text{t}}^2 + \sigma_{\text{ext}}^2 \,, \tag{15}$$

where the distribution $N_{\mu,\sigma}(x)$ is of the membrane potential x is a Gaussian, see (6), with mean $\mu = 0$ and variance σ^2. A variable transformation $ax = z$ leads to

$$\sigma_{\text{t}}^2 = \int_{-\infty}^{\infty} \mathrm{dz} \tanh^2(z) \frac{1}{\sqrt{2\pi\sigma_{\text{a}}^2}} \mathrm{e}^{-z^2/2\sigma_{\text{a}}^2} \qquad \sigma_{\text{a}}^2 = a^2\left(\sigma_{\text{w}}^2 \sigma_{\text{t}}^2 + \sigma_{\text{ext}}^2\right), \tag{16}$$

where the renormalized variance σ_{a}^2 can be written as

$$\sigma_{\mathrm{a}}^2 = a^2\sigma_{\mathrm{w}}^2\left(\sigma_{\mathrm{t}}^2 + \sigma_{\mathrm{ext}}^2/\sigma_{\mathrm{w}}^2\right). \tag{17}$$

The gain a is adapted such that (16) is fulfilled, defining a 2d-manifold in the $(\sigma_{\mathrm{t}}, \sigma_{\mathrm{ext}}, a)$ space. The relation (17) shows furthermore that the solution of (16) depends on the ratio $\sigma_{\mathrm{ext}}/\sigma_{\mathrm{w}}$, which measures the relevance of the external driving with respect the recurrent synaptic connections. For convenience, we thus introduce $\sigma'_{\mathrm{ext}} \equiv \sigma_{\mathrm{ext}}/\sigma_{\mathrm{w}}$

5.1 Gaussian Approximation

A polynomial approximation of $\tanh^2$ to fourth order captures the right behavior close to the origin, but does not account for the fact that $\tanh^2$ converges to 1 for large absolute values of the membrane potential. Alternatively, an approximation with the correct scaling to second order as well as the right convergence is

$$\tanh^2(x) \approx 1 - \exp\left(-x^2\right). \tag{18}$$

Using this approximation in (16) and solving the integral yields

$$\sigma_{\mathrm{t}}^2 = 1 - \sqrt{1 + 2a^2\sigma_{\mathrm{w}}^2\left(\sigma_{\mathrm{t}}^2 + \sigma'^2_{\mathrm{ext}}\right)}^{-1}. \tag{19}$$

Solving this equation for a gives

$$a = \sigma_{\mathrm{w}}^{-1}\sqrt{\frac{1-\left(1-\sigma_{\mathrm{t}}^2\right)^2}{2\left(1-\sigma_{\mathrm{t}}^2\right)^2\left(\sigma_{\mathrm{t}}^2 + \sigma'^2_{\mathrm{ext}}\right)}}. \tag{20}$$

As shown in Fig. 4B, this approximation matches the exact solution to a high accuracy.

We can also derive an approximation for the critical transition from (19), which is given by

$$\sigma'_{\mathrm{ext}} = \sqrt{\frac{1}{2\left(1-\sigma_{\mathrm{t}}^2\right)^2} - \sigma_{\mathrm{t}}^2 - \frac{1}{2}}. \tag{21}$$

See Fig. 5 for a comparison. One can obtain an even simpler approximation by

$$\sigma'_{\mathrm{ext}} = \sqrt{\frac{1}{2\left(1-\sigma_{\mathrm{t}}^2\right)^2} - \sigma_{\mathrm{t}}^2 - \frac{1}{2}} \tag{22}$$

$$= \sqrt{2}^{-1}\sqrt{\frac{1}{\left(1-\sigma_{\mathrm{t}}^2\right)^2} - 2\sigma_{\mathrm{t}}^2 - 1} \tag{23}$$

$$= \sqrt{2}^{-1}\frac{\sqrt{3\sigma_{\mathrm{t}}^4 - 2\sigma_{\mathrm{t}}^6}}{1-\sigma_{\mathrm{t}}^2} \tag{24}$$

$$\approx \sqrt{\frac{3}{2}}\frac{\sigma_{\mathrm{t}}^2}{1-\sigma_{\mathrm{t}}^2} \tag{25}$$

where we ignored the sixth order term. (25) is also shown in Fig. 5 for comparison. This simple form also allows us conveniently state the inverse relation, which is what is required for tuning the network activity into an optimal state:

$$\sigma_{\mathrm{t}} \approx \left[\sqrt{\frac{3}{2}}\frac{\sigma_{\mathrm{w}}}{\sigma_{\mathrm{ext}}} + 1\right]^{-1/2}. \tag{26}$$

5.2 ESP Transition

The so called *echo state property* was formally described in Jaeger (2001). In short, it states that there exists at least one left-infinity input sequence for which the state sequence of the recurrent network is uniquely determined by this input sequence. In particular, this means that the effect of different initial conditions of the recurrent network will decay. The transition shown in Fig. 3D by the red striped line was determined using this effect by initializing two copies of a network with a small perturbation in the initial conditions and then observing the euclidean distance between both network states while feeding in the same quenched random input.

Assuming that the perturbation is sufficiently small, we model the evolution of this initial offset δ_i by a linear approximation:

$$\delta_i(t+1) \approx a y_i'(t+1) \sum_j w_{ij} \delta_j(t) \tag{27}$$

$$= a \left(1 - \tanh\left(a\left(x_i(t+1) - b_i\right)\right)^2\right) \sum_j w_{ij} \delta_j(t) \tag{28}$$

Furthermore, we make the assumption that the trajectory of the recurrent network state is sampling from phase space in a quasi-random way. This means that we assume neural activity to be essentially uncorrelated in time and across the population, which was also the assumption made for (15). With respect to simplifying (28), this assumption allows us to interpret $\left(1 - \tanh\left(a\left(x_i - b_i\right)\right)^2\right)$ as an independent random vector drawn from $\left(1 - \tanh\left(ax\right)^2\right)$, where x is Gaussian distributed with the usual $\mu_{\mathrm{x}} = 0$, $\sigma_{\mathrm{x}}^2 = \sigma_{\mathrm{w}}^2 \sigma_{\mathrm{t}}^2 + \sigma_{\mathrm{ext}}^2$.

We can write (27) as

$$\delta_i(t+1) = \sum_j A_{ij}(t+1) \delta_j(t) \tag{29}$$

$$A_{ij} := a \left(1 - \tanh\left(a\left(x_i(t+1) - b_i\right)\right)^2\right) w_{ij} \tag{30}$$

To determine the long term evolution of $\|\delta\|(t)$, we would like to evaluate the average growth factor

$$\gamma = \left\langle \frac{\|\delta(t+1)\|}{\|\delta(t)\|} \right\rangle_{\mathrm{t}} \tag{31}$$

Since (27) is a linear system, we can rescale terms in (31) such that the denominator becomes one and only evaluate the change in length for each time step:

$$\gamma = \langle \|\Delta(t)\| \rangle_t \tag{32}$$

$$\Delta(t) := \widehat{A}(t) \delta'(t-1) = \widehat{A}(t) \delta(t-1) / \|\delta(t-1)\| \tag{33}$$

In principle, $\delta(t)$ and thus also $\Delta(t)$ are not independent in time, since they are iteratively generated. However, since all $A(t)$ are assumed to be independent, we can also regard $\delta(t)$ as statistically independent in time (THIS IS VERY HAND-WAVY, BUT I'M NOT SURE HOW TO FURTHER JUSTIFY THIS).

We express (32) as

$$\gamma = \left\langle \sqrt{\sum_i \left(\sum_j A_{ij}(t+1)\delta'_j(t) \right)^2} \right\rangle_{\mathrm{t}} . \tag{34}$$

For large N, the inner sum follows a Gaussian distribution with $\mu_\sum = N\mu_A\mu_{\delta'}$ and $\sigma^2_\sum = N\sigma^2_A\sigma^2_{\delta'}$. Since $\mu_A = 0$, we have

$$\gamma = \left\langle \sqrt{\sum_i X_i^2(t)} \right\rangle_{\mathrm{t}} \tag{35}$$

$$= \sigma_\sum \left\langle \sqrt{\sum_i X_i^2(t)} \right\rangle_{\mathrm{t}} \tag{36}$$

where X follows a standard Gaussian distribution. The remaining sum obeys a Chi-squared distribution which, for large N, again follows a Gaussian distribution with mean N and variance $2N$, which allows us to ignore fluctuation around the mean for large N:

$$\gamma = \sigma_\sum \left\langle \sqrt{\sqrt{2N}X(t) + N} \right\rangle_{\mathrm{t}} \tag{37}$$

$$\approx \sigma_\sum \sqrt{N} \tag{38}$$

$$= N\sigma_A\sigma_{\delta'} \tag{39}$$

The empirical variance of δ' is

$$\sigma^2_{\delta'} = \frac{1}{NT}\sum_{i,t} \delta'_i(t)^2 - \mu^2_{\delta'} . \tag{40}$$

For all t, the sum over i is one by definition. $\delta'(t)$ is a sequence of points on a N-dimensional unit sphere. A non-zero mean value would imply that there is a "preferred direction" in the sequence distribution. However, since we have assumed that the mapping $\widehat{A}(t)$ is randomly drawn from statistics that are the same for each coordinate (symmetric under exchange of coordinates), this mapping will generate a distribution with the same symmetrical property (VERY BOLD STATEMENT), thus giving zero mean.

We now find for γ

$$\gamma = \sqrt{N}\sigma_A \tag{41}$$

$$= \sqrt{N}a\sqrt{\mathrm{E}[w^2]\mathrm{E}[y'^2] - \mathrm{E}[w]^2\mathrm{E}[y']^2} \tag{42}$$

$$= \sqrt{N}a\sqrt{\mathrm{E}[w^2]\mathrm{E}[y'^2]} \tag{43}$$

$$= a\sigma_{\mathrm{w}}\sqrt{\mathrm{E}[y'^2]} \tag{44}$$

$$= a\sigma_{\mathrm{w}}\sqrt{\int_{-\infty}^{\infty} \mathrm{dx}\left(1 - \tanh^2(ax)\right)^2 N_{\mu=0,\sigma}(x)}, \qquad \sigma^2 = \sigma^2_{\mathrm{w}}\left(\sigma^2_{\mathrm{t}} + {\sigma'_{\mathrm{ext}}}^2\right) . \tag{45}$$

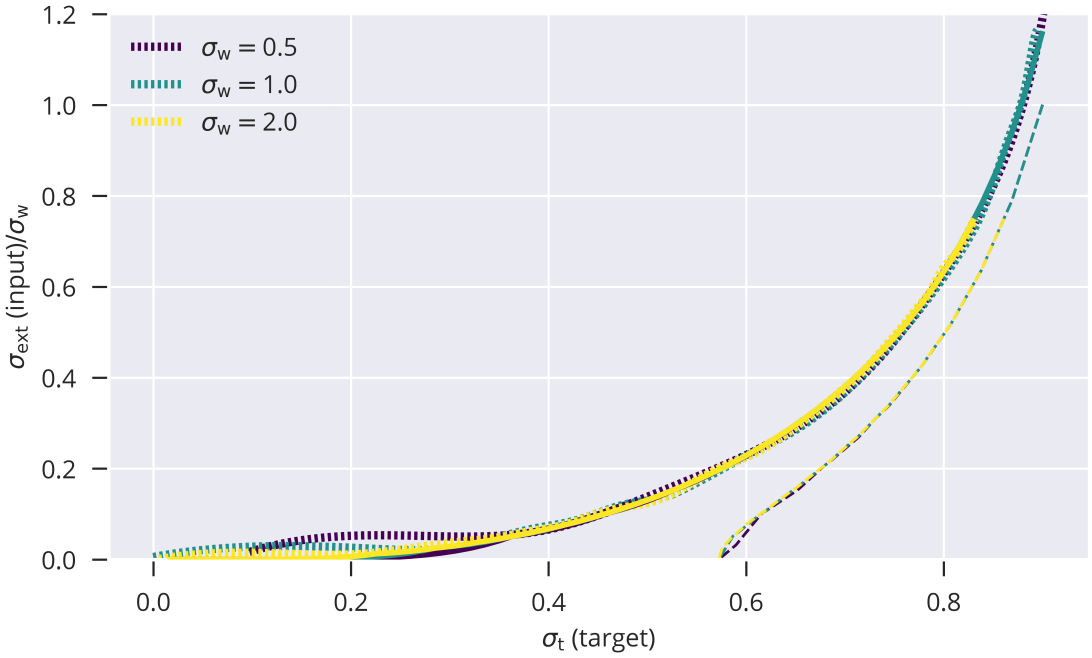

Figure 7: Striped lines: ESP-transition for full simulations. Dashed lines: approximation given by (46). Full lines: Exact solution defined by (45) and (15).

Similar to the approximation used to find an explicit equation for $a(\sigma_w, \sigma'_{ext})$, we use $\left(1 - \tanh^2(ax)\right)^2 \approx \exp\left(-2a^2x^2\right)$, leading to

$$\gamma \approx a\sigma_w \left(1 + 4a^2\sigma^2\right)^{-1/4} . \tag{46}$$

As Fig. 7 shows, numerically evaluating the exact solution of (45) and (15) for $\gamma = 1$ accurately describes the ESP-transition found in the simulations. However, using the aforementioned approximations leads to a significant mismatch. In particular, it does not correctly describe the transition for external inputs going to zero.

Conclusion

We have illustrated the basic format to the manuscript that you consider to submit to Neural Computation. We hope this is helpful to the authors.

Acknowledgments

The people you want to acknowledge. For this document, we appreciate Jrg Lcke, author of an accepted paper who generously allowed us to use his template.

Appendix

You should put the details that are not required in the main body into this Appendix.

References

Arviv, O., A. Goldstein, and O. Shriki
2015. Near-critical dynamics in stimulus-evoked activity of the human brain and its relation to spontaneous resting-state activity. *Journal of Neuroscience*, 35(41):13927–13942.

Berkes, P., G. Orbán, M. Lengyel, and J. Fiser
2011. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–87.

Boedecker, J., O. Obst, J. T. Lizier, N. M. Mayer, and M. Asada
2012. Information processing in echo state networks at the edge of chaos. *Theory in Biosciences*, 131(3):205–213.

Boedecker, J., O. Obst, N. M. Mayer, and M. Asada
2009. Initialization and self-organized optimization of recurrent neural network connectivity. *HFSP journal*, 3(5):340–349.

Caluwaerts, K., F. Wyffels, S. Dieleman, and B. Schrauwen
2013. The spectral radius remains a valid indicator of the echo state property for large reservoirs. In *Neural Networks (IJCNN), The 2013 International Joint Conference on*, Pp. 1–6. IEEE.

Cannon, J. and P. Miller
2017. Stable control of firing rate mean and variance by dual homeostatic mechanisms. *The Journal of Mathematical Neuroscience*, 7(1):1.

Caraballo, T., G. Łukaszewicz, and J. Real
2006. Pullback attractors for asymptotically compact non-autonomous dynamical systems. *Nonlinear Analysis: Theory, Methods & Applications*, 64(3):484–498.

Cocchi, L., L. L. Gollo, A. Zalesky, and M. Breakspear
2017. Criticality in the brain: A synthesis of neurobiology, models and cognition. *Progress in neurobiology*, 158:132–152.

Echeveste, R. and C. Gros
2014. Generating functionals for computational intelligence: the fisher information as an objective function for self-limiting hebbian learning rules. *Frontiers in Robotics and AI*, 1:1.

Enel, P., E. Procyk, R. Quilodran, and P. F. Dominey
2016. Reservoir computing properties of neural dynamics in prefrontal cortex. *PLoS computational biology*, 12(6):e1004967.

Farkaš, I., R. Bosák, and P. Gergel'
2016. Computational analysis of memory capacity in echo state networks. *Neural Networks*, 83:109–120.

Gallicchio, C. and A. Micheli
2017. Echo state property of deep reservoir computing networks. *Cognitive Computation*, 9(3):337–350.

Gros, C.
1990. Criterion for a good variational wave function. *Physical Review B*, 42(10):6835.

Gros, C.
2009. Cognitive computation with autonomously active neural networks: an emerging field. *Cognitive Computation*, 1(1):77–90.

Gros, C.
2015. *Complex and adaptive dynamical systems: A primer*. Springer.

Jaeger, H.
2001. The "echo state" approach to analysing and training recurrent neural networks. GMD Report 148, GMD - German National Research Institute for Computer Science.

Jaeger, H.
2002a. Short Term Memory in Echo State Networks. GMD Report 152, Fraunhofer Institute for Autonomous Intelligent Systems.

Jaeger, H.
2002b. *Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the" echo state network" approach*, volume 5. GMD-Forschungszentrum Informationstechnik Bonn.

Jaeger, H.
2005. Reservoir riddles: Suggestions for echo state network research. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 3, Pp. 1460–1462. IEEE.

Kloeden, P. E.
2000. Pullback attractors in nonautonomous difference equations. *Journal of Difference Equations and Applications*, 6(1):33–52.

Linkerhand, M. and C. Gros
2013. Self-organized stochastic tipping in slow-fast dynamical systems. *Mathematics and Mechanics of Complex Systems*, 1(2):129–147.

Livi, L., F. M. Bianchi, and C. Alippi
2018. Determination of the edge of criticality in echo state networks through fisher information maximization. *IEEE Transactions on Neural Networks and Learning Systems*, 29(3):706–717.

Lukoševičius, M. and H. Jaeger
2009. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149.

Maass, W., T. Natschläger, and H. Markram
2002. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, 14(11):2531–2560.

Manjunath, G. and H. Jaeger
2013. Echo state property linked to an input: Exploring a fundamental characteristic of recurrent neural networks. *Neural computation*, 25(3):671–696.

Marder, E. and J.-M. Goaillard
2006. Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience*, 7(7):563.

Markovic, D. and C. Gros
2010. Self-organized chaos through polyhomeostatic optimization. *Physical Review Letters*, 105(6):068702.

Mitra, A. and M. E. Raichle
2016. How networks communicate: propagation patterns in spontaneous brain activity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1705):20150546.

Nikolić, D., S. Häusler, W. Singer, and W. Maass
2009. Distributed fading memory for stimulus properties in the primary visual cortex. *PLoS biology*, 7(12):e1000260.

Ozturk, M. C., D. Xu, and J. C. Príncipe
2007. Analysis and design of echo state networks. *Neural computation*, 19(1):111–138.

Petermann, T., T. C. Thiagarajan, M. A. Lebedev, M. A. Nicolelis, D. R. Chialvo, and D. Plenz
2009. Spontaneous cortical activity in awake monkeys composed of neuronal avalanches. *Proceedings of the National Academy of Sciences*, 106(37):15921–15926.

Schrauwen, B., M. Wardermann, D. Verstraeten, J. J. Steil, and D. Stroobandt
2008. Improving reservoirs using intrinsic plasticity. *Neurocomputing*, 71(7-9):1159–1171.

Schuecker, J., S. Goedeke, and M. Helias
2018. Optimal sequence memory in driven random networks. *Physical Review X*, 8(4):041029.

Sompolinsky, H., A. Crisanti, and H.-J. Sommers
1988. Chaos in random neural networks. *Physical review letters*, 61(3):259.

Triesch, J.
2005. A gradient rule for the plasticity of a neurons intrinsic excitability. In *International Conference on Artificial Neural Networks*, Pp. 65–70. Springer.

Wainrib, G. and M. N. Galtier
2016. A local echo state property through the largest lyapunov exponent. *Neural Networks*, 76:39–45.

Wernecke, H., B. Sándor, and C. Gros
2019. Chaos in time delay systems, an educational review. *arXiv preprint arXiv:1901.04826*.

Yildiz, I. B., H. Jaeger, and S. J. Kiebel
2012. Re-visiting the echo state property. *Neural networks*, 35:1–9.