Evaluation of Information-Theoretic Measures in Echo State Networks on the Edge of Stability

Miloslav Torda and Igor Farkaš
Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava
Mlynská dolina, 84248 Bratislava, Slovak Republic
Email: farkas@fmph.uniba.sk

***Abstract*—It has been demonstrated that the computational capabilities of echo state networks are maximized when the recurrent layer is close to the border between a stable and an unstable dynamics regime, the so called edge of stability, or criticality. The maximization of performance is computationally useful, leading to minimal prediction error or maximal memory capacity, and has been shown to lead to maximization of information-theoretic measures, such as transfer entropy and active information storage in case of some datasets. In this paper, we take a closer look at these measures, using Kraskov–Grassberger–Stögbauer estimator with optimized parameters. We experiment with four datasets differing in the data complexity, and discover interesting differences, compared to the previous work, such as more complex behavior of the information-theoretic measures. We also investigate the effect of reservoir orthogonalization, that has been shown earlier to maximize memory capacity, on the prediction accuracy and the above mentioned measures.**

I. Introduction

Reservoir computing (RC) has received considerable attention regarding the effect of reservoir properties on information processing in the neural network models employed in various tasks such as time series prediction or input reconstruction (reflecting the memory properties of the model). It has been shown [1], and further confirmed in subsequent studies, e.g. [2], [3], that the task performance is maximized when the network is operating in a state near the edge of chaos, or stability. It concerns the critical state between a stable (ordered) regime when disturbances are attenuated and an unstable (chaotic) regime where disturbances are amplified (hence deteriorating performance). The critical regime seems interesting also from a biological perspective, since it has been hypothesized that cortical circuits may be tuned to criticality for optimized behavior [4], [5] (but see also [6] for critical assessment).

In the case of the echo state networks (ESNs) having discrete-time dynamics, a lot of work has been done regarding proper initialization of the reservoir matrix, its proper tuning, including the reservoir orthogonalization (see overview in [7]). Regarding the memory properties of ESNs, Jaeger [8] defined and quantified the short-term memory capacity (MC) that measures the network ability to reconstruct the past input information from the reservoir on the network output by computing squared correlations. Orthogonal ESNs have been shown to increase the MC to a certain degree and we recently showed [9], using two gradient-based orthogonalization procedures, that this increase can approach the theoretical limit proved by Jaeger [8] for linear reservoirs to be equal to the reservoir size.

The critical regime has also been investigated from the view of information processing by evaluating the information-theoretic measures such as active information storage [10] and transfer entropy [11]. It was shown, in case of two datasets [12], that both measures are maximized near the edge of stability (criticality). In this work, we replicated these results, but extended the investigation using four datasets, and taking a closer look at these measures, leading to a more complex interpretation of the results. In addition, we investigate the effect of reservoir orthogonalization on these measures as well as on ESN performance.

The paper is organized as follows. In Section II we provide background information about the underlying theory and the methods used. Section III presents results of experiments. Section IV concludes the paper.

II. Theory and methods

Here we introduce the model of ESN we used, as well as measures used for its evaluation on two different tasks (memory capacity and time series prediction) using four datasets.

A. Echo state network model

For the purpose of experiments, we assume an ESN model with a single input $u(t)$, N reservoir neurons and Q output neurons. Reservoir activations $\mathbf{x}(t) = (x_1(t), \ldots, x_N(t))^\top$ and output activations $\mathbf{y}(t) = (y_1(t), \ldots, y_Q(t))^\top$ are updated according to ESN dynamics given by the formulas

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{f}(\mathbf{w}^{\text{in}} u(t) + \mathbf{W}\mathbf{x}(t-1)) \\ \mathbf{y}(t) &= \mathbf{f}^{\text{out}}(\mathbf{W}^{\text{out}}\mathbf{x}(t)) \end{aligned}$$

where $\mathbf{f} : \mathbb{R}^N \to \mathbb{R}^N$ and $\mathbf{f}^{\text{out}} : \mathbb{R}^N \to \mathbb{R}^Q$ are suitable activation functions. We use nonlinear $f = \tanh$ and the linear readout $\mathbf{f}^{\text{out}} = \text{id}$ (both applied element-wise). The weight vector $\mathbf{w}^{\text{in}}$ refers to input weights, $\mathbf{W}$ and $\mathbf{W}^{\text{out}}$ are recurrent and output weight matrices, respectively. Readout weights are computed as $\mathbf{W}^{\text{out}} = \mathbf{U}\mathbf{X}^+$, where the matrix $\mathbf{U}$ is created either by concatenation of the target vectors (corresponding to past inputs with different delays) in MC task, or as a single target vector (in case of the prediction task), and $\mathbf{X}^+ = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-\mathbf{1}}$ is the Moore–Penrose pseudoinverse matrix of concatenated state vectors.

B. Information-theoretic measures

In order to get deeper insight into information processing in an ESN, we use fundamental quantities of information theory,

 978-1-5090-6014-6/18/$31.00

capturing the dynamics of information in the network. The fundamental question the measures address is: "where does the information in a random variable V_{t+1} in a time series come from?" [13]. This question is addressed in terms of information from the past of process V (the information storage), or information contributed from other source processes Z (the information transfer). The *active information storage* (AIS) A_V [10] measures how much of the information is observed to be in use in computing its next state. A_V is the local mutual information between realizations $\mathbf{v}_t^{(k)} = [v_t, v_{t-1}, ..., v_{t-k+1}]$ of the past state $\mathbf{V}_t^{(k)}$ and the corresponding realizations v_{t+1} of the next value V_{t+1}. Formally, local AIS a_V is defined as

$$a_V(k, t+1) = \log_2 \frac{p(v_{t+1}|\mathbf{v}_t^{(k)})}{p(v_{t+1})}$$

The local values of AIS measure the dynamics of information storage at different time points within a system, revealing how the use of memory fluctuates during a process. The average $A_V^{(k)}$ is the time-average of the local values $a_V(t+1, k)$.

Information transfer, as a directed measure, is defined as the amount of information that a source process provides about a target (or destination) process' next state that was not contained in the target's past. The *transfer entropy* (TE) [11] captures the average mutual information (MI) from realizations $\mathbf{z}_t^{(l)}$ of the state $\mathbf{Z}_t^{(l)}$ of a source time-series process Z to the corresponding realizations z_{t+1} of the next value Z_{t+1} of the target time-series process V, conditioned on realizations $\mathbf{v}_t^{(k)}$ of the previous state $\mathbf{V}_t^{(k)}$, i.e. $T_{Z\to V}(k, l) = \mathrm{MI}(V_{t+1}; \mathbf{Z}_t^{(l)}|\mathbf{V}_t^{(k)})$. The local TE quantifies the amount of information transfer attributed to the specific configuration or realization $(v_{t+1}, \mathbf{v}_t^{(k)}, \mathbf{z}_t^{(l)})$ at time step $t+1$:

$$t_{Z\to V}(k, l) = \log_2 \frac{p(v_{t+1}|\mathbf{v}_t^{(k)}, \mathbf{z}_t^{(l)})}{p(v_{t+1}|\mathbf{v}_t^{(k)})}\,.$$

These local information transfer values measure the dynamics of transfer in time between any given pair of processes within a system.

C. Estimation of information-theoretic measures

For approximation of TE we use the Kraskov–Grassberger–Stögbauer estimator [14] that has been formulated and argued in [13] (p. 48) to be more data efficient and accurate than other techniques. Precisely,

$$\widehat{\mathrm{TE}}_{Z\to V}^{(k,l)} = \Psi(K) + \\ + \langle \Psi(n_{\mathbf{v}_t^{(k)}} + 1) - \Psi(n_{v_{t+1}, \mathbf{v}_t^{(k)}} + 1) - \Psi(n_{\mathbf{v}_t^{(k)}, \mathbf{z}_t^{(l)}} + 1) \rangle_t$$

where Ψ denotes the digamma function, $n_{(\cdot)}$ denotes the number of nearest neighbors in ϵ-hypercubes centered at $(\cdot)$ in the marginal spaces where ϵ is given by the Chebyshev distance of the realization $(v_{t+1}, \mathbf{v}_t^{(k)}, \mathbf{z}_t^{(l)})$ at time step t to its Kth nearest neighbor in the joint space, and $\langle\cdot\rangle_t$ denotes the time-average. Similarly the AIS estimator is written as

$$\widehat{\mathrm{AIS}}_V^{(k)} = \Psi(K) + \Psi(N) - \langle \Psi(n_{\mathbf{v}_t^{(k)}} + 1) + \Psi(n_{v_{t+1}} + 1) \rangle_t.$$

where N is the sample size.

For finding the optimal parameters of the KGS estimator, we apply the Ragwitz–Kantz criterion [15] that is based on scanning the (k, τ) plane to identify the point in that plane that minimizes the locally constant predictor error, i.e.

$$(k, \tau) = \arg\min \|\mathbf{v} - \hat{\mathbf{v}}(k, \tau)\| \tag{1}$$

where $k \in \mathcal{Z}, \tau \in \mathcal{Z}$, $\mathbf{v}(k, \tau)$ is the entire tested sequence (observed on a single neuron), $\hat{\mathbf{v}}(k, \tau)$ is its estimate and the embedded state vector $\mathbf{v}_t^{(k)} = [v_t, v_{t-\tau}, ..., v_{t-(k-1)\tau}]$. The estimated state vectors are found as

$$\hat{\mathbf{v}}_{t+1} = \frac{1}{\mathrm{card}(\mathcal{U}_t)} \sum_{\mathbf{v}_l \in \mathcal{U}_t} \mathbf{v}_{l+1} \tag{2}$$

where $\mathcal{U}_t = \{\mathbf{v}_l : \|\mathbf{v}_l - \mathbf{v}_t\| \le \epsilon\}$. The parameter ϵ results from the number (K) of chosen nearest neighbors (NN).

Since we are using ESNs, we assume that only the most recent activity of the source unit is a causal contributor to the activity of the target unit, i.e. $l = 1$. Hence, no parameter search is made for optimal embedding of the source unit.

D. Relative entropy of reservoir transfer entropy

In order to assess the change in TE due to the reservoir scaling, we computed relative entropy, known as Kullback–Leibler divergence, that measures the distance of the tested distribution from the uniform distribution on the interval [0;1]. [1] This can be used to quantify the TE differences depending on the reservoir scaling (the higher the KL value, the larger the distance). The estimator for relative entropy of the reservoir transfer entropy in our case is computed as

$$\widehat{D_{\mathrm{KL}}}(\mathrm{TE}_{\mathrm{RES}}^{(k,l)} \,\|\, \mathcal{U}(0;1)) = \ln(1) - \widehat{H_{\mathrm{KL}}}(\mathrm{TE}_{\mathrm{RES}}^{(k,l)})$$

where $\ln(1)$ is the exact differential entropy of the uniform distribution on $[0; 1]$, $\widehat{H_{\mathrm{KL}}}(\cdot)$ is the Kozachenko–Leonenko entropy estimator and $\mathrm{TE}_{\mathrm{RES}}^{(k,l)}$ is the distribution of the estimates of transfer entropies in the reservoir.

E. Memory capacity

Jaeger [8] introduced (short term) memory capacity (MC), as a measure for the ability of the reservoir to store and recall previous inputs fed into the network. Jaeger defined it as

$$\mathrm{MC} = \sum_{q=1}^{q_{\max}} \mathrm{MC}_q = \sum_{q=1}^{q_{\max}} \frac{\mathrm{cov}^2(u(t-q), y_q(t))}{\mathrm{var}(u(t)) \cdot \mathrm{var}(y_q(t))} \tag{3}$$

where 'cov' denotes covariance (of the two time series), 'var' means variance, $q_{\max} = \infty$, $u(t-q)$ is the input presented q-steps before the current input, and $y_q(t) = \mathbf{w}_q^{\mathrm{out}}\mathbf{x}(t) = \tilde{u}(t-q)$ is its reconstruction at the network output (using linear readout), where $\mathbf{w}_q^{\mathrm{out}}$ is the weight vector of q-th output unit. The computation of MC is approximated using $q_{\max} = Q$ (i.e. given by the number of output neurons). The concept of MC is based on the network ability to retrieve the past information (for various delays q) from the reservoir using the linear combinations of reservoir unit activations observed at the output (quantified by MC_q).

[1] We chose the uniform distribution as it maximizes the differential entropy when no prior knowledge about distribution is available.

F. Reservoir setting

Memory capacity depends on reservoir properties. Papers [16] and [7] provide a concise overview of practical tips on the reservoir initialization in ESNs (but see also [17], [18] for more recent results). Spectral radius is not a universally acceptable indicator of (non)existence of echo states. Nevertheless, $\rho(\mathbf{W}) \approx 1$ tends to lead to higher MC, as investigated also in [9], [19], where we also investigated the effect of two iterative orthogonalization procedures (OG and ON) of reservoir weight vectors on memory capacity of an ESN. We showed that both procedures helped MC increase to almost reach the theoretical limit (N). Here we test the effect of these procedures on information measures and the ESN performance in case of four different datasets.

G. Estimating the criticality

In order to monitor the changes of information measures, one can look at the stability properties of the reservoir. The well-known approach from the literature is the (characteristic) Lyapunov exponent (LE, or λ), based on evaluating the average sensitivity to perturbations of the initial conditions [1]. LE is computed for trained ESNs, considering all reservoir neurons, one at a time, and averaging over their sensitivity to perturbations over the large enough temporal interval. Ordered state in ESN occurs for $\lambda < 0$, whereas $\lambda > 0$ implies unstable state. Hence, a bifurcation occurs at $\lambda \approx 0$ (the critical point, or the edge of stability). Since λ is by definition an asymptotic quantity, it has to be estimated for most dynamical systems. We used the method described in [20] and replicated in [9].

III. Experiments

A. Experimental setup

We use the ESNs with $N = 100$ reservoir units, a single input unit and the number of output units Q dependent on the task (120 units for the MC task, and a single unit for the prediction task). The elements of the input weight vector $\mathbf{w}^{\text{in}}$ were initialized from the uniform distribution $\mathcal{U}(-0.1; 0.1)$ and elements of the recurrent weight matrix $\mathbf{W}$ from the normal distribution $\mathcal{N}(0; 0.5)$, in order to set the reservoir to a unstable regime in all four dataset cases. For reservoir scaling, we set the spectral radius ρ to the desired values 0.6 and 0.95, in order to get to a stable and a close-to-critical regimes, respectively. The reservoirs scaled to $\rho = 0.95$ were further orthogonalized by the OG or ON method. In all experiments, after discarding the first 100 samples to get rid of transients, we used 1000 samples for setting the readout weights and the next 2000 samples of the time series for testing the model performance.

For experiments, we use four datasets. For testing memory capacity (MC), we consider (as in our previous work) an unstructured one-dimensional input: a sequence of independent and identically distributed (i.i.d.) real numbers from the interval $[-1; 1]$. For time series prediction, we use three benchmark datasets: (1) a nonlinear NARMA model, generated by the equation $u(t+1) = 0.2\,u(t) + 0.004\,u(t)\sum_{i=0}^{29} u(t-i) + 1.5\,q(t-29)q(t) + 0.001$, where the driving input $q(t)$ is sampled from a uniform random distribution $\mathcal{U}(0; 0.5)$, (2) Mackey–Glass (M–G) system [21] with parameters $\tau = 17, \beta = 0.2, \gamma = 0.1, n = 10$, and (3) the x-coordinate of the Lorenz system [22] with parameters $\sigma = 10, \rho = 28, \beta = 8/3$. The reason for using three different time series was to investigate potential differences in TE and AIS depending on the reservoir scaling, and the potential effect of reservoir orthogonalization(s). Regarding the complexity, NARMA is known to be more complex than the other two time series. The prediction performance is measured by normalized root mean squared error computed as

$$\text{NRMSE} = \sqrt{\frac{\langle (\hat{y}(t) - y(t))^2 \rangle_t}{\langle (y(t) - \langle y(t) \rangle_t)^2 \rangle_t}}$$

where $\hat{y}(t)$ denotes the predicted output and $y(t)$ is the true output.

B. Information-theoretical measures around criticality

As a first step, we looked at TE behavior as a function of the estimated Lyapunov exponent (introduced in Section II-G), whose values can be used for monitoring the reservoir states around the critical regime. Using the same initialization of $\mathbf{w}^{\text{in}}$ and $\mathbf{W}$, we increased σ, with 5 simulations per value, such that $\log\sigma$ varied within the interval $[-1.5; -0.25]$, with a step 0.1. In order to get different values of λ, we increased the variance σ of the distribution from which the elements of $\mathbf{W}$ are drawn, with 5 simulations per value, such that $\log\sigma$ varied within the interval $[-1.5; -0.25]$, with a step 0.1 using the same initialization of $\mathbf{w}^{\text{in}}$ and we computed λ afterwards. Regarding the KGS estimator needed for calculation of information measures, we used $k = 2, \tau_k = 1, l = 1, \tau_l = 1$, and 4-NN (skipping the grid search for optimal values, to be done in the next step). Since the KGS estimator has some inherent systematic error, for TE and AIS close to zero the $\widehat{\text{TE}}$ and $\widehat{\text{AIS}}$ estimates can be negative. From the definition of TE and AIS as special cases of Kullbeck–Leibler divergence, the values can be only non-negative. Therefore we have set all the negative values of TE and AIS in the experiments to zero since the purpose of the experiments was not to test whether there is some information transfer but to compare changes of information transfer in various settings.

Figure 1 provides the results. In MC and NARMA tasks, the performance peaks when TE is increased, as observed in [12], but in the stable regime, when approaching $\lambda = 0$, both TE and AIS decrease while MC grows (as seen from a more detailed inset). Hence, there exists a non-monotonous relationship between the measures and the performance. For $\lambda > 0$ the dependence looks similar to [12], i.e. significant decrease of both measures and the performance (i.e. error growth in NARMA task). Another difference between our results and [12] is that our AIS maximum at criticality in NARMA dataset is much higher (3 versus 1.6). All differences may arise from the fact that they used kernel estimation with a fixed radius of 0.2 (and $k = 2$ as we did), whereas we used K-NN estimator for information measures. The other two datasets, M–G and Lorenz attractor, not tested in [12], yield quite different results. First, in both cases, TE peaks at criticality, but grows from left (unlike previous datasets), manifesting monotonous relationship. Second, AIS does not peak at criticality but keeps its high values all over the stable regime. Third, for $\lambda > 0$, the performance degrades rapidly in case of M–G, but only gradually in case of the Lorenz system, which implies that the latter dataset is less sensitive to disturbances in an unstable reservoir.

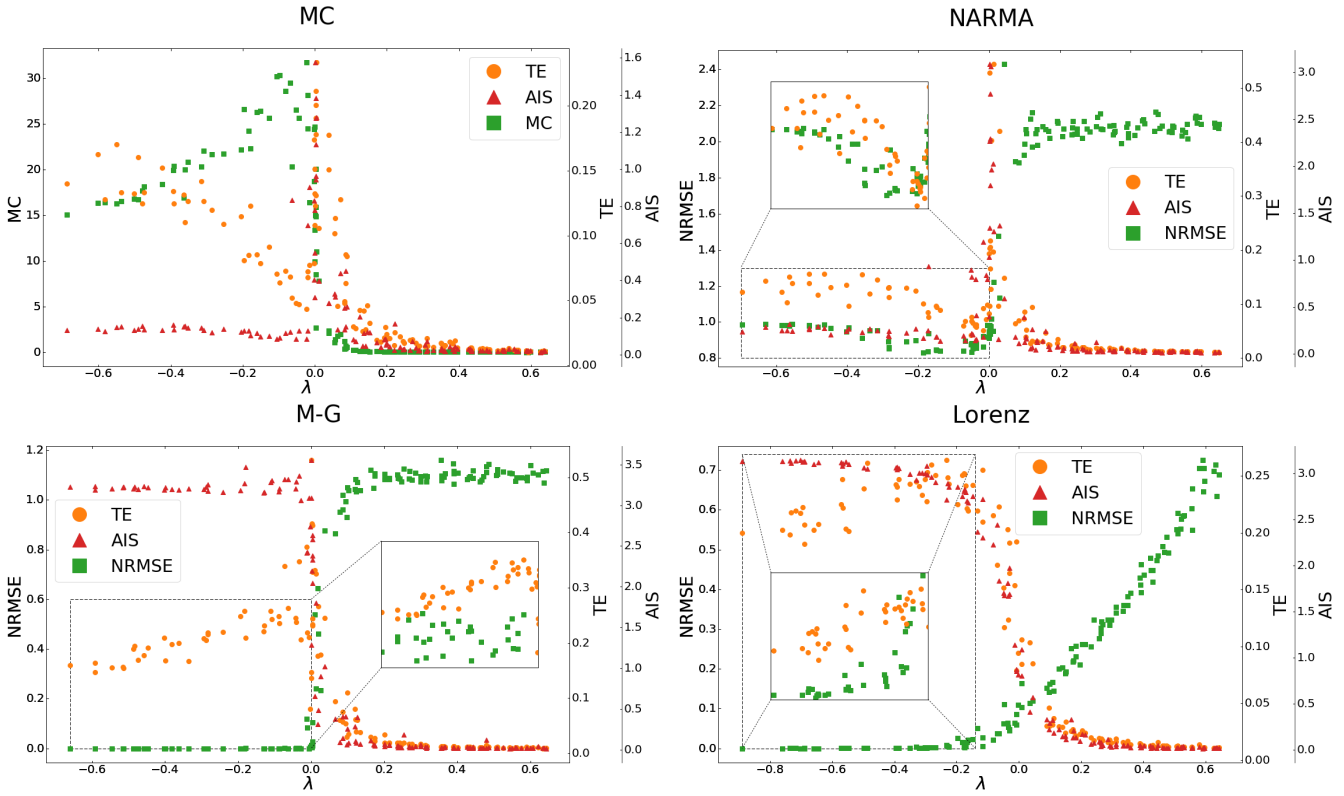

Fig. 1. Average values of TE and AIS in the reservoir and the task performance as a function of the Lyapunov exponent, in case of four datasets. For interpretation, see the text.

C. Optimization of the KGS estimator

The second step was to get a more detailed view at information flow in the reservoir, and, hence, TE. Before doing so, we searched for optimal parameters of the KGS estimator using the simplifying assumption (to reduce the search space) that only the most recent value of the source unit causally affects the target unit, hence $l = 1$. We varied both parameters k and τ from 1 to 6, and for each pair, 100 instances were run. Results are shown in Figure 2. The dependences have single maxima for all unscaled reservoirs, and remain unimodal after scaling in all cases (M–G shows a slightly bimodal pattern), with $\tau = 1$ (left column). The pattern for Lorenz dataset is less focused, especially for $\rho = 0.95$. Optimal values of the KGS estimator that we chose in further experiments are shown in Table I. Using these values implies that the maximum possible amount of information is extracted from the unit's history, while the rest has to be extracted from the source units.

TABLE I. OPTIMAL VALUES OF PARAMETERS, USED IN SUBSEQUENT EXPERIMENTS FOR ESTIMATING INFORMATION-THEORETICAL MEASURES.

Scaling \ Task	MC		NARMA		M–G		Lorenz	
	k	τ	k	τ	k	τ	k	τ
init	6	1	6	1	6	1	6	1
$\rho = 0.6$	4	1	4	1	2	1	2	3
$\rho = 0.95$	6	1	6	1	2	1	2	2

D. A closer look at transfer entropy

Having found optimal values of KGS estimator, we can proceed to computing transfer entropy. We generated one instance of an input weight vector $\mathbf{w}^{\text{in}}$ and a recurrent weight matrix $\mathbf{W}$ and used them in all subsequent scalings across all tasks for comparability reasons. We plot the matrices of the individual TE values (i.e. for each pair of reservoir neurons) for the scaled reservoirs. All initialized reservoirs had flat TE matrices with small values (not shown), but the four data sets yield different TE matrices as seen in Figure 3. TE for unscaled is not shown, since in all cases it looks flat with very low

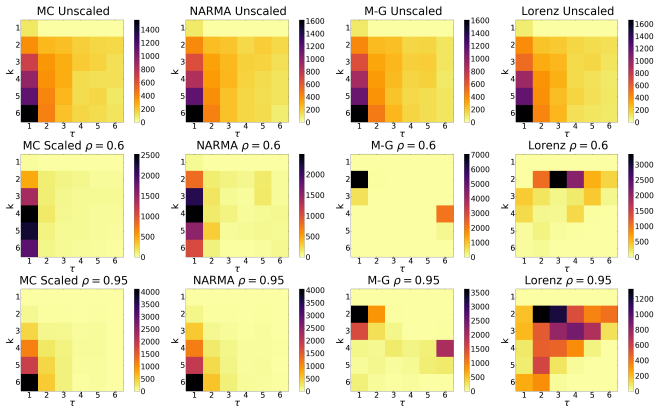

Fig. 2. Grid search of (k, τ) pairs (with k growing from the top on y-axis, τ is on x-axis starting from the left), for various scalings of the reservoir, in case of all four datasets (MC task, NARMA, M–G, Lorenz). Each cell denotes the number of occurrences when the task performance was best. We used this maximum value for determining the optimal pair.

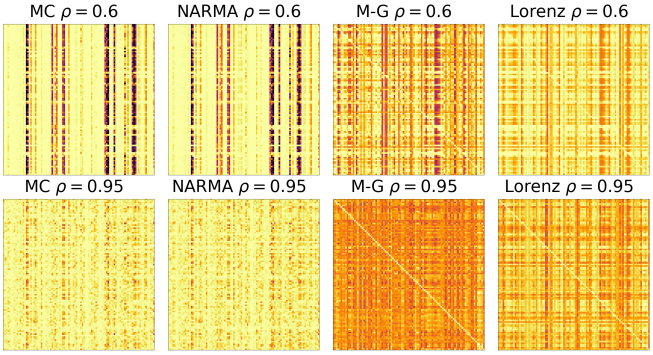

Fig. 3. The matrices ($N \times N$) of TE values in the reservoir for two scalings (spectral radii), in four datasets. Rows of the matrix denote source units and columns denote target units.

values. Scaling to $\rho = 0.6$ leads to a sparse distribution of information transfer "hubs" (fan-in), i.e. neurons with large TE values (visible as darker columns) from other sources. As opposed to "vertical" matrix structure in MC and NARMA, M–G and Lorenz matrices partially reveal also a "horizontal" structure, due to neurons with higher values of TE leading to various targets (fan-out). Scaling close to the criticality ($\rho = 0.95$) leads to different effects depending on the dataset. It either disturbs the hubs (MC, NARMA), or extends the distribution to more units (articulated more for M-G, less for Lorenz). Quantitative differences are provided in Table II.

E. Changes in information measures

Consistently with previous results, here we plot the distributions of TE resulting from different scalings of the reservoir. This is shown in Figure 4. The initial distribution of TE is the same (with small values) for all datasets, but changes arise due to scaling. For $\rho = 0.6$, the span ranges from the entire interval [0;1] (MC) to the lower half. Criticality leads to decrease of maximal values in case of MC, NARMA and M–G datasets. In case of M–G, the distribution of TE differs most noticeably from other cases.

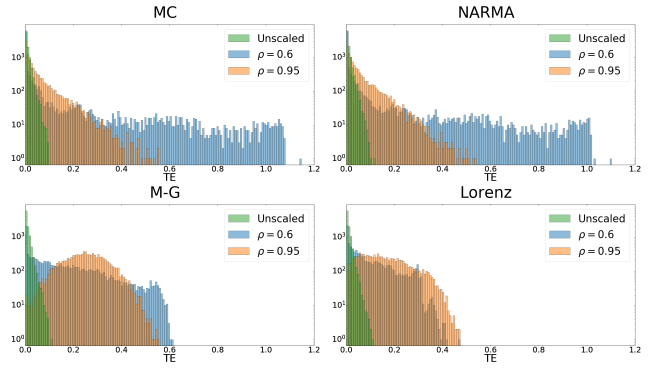

Fig. 4. Changes in TE distribution due to reservoir scaling in case of four data sets (y-axis is in the log plot). Initial TE distribution is always the same (with small values) but the differences arise in the scaled reservoirs, both in the x-range and the shape, depending on the data set.

Analogical results for AIS are plotted in Figure 5. It can be observed that AIS value have similar spread but different mean values, contrasting 0.5 (top row) versus 3 (bottom row).

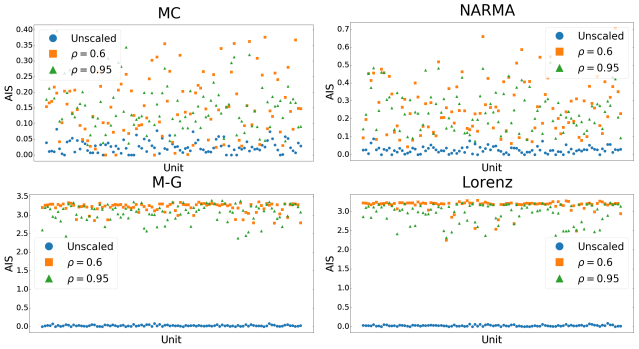

Fig. 5. AIS changes in various reservoir scalings in case of four data sets. Initial AIS distribution is always the same but evident differences arise between two pairs of data sets (top and bottom).

TABLE II. QUANTITATIVE MEASURES FOR ALL DATASETS, IN CASE OF INITIALIZED, SCALED AND ORTHOGONALIZED RESERVOIRS.

MC task	Unscaled	$\rho = 0.6$	$\rho = 0.95$	OG	ON
MC	0.06	17.8	32.8	47.4	33.3
Average TE	0.009	0.092	0.048	0.042	0.062
Average AIS	0.027	0.146	0.151	0.12	0.148
Rel. entr. of TE	3.37	1.19	2.11	2.35	1.73
LE	0.53	−0.52	−0.06	−0.09	−0.16
NARMA	Unscaled	$\rho = 0.6$	$\rho = 0.95$	OG	ON
NRMSE	2.07	0.98	0.83	0.82	0.84
Average TE	0.0093	0.088	0.053	0.044	0.067
Average AIS	0.023	0.267	0.247	0.232	0.257
Rel. entr. of TE	3.35	1.55	1.89	2.05	1.66
LE	0.52	−0.53	−0.062	−0.089	−0.16
Mackey–Glass	Unscaled	$\rho = 0.6$	$\rho = 0.95$	OG	ON
NRMSE	1.12	0.00025	0.00026	0.0003	0.00026
Average TE	0.0093	0.15	0.25	0.26	0.24
Average AIS	0.029	3.204	3.073	3.142	3.115
Rel. entr. of TE	3.36	0.81	0.99	1.23	0.88
LE	0.53	−0.52	−0.062	−0.09	−0.16
Lorenz	Unscaled	$\rho = 0.6$	$\rho = 0.95$	OG	ON
NRMSE	0.56	0.00012	0.0013	0.0034	0.00097
Average TE	0.009	0.087	0.16	0.15	0.12
Average AIS	0.025	3.157	2.971	2.932	3.002
Rel. entr. of TE	3.37	1.45	1.01	1.09	1.24
LE	0.52	−0.74	−0.28	−0.35	−0.32

This reflects the fact that reservoir units have much higher predictability of their future values in case of less complex datasets (M–G and Lorenz system).

F. Quantification of used measures

As a final step, we evaluated information-theoretic measures in case of initialized reservoirs (corresponding to an unstable regime) and two scalings (0.6 and 0.95). In addition, these values are also computed for orthogonalized reservoirs, where we used the learning rates according to [9]. Consistently with Figure 1, Table II confirms several observations. First, reservoir scaling improves performance in all cases, but for $\rho = 0.95$ the performance is not maximized in case of M–G and Lorenz datasets. Second, orthogonalization methods keep all models in the stable regime but different effects on various measures across the datasets. Third, in some cases we observe a small but statistically significant decrease of NRMSE by one of the methods, as tested by Wilcoxon signed-rank test of differences between two sample means. Concretely, OG method improves NARMA, whereas ON method improves Lorenz prediction. On the other hand, the orthogonalization

does not improve MC in our simulations, as it should, when we compare it to [9] (where both OG and ON methods led to $\mathrm{MC} > 90$). The reason is due to different unscaled initialization of the reservoir matrix: here we used $\sigma = \sqrt{0.5}$ and $\mathcal{U}(-0.1; 0.1)$ for input weights (commonly for all datasets), as opposed to $\sigma = 0.092$ found to be optimal by grid search, and $\mathcal{U}(-0.01; 0.01)$. It appears that an optimal value of σ for a concrete dataset, prior to scaling and subsequent orthogonalization is crucial for best performance, together with suitable input weights scaling (also investigated in [9]).

IV. Conclusion

Our results partially confirm those in [12], but extend them in multiple ways. We took a closer look at information measures (transfer entropy and active information storage), quantifying the information flow in ESNs with differently scaled reservoirs. The differences among the datasets arise from the complexity of the input sequences. The results conformed the previous findings that both AIS and TE increase significantly at criticality. However, our computational analyses revealed in addition that both measures actually decrease before they peak at criticality. Hence, the dependence of performance is more complex, nonmonotonous, in case of these two datasets than previously thought. Interestingly, the two additional data sets (Mackey–Glass and Lorenz system) reveal that AIS is (equally) high also outside the criticality (in a stable regime), and to a small degree also TE. At criticality, the performance is usually maximized, but it seems that (deterministic) Mackey–Glass and Lorenz systems do not benefit from criticality because they achieve equivalent performance also in the stable regime.

Results from the search for optimal embedding dimension and the time delay using a locally constant predictor also provide some insight into the complexity and dynamics of the reservoir units in case of our four datasets. In [15] it is argued that a locally constant predictor is in fact a specific Markov predictor and the embedding dimension that minimizes the locally constant prediction error of a time series can be viewed as the order of some scalar Markov process. In the light of the above, if we describe the signal of a reservoir unit as a realization of a Markov process, then the most frequent orders, as shown in Figure 2, are the same for MC and NARMA tasks, as well as for M–G and Lorenz tasks. This seems to support our findings that two complexity classes are present in our four tasks where different behaviors regarding TE, AIS, NRMSE/MC are observed in the neighborhood of the criticality. More detailed investigations are needed for better understanding of these differences, as well as the optimization of reservoir orthogonalization aimed at potential improvement of prediction performance.

Acknowledgment

This work was supported by the Slovak Grant Agency for Science (VEGA), project 1/0686/18 and by KEGA project 017UK-4/2016.

References

[1] N. Bertschinger and T. Natschläger, "Real-time computation at the edge a of chaos in recurrent neural networks," *Neural Computation*, vol. 16, no. 7, pp. 1413–1436, 2004.

[2] R. Legenstein and W. Maass, "Edge of chaos and prediction of computational performance for neural circuit models," *Neural Networks*, vol. 20, pp. 323–334, 2007.

[3] L. Büsing, B. Schrauwen, and R. Legenstein, "Connectivity, dynamics, and memory in reservoir computing with binary and analog neurons," *Neural Computation*, vol. 22, no. 5, pp. 1272–1311, 2010.

[4] D. Chialvo, "Critical brain networks," *Physica A*, vol. 340, no. 4, pp. 756–765, 2004.

[5] J. Beggs, "The criticality hypothesis: how local cortical networks might optimize information processing," *Philosophical Transactions of the Royal Society A*, vol. 366, no. 1864, pp. 329–343, 2008.

[6] J. Beggs and N. Timme, "Being critical of criticality in the brain," *Frontiers in Physiology*, vol. 3, no. 163, 2012.

[7] M. Lukoševičius, *A practical guide to applying echo state networks*, 2nd ed., ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7700, pp. 659–686.

[8] H. Jaeger, "Short term memory in echo state networks," German National Research Center for Information Technology, Tech. Rep. GMD Report 152, 2001.

[9] I. Farkaš, R. Bosák, and P. Gergeľ, "Computational analysis of memory capacity in echo state networks," *Neural Networks*, vol. 83, pp. 109–120, 2016.

[10] J. Lizier, M. Prokopenko, and A. Zomaya, "Local measures of information storage in complex distributed computation," *Information Science*, vol. 208, pp. 39–54, 2012.

[11] T. Schreiber, "Measuring information transfer," *Physics Review Letters*, vol. 85, no. 2, pp. 461–464, 2000.

[12] J. Boedecker, O. Obst, J. Lizier, N. Mayer, and M. Asada, "Information processing in echo state networks at the edge of chaos," *Theory in Biosciences*, vol. 131, pp. 205–213, 2012.

[13] M. Wibral, R. Vicente, and J. Lizier, Eds., *Directed Information Measures in Neuroscience*. Springer, 2014.

[14] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physics Reviews E: Stat. Nonlin. Soft Matter Phys.*, vol. 69, no. 6, Part 2, p. 066138, 2004.

[15] M. Ragwitz and H. Kantz, "Markov models from data by simple nonlinear time series predictors in delay embedding spaces," *Physics Reviews E: Stat. Nonlin. Soft Matter Phys.*, vol. 65, no. 5, Part 2, p. 056201, 2002.

[16] M. Lukoševičius and H. Jaeger, "Survey: Reservoir computing approaches to recurrent neural network training," *Computer Science Reviews*, vol. 3, no. 3, pp. 127–149, 2009.

[17] I. Yildiz, H. Jaeger, and S. Kiebela, "Re-visiting the echo state property," *Neural Networks*, vol. 35, pp. 1–9, 2012.

[18] G. Manjunath and H. Jaeger, "Echo state property linked to an input: Exploring a fundamental characteristic of recurrent neural networks," *Neural Computation*, vol. 25, pp. 671–696, 2013.

[19] I. Farkaš and P. Gergeľ, "Maximizing memory capacity of echo state networks with orthogonalized reservoirs," in *International Joint Conference on Neural Networks*, 2017, pp. 2437–2442.

[20] J. Sprott, *Chaos and Time-Series Analysis*. Oxford University Press, 2003.

[21] L. Glass and M. Mackey, *Scholarpedia*, vol. 5, no. 3, 2010.

[22] "Lorenz system," https://en.wikipedia.org/wiki/Lorenz_system.