

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

Miloslav Torda
Supervisor: Igor Farkaš

Artificial neural networks:

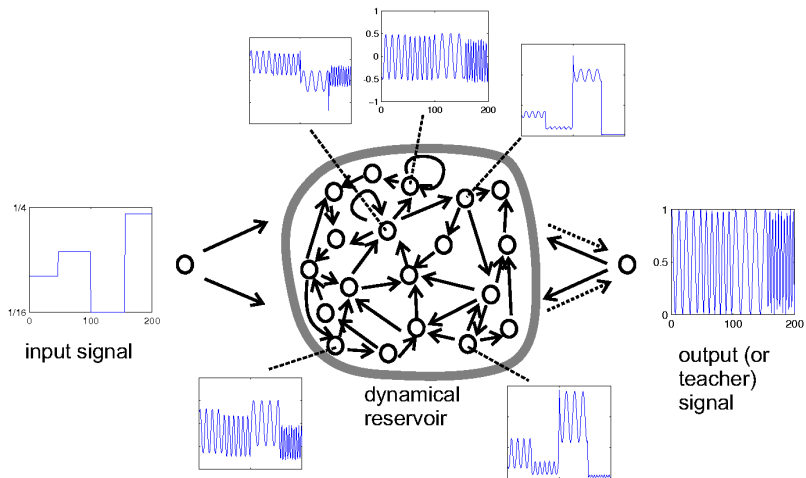
- Task performance is maximized when the network is operating in a state near the edge of chaos, or stability (Bertschinger and Natschläger, 2004)
- Information measures are maximized near the edge of stability (Boedecker et. al., 2012)

Biological neural networks:

- Cortical circuits may be tuned to criticality for optimized behavior (Beggs, 2008)

Echo state network: Model

Model



Echo state network

General model

Update equations:

$$\mathbf{x}(t) = \tanh(\mathbf{w}^{\text{in}}u(t) + \mathbf{W}\mathbf{x}(t-1))$$

$$\mathbf{x}(t) \in \mathbb{R}^{N_x}, \mathbf{w}^{\text{in}} \in \mathbb{R}^{N_x}, \mathbf{W} \in \mathbb{R}^{N_x \times N_x}$$

Output layer:

$$\mathbf{y}(t) = \mathbf{W}^{\text{out}}\mathbf{x}(t)$$

$$\mathbf{y}(t) \in \mathbb{R}^{N_y}, \mathbf{W}^{\text{out}} \in \mathbb{R}^{N_y \times N_x}$$

Learning:

$$\mathbf{W}^{\text{out}} = \mathbf{U}\mathbf{X}^+$$

$$\mathbf{X}^+ = \mathbf{X}^{\text{T}}(\mathbf{X}\mathbf{X}^{\text{T}})^{-1}$$

Information measures

Transfer entropy:

$$\begin{aligned} TE_{Y \rightarrow X}^{(k,l)} &= I(X_t : \mathbf{Y}_{t-u}^{(l)} | \mathbf{X}_{t-1}^{(k)}) \\ &= \sum_{\mathbf{x}_{t-1}^{(k)}, \mathbf{y}_{t-u}^{(l)}} p(\mathbf{x}_{t-1}^{(k)}, \mathbf{y}_{t-u}^{(l)}) \sum_{x_t} p(x_t | \mathbf{x}_{t-1}^{(k)}, \mathbf{y}_{t-u}^{(l)}) \log_2 \frac{p(x_t | \mathbf{x}_{t-1}^{(k)}, \mathbf{y}_{t-u}^{(l)})}{p(x_t | \mathbf{x}_{t-1}^{(k)})} \end{aligned}$$

Active information storage:

$$\begin{aligned} AIS_X^{(k)} &= I(X_t : \mathbf{X}_{t-1}^{(k)}) \\ &= \sum_{x_t, \mathbf{x}_{t-1}^{(k)}} p(x_t, \mathbf{x}_{t-1}^{(k)}) \log_2 \frac{p(x_t, \mathbf{x}_{t-1}^{(k)})}{p(x_t) p(\mathbf{x}_{t-1}^{(k)})} \end{aligned}$$

where

$$\mathbf{X}_{t-1}^{(k)} = (X_{t-1}, X_{t-1-\tau_k}, X_{t-1-2\tau_k}, \dots, X_{t-1-(k-1)\tau_k})$$

$$\mathbf{Y}_{t-u}^{(l)} = (Y_{t-u}, Y_{t-u-\tau_l}, Y_{t-u-2\tau_l}, \dots, Y_{t-u-(l-1)\tau_l})$$

Transfer entropy:

$$\widehat{\text{TE}}_{Y \rightarrow X}^{(k,l)} = \Psi(K) + \langle \Psi(n_{\mathbf{x}_{t-1}^{(k)}} + 1) - \Psi(n_{\mathbf{x}_t, \mathbf{x}_{t-1}^{(k)}} + 1) - \Psi(n_{\mathbf{x}_{t-1}^{(k)}, \mathbf{y}_{t-1}^{(l)}} + 1) \rangle_t$$

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 $(\mathbf{x}_t, \mathbf{x}_{t-1}^{(k)}, \mathbf{y}_{t-1}^{(l)})$ at time step t to its K th nearest neighbor in the joint space, and $\langle \cdot \rangle_t$ denotes the time-average.

Active information storage:

$$\widehat{\text{AIS}}_X^{(k)} = \Psi(K) + \Psi(N) - \langle \Psi(n_{\mathbf{x}_{t-1}^{(k)}} + 1) + \Psi(n_{\mathbf{x}_t} + 1) \rangle_t$$

Estimating criticality

Maximum Lyapunov characteristic exponent

Let $\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t))$ be a discrete time dynamical system and $\mathbf{x}'(0) = \mathbf{x}(0) + \delta\mathbf{x}(0)$ be the initial conditions a trajectory $\mathbf{x}'(t)$ obtained by an infinitesimal displacement from $\mathbf{x}(0)$ such that $\gamma_0 = \|\delta\mathbf{x}(0)\| \ll 1$. Then the maximum Lyapunov characteristic exponent is

$$\lambda = \lim_{t \rightarrow \infty} \frac{1}{t} \ln \left(\frac{\gamma_t}{\gamma_0} \right) < \infty$$

where $\gamma_t = \|\mathbf{x}'(t) - \mathbf{x}(t)\|$.

$$\gamma_t \sim \gamma_0 e^{\lambda t} \Rightarrow \begin{cases} \lambda > 0 & \text{sensitive to initial conditions} \\ \lambda \approx 0 & \text{edge of criticality} \\ \lambda < 0 & \text{sub-critical} \end{cases}$$

Experimental setup

ESN model

- $N=100$ reservoir units
- single input unit
- output units
 - 120 for the MC task
 - single unit for the prediction task
- $w_j^{in} \in \mathcal{U}(-0.1; 0.1), \forall j = 1, \dots, N$
- $w_{ij}^{res} \in \mathcal{N}(0; 0.5), \forall i, j = 1, \dots, N$
- 100 samples transients
- 1000 samples training
- 2000 samples testing

Experimental setup

Performance measures

Prediction tasks:

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

MC task:

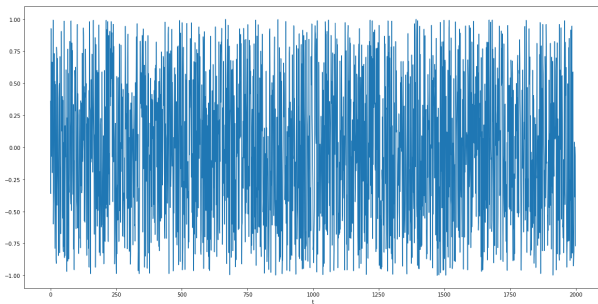
$$\text{MC} = \sum_{q=1}^{q_{\max}} \text{MC}_q = \sum_{q=1}^{q_{\max}} \frac{\text{cov}^2(u(t-q), y_q(t))}{\text{var}(u(t)) \cdot \text{var}(y_q(t))}$$

Benchmark tasks

MC

$$\forall t : u(t) \sim \mathcal{U}(-1; 1)$$

Uniform



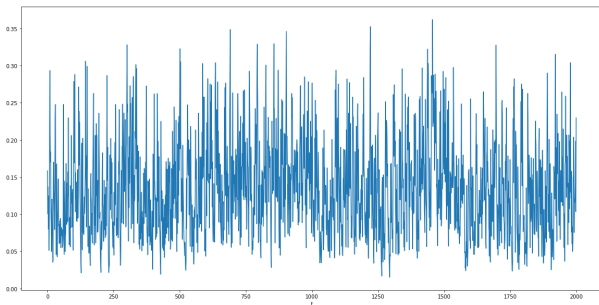
Benchmark tasks

NARMA

$$u(t+1) = 0.2 u(t) + 0.004 u(t) \sum_{i=0}^{29} u(t-i) + 1.5 q(t-29)q(t) + 0.001$$

where $\forall t : q(t) \sim \mathcal{U}(0; 0.5)$

NARMA



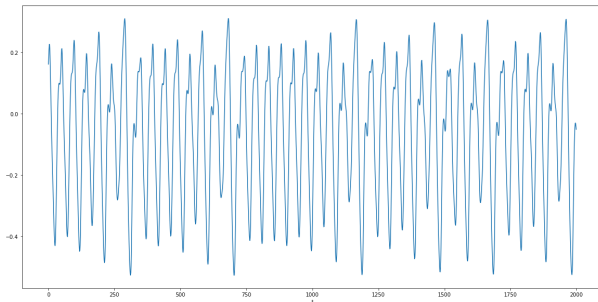
Benchmark tasks

Mackey-Glass system

$$\frac{dy}{dt} = 0.2 \frac{y_{17}}{1 + y_{17}^{10}} - 0.1y$$

where u_{17} is the value of u at time $t - 17$.

Mackey-Glass



Benchmark tasks

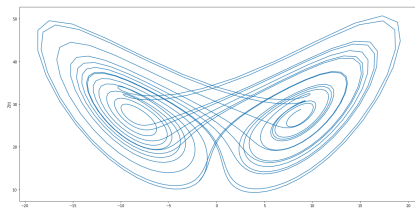
Lorenz system

$$\frac{dx}{dt} = 10(y - x)$$

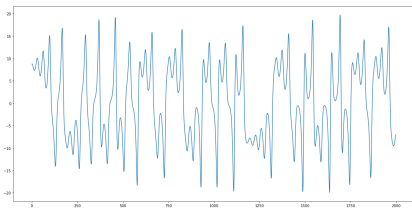
$$\frac{dy}{dt} = x(28 - z) - y$$

$$\frac{dz}{dt} = xy - \frac{8}{3}z$$

Lorenz attractor



Lorenz attractor



ESN:

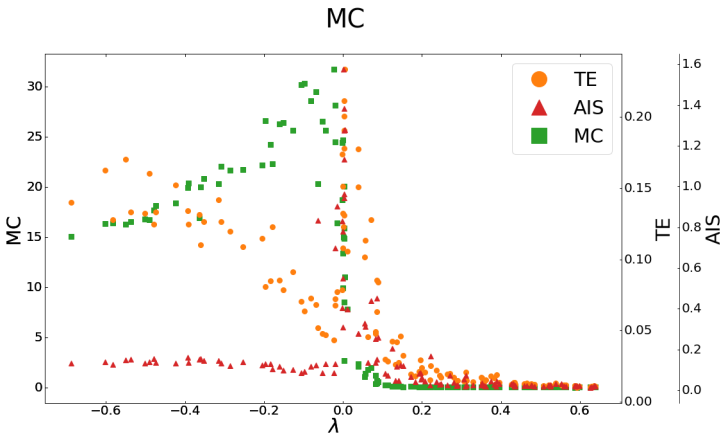
- fixed \mathbf{w}^{in}
- $w_{ij}^{\text{res}} \in \mathcal{N}(0; \sigma)$ such that $\log \sigma \in [-1.5; -0.25]$ in 26 steps 5 instances per value σ

Information measures:

- TE
 - Target - $k = 2, \tau_k = 1$
 - Source - $l = 1, \tau_l = 1$
- AIS - $k = 2, \tau_k = 1$
- 4-NN

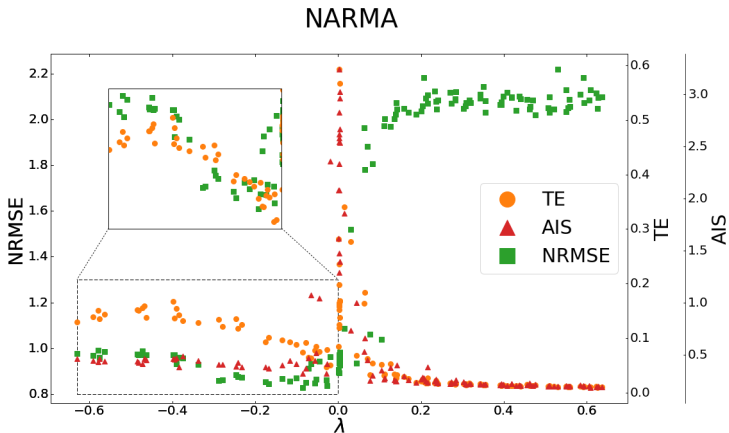
Information-theoretical measures around criticality

MC task



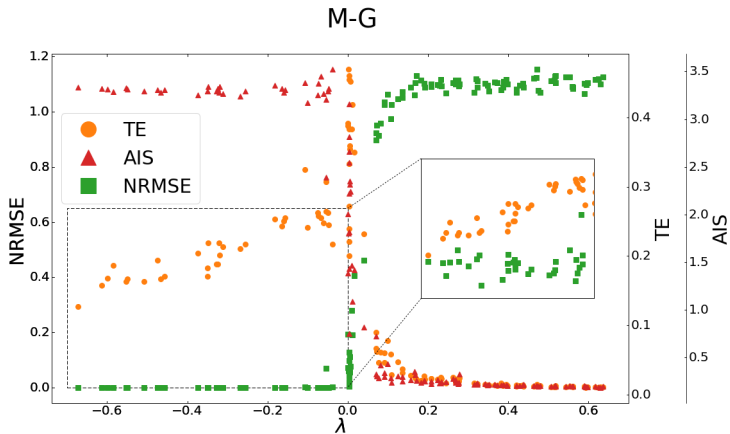
Information-theoretical measures around criticality

NARMA task



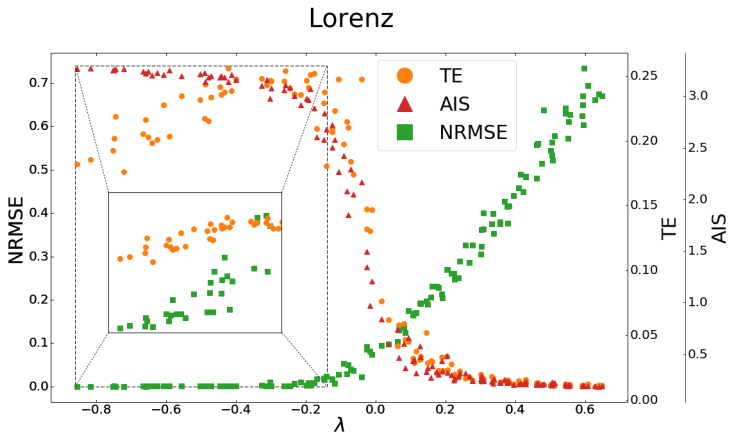
Information-theoretical measures around criticality

M-G task



Information-theoretical measures around criticality

Lorenz task



A closer look at information measures

Reservoir neuron signal embedding

Ragwitz-Kantz criterion:

$$(k, \tau) = \operatorname{argmin}_{k \in \mathcal{Z}, \tau \in \mathcal{Z}} \|\mathbf{x} - \hat{\mathbf{x}}(k, \tau)\|$$

where

$$\mathbf{x} = (x_1, \dots, x_n), \hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_n)$$

Locally constant predictor:

$$\hat{x}_{t+1} = \frac{1}{\operatorname{card}(\mathcal{U}_t)} \sum_{\mathbf{x}_l \in \mathcal{U}_t} x_{l+1}$$

$$\mathcal{U}_t = \{\mathbf{x}_l : \|\mathbf{x}_l - \mathbf{x}_t\| \leq \epsilon\}, \mathbf{x}_t = (x_t, x_{t-\tau_k}, x_{t-2\tau_k}, \dots, x_{t-(k-1)\tau_k})$$

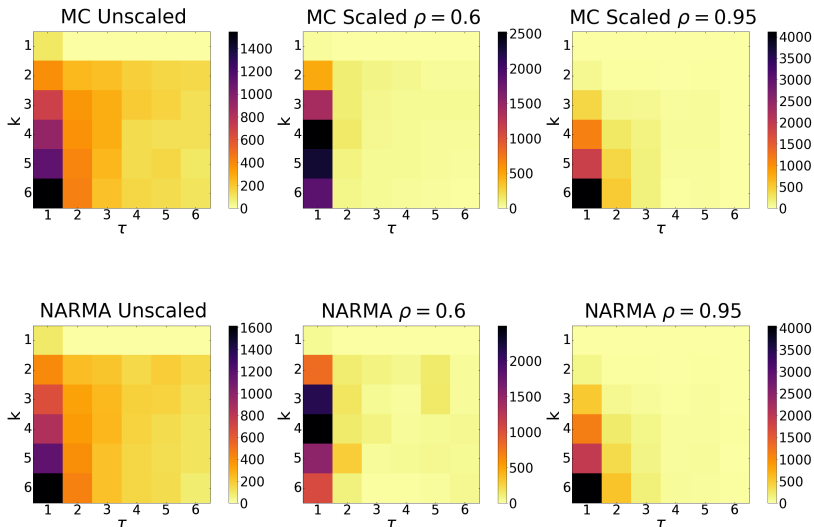
A closer look at information measures

Reservoir neuron signal embedding: Experimental setup

- Ragwitz-Kantz search constrains:
 - $k = 1, \dots, 6$
 - $\tau = 1, \dots, 6$
- Reservoir spectral radius scalings:
 - Unstable: Unscaled
 - Stable: $\rho = 0.6$
 - Close to criticality: $\rho = 0.95$
- 100 instances per scaling and task

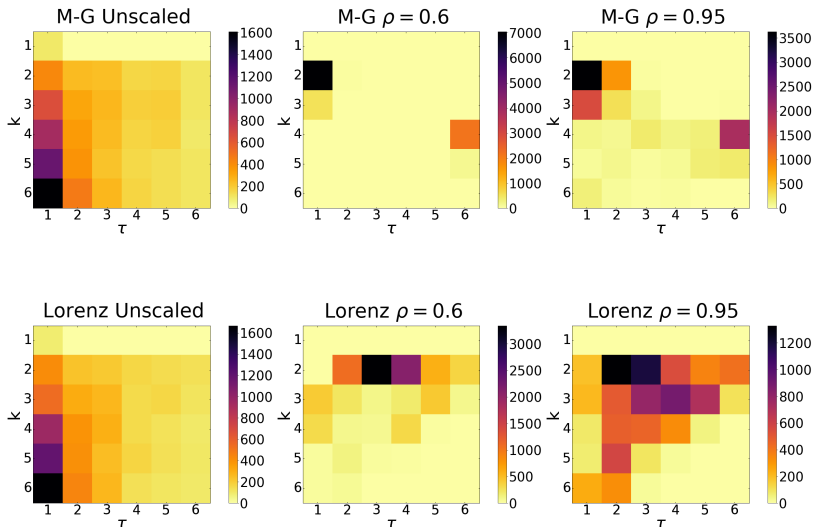
A closer look at information measures

Reservoir neuron signal embedding: MC and NARMA tasks



A closer look at information measures

Reservoir neuron signal embedding: M-G and Lorenz tasks



A closer look at information measures

Reservoir neuron signal embedding

Most frequent (k, τ) pairs for each task and scaling

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

A closer look at information measures

Experimental setup

ESN:

- 1 instance for all tasks and scalings
 - $w_j^{in} \in \mathcal{U}(-0.1; 0.1), \forall j = 1, \dots, N$
 - $w_{ij}^{res} \in \mathcal{N}(0; 0.5), \forall i, j = 1, \dots, N$
- Reservoir spectral radius scalings
 - Unstable: Unscaled
 - Stable: $\rho = 0.6$
 - Close to criticality: $\rho = 0.95$

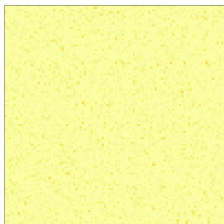
Information measures:

- TE
 - Target - determined by Ragwitz-Kantz criterion
 - Source - $l = 1, \tau_l = 1$
- AIS - determined by Ragwitz-Kantz criterion
- 4-NN

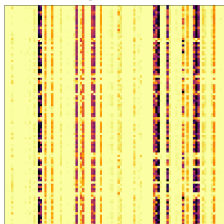
A closer look at information measures

TE matrices: MC and NARMA tasks

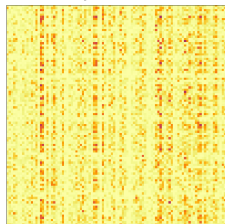
MC Unscaled



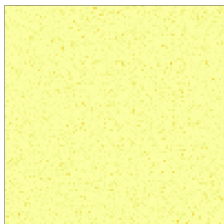
MC $\rho = 0.6$



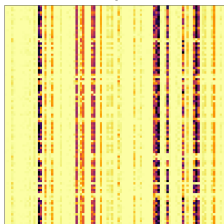
MC $\rho = 0.95$



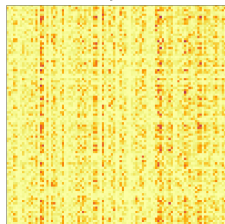
NARMA Unscaled



NARMA $\rho = 0.6$



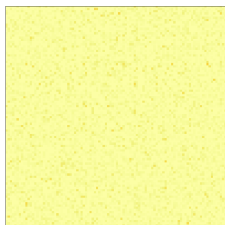
NARMA $\rho = 0.95$



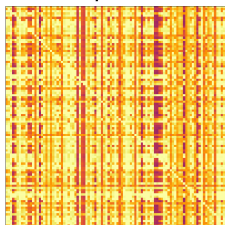
A closer look at information measures

TE matrices: M-G and Lorenz tasks

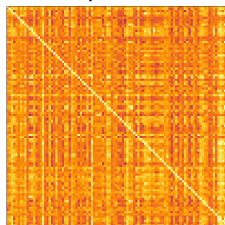
M-G Unscaled



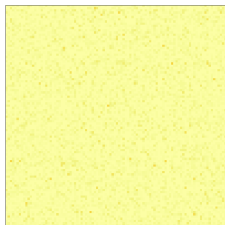
M-G $\rho = 0.6$



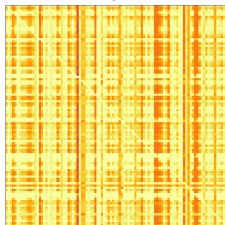
M-G $\rho = 0.95$



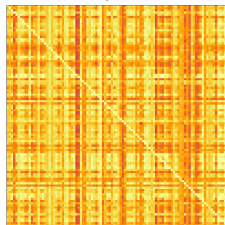
Lorenz Unscaled



Lorenz $\rho = 0.6$



Lorenz $\rho = 0.95$



A closer look at information measures

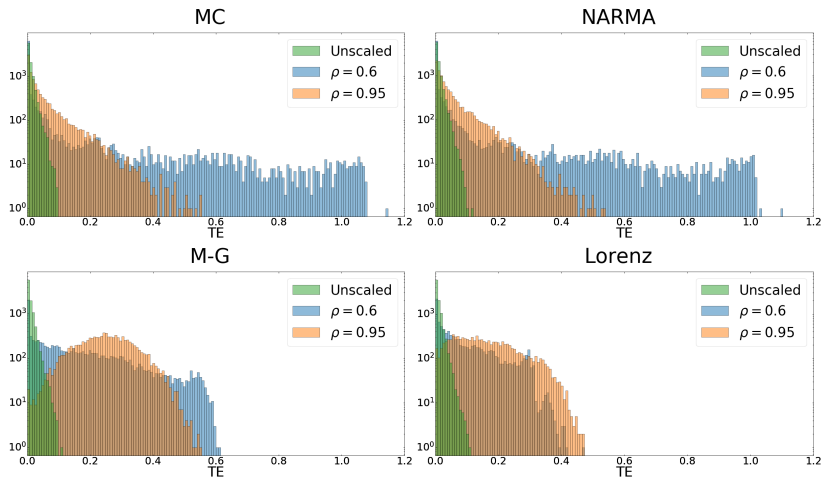
Relative entropy of TE

$$\begin{aligned} \left| \overline{H}(\text{TE}_{\text{RES}}^{(k,l)}) \right| &= \widehat{D}_{\text{KL}}(\text{TE}_{\text{RES}}^{(k,l)} \parallel \mathcal{U}(0; \max \text{TE}_{\text{RES}}^{(k,l)})) \\ &= \ln(\max \text{TE}_{\text{RES}}^{(k,l)}) - \widehat{H}_{\text{KL}}(\text{TE}_{\text{RES}}^{(k,l)}) \end{aligned}$$

where $\ln(\max \text{TE}_{\text{RES}}^{(k,l)})$ is the exact differential entropy of the uniform distribution with the support in $[0; \max \text{TE}_{\text{RES}}^{(k,l)}]$, $\widehat{H}_{\text{KL}}(\cdot)$ is the Kozachenko-Leonenko differential entropy estimator and $\text{TE}_{\text{RES}}^{(k,l)}$ is the distribution of the estimates of transfer entropies in the reservoir.

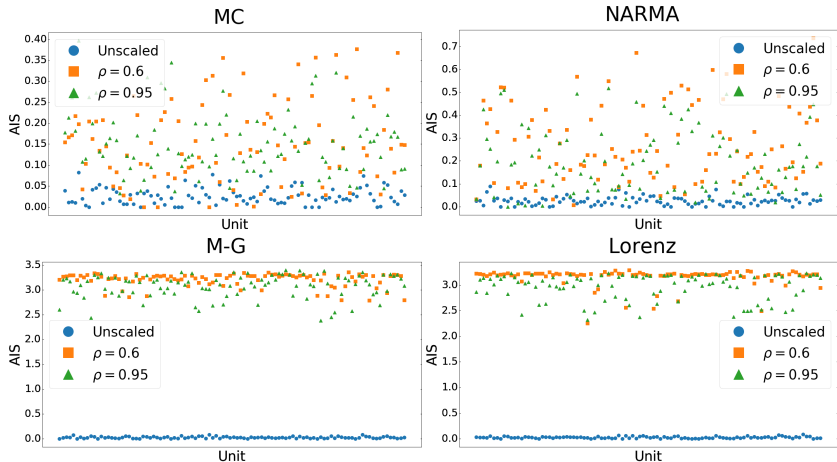
A closer look at information measures

TE distribution



A closer look at information measures

Active Information Storage



A closer look at information measures

Quantification of used measures

MC task	Unscaled	$\rho = 0.6$	$\rho = 0.95$	NARMA	Unscaled	$\rho = 0.6$	$\rho = 0.95$
MC	0.06	17.8	32.8	NRMSE	2.07	0.98	0.83
Average TE	0.009	0.092	0.048	Average TE	0.0093	0.088	0.053
Average AIS	0.027	0.146	0.151	Average AIS	0.025	0.269	0.191
Rel. entr. of TE	1.010	1.323	1.354	Rel. entr. of TE	1.175	1.642	1.271
LE	0.53	-0.52	-0.06	LE	0.52	-0.53	-0.062

Mackey-Glass	Unscaled	$\rho = 0.6$	$\rho = 0.95$	Lorenz	Unscaled	$\rho = 0.6$	$\rho = 0.95$
NRMSE	1.12	0.00025	0.00026	NRMSE	0.56	0.00012	0.0013
Average TE	0.0093	0.15	0.25	Average TE	0.009	0.087	0.16
Average AIS	0.025	3.204	3.073	Average AIS	0.025	3.157	2.971
Rel. entr. of TE	1.166	0.315	0.391	Rel. entr. of TE	1.148	0.694	0.254
LE	0.53	-0.52	-0.062	LE	0.52	-0.74	-0.28

Conclusion

- Two complexity classes in tested tasks
 - stochastic: MC/NARMA - benefit from criticality
 - deterministic: M-G/Lorenz - do not benefit from criticality
- Information measures are maximized at phase transition from stable to unstable regime
- Between complexity classes: increase in performance is associated with increase in global TE and AIS, and decrease of relative entropy of TE
- Within tasks: in stable regions there seems to be an inverse relationship between performance and global TE.
- Higher information flow does not mean better performance.
- Task specific

THANK YOU