Exploring the Impact of Pseudospectra on the Stability of Echo State Networks

Sebastián Basterrech

DTU-Compute, Technical University of Denmark
Kongens Lyngby, Denmark

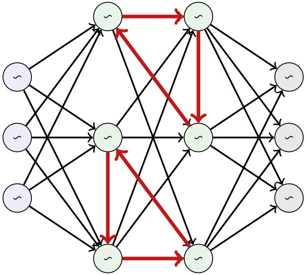
How to Apply ML to Experimental and Theoretical Physics
HAMLET-Physics'2024

Recurrent Neural Networks

A *Recurrent Neural Network (RNN)*:

- Concept: a RNN is a dynamical system (fundamental difference with FNNs).
- Discrete-time state-space model:
  - Implicit memorization.
  - High dimensional distributed representation.
- *Nice* properties:
  - Universal approximation property.
  - Biological plausibility.

Graph with circuits:

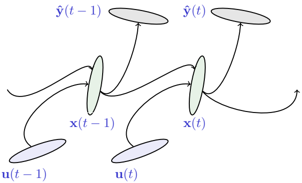


Recurrent Neural Networks

Formalization:

- A generic RNN with an input $\mathbf{u}(t)$ and state $\mathbf{x}(t)$ for time step t is given by:

$$\mathbf{x}(t) = T(\mathbf{x}(t-1), \mathbf{u}(t), \boldsymbol{\theta}). \qquad (1)$$



- For instance:

$$\mathbf{x}(t) = \mathbf{W}^{\mathrm{h}} \tanh(\mathbf{x}(t-1)) + \mathbf{W}^{\mathrm{in}}\mathbf{u}(t), \qquad (2)$$

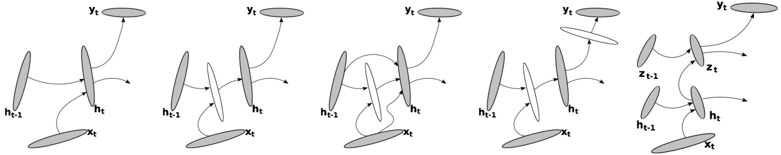
with $\mathbf{W}^{\mathrm{h}}$: weights with a recurrent structure, $\mathbf{W}^{\mathrm{in}}$: weights among inputs and recurrences.

Figure taken from Pascanu *et al.*, How to construct Deep RNNs, ICLR, 2014.

R. Pascanu, T. Mikolov, Y. Bengio, "On the difficulty of training recurrent neural networks," ICML, 2013.

RNNs

Examples of RNN topologies



(a) RNN (b) DT-RNN (b*) DT(S)-RNN (c) DOT-RNN (d) Stacked RNN

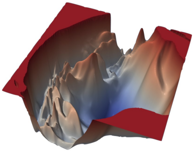
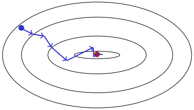
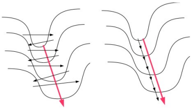
Figure 2: Illustrations of four different recurrent neural networks (RNN). (a) A conventional RNN. (b) Deep Transition (DT) RNN. (b*) DT-RNN with shortcut connections (c) Deep Transition, Deep Output (DOT) RNN. (d) Stacked RNN

Visualization of several RNN types. Figure taken from Pascanu *et al.*, How to construct Deep RNNs, 2014.

Obstacles of learning RNNs

Vanishing-exploding gradient

- Vanishing gradient problem (gradient norm tends fast towards zero).
- Exploding gradient problem (gradient norm tends to get very large).

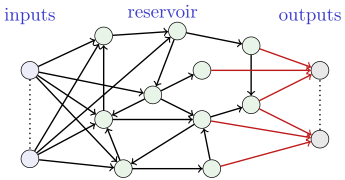
- Now exist "good" alternatives to train RNNs: truncated BPTT, Hessian-Free, clipping gradients.

Figure taken from: Hao Li *et al.*, "Visualizing the Loss Landscape of Neural Nets", NIPS, 2018.

Echo State Networks

Motivations

- Attempt to develop a model that uses the potential for *memorization* of RNNs without the difficulties in the training process.
- Reservoir Computing (2000s till now): specific design and training, only the effort is made over the memoryless structure (Herbert Jaeger).
- **How to initialize the network and where to train?**



An untrained recurrent part called reservoir.
A memory-less supervised learning tool called readout.

Herbert Jaegger and Harald Hass, "Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication", Science, 2004.

ESN

Other examples of learning simplification

Simplification of the training process in other NNs:

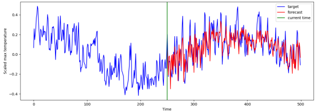
- Extreme Learning Machines: specific FFNN design, only a subset of the weights are trained.
- Shared weights principle: significant in CNNs, common strategy in deep learning, kernels with same parameters, translation invariance, generalization, dimensionality reduction (it is not biologically plausible).

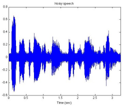
Jordan Ott *et al.* Learning in the machine: To share or not to share?, *Neural Networks*, 2020.

Guang Huang *et al.* Extreme learning machines: theory and applications, *Neurocomputing*, 2006.

ESN

Where/how to use it?





$\psi(\cdot)$ $\longrightarrow$ "En un lugar de la Mancha, de cuyo nombre no quiero acordarme"

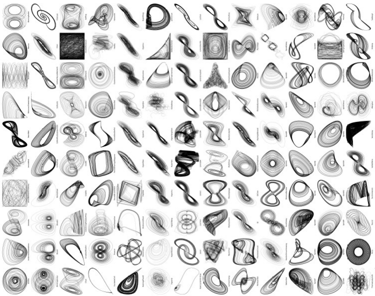
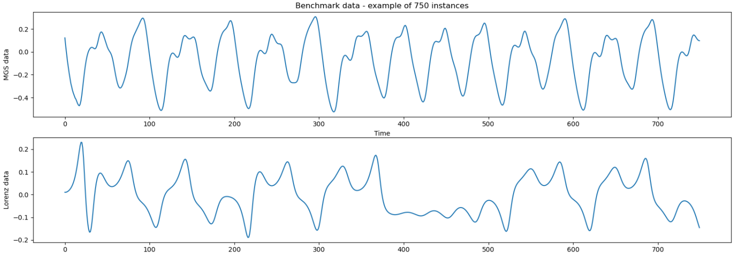


Figure taken from: William Gilpin, "Chaos as an interpretable benchmark for forecasting and data-driven modelling", NIPS'2021.

ESN

Applications in physics

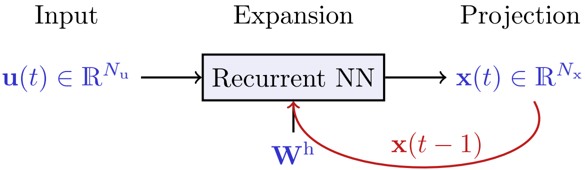


- Lorenz system: dynamics of thermo-rotation, thermohaline ocean flow circulation.
- Mackey-Glass (MGS) system: dynamical system used for analyzing bifurcations in physiological applications.

ESN

Definitions

- Expansion projection (memorization)



$$(N_\mathrm{u} \ll N_\mathrm{x})$$

- The recurrent part of the network at any time t is given by:

$$\mathbf{x}(t) = \tanh\left(\mathbf{W}^\mathrm{in}\mathbf{u}(t) + \mathbf{W}^\mathrm{h}\mathbf{x}(t-1)\right). \tag{3}$$

ESN

Conceptual separation

- The expansion can enhance the linear separability of the input information.
- The memory-less structure is fast and robust (for instance: linear model).
- Readout structure:

Projection   Training   Output

$$\mathbf{x}(t) \in \mathbb{R}^{N_x} \longrightarrow \boxed{\text{Supervised learning}} \longrightarrow \hat{\mathbf{y}}(t) \in \mathbb{R}^{N_y}$$

$$\uparrow$$

$$\mathbf{W}^{\text{out}}$$

- Only the $\mathbf{W}^{\text{out}}$ weights are updated in the training process (gradient descent methods, ridge regression, quasi-Newton, etc.

Properties

Main global parameters

Main adjustable parameters of an ESN model:

- Number of neurons: influences in the accuracy (larger networks increases the accuracy).

less accuracy more accuracy

$$- \xleftrightarrow[\text{size}]{} +$$

- Spectral radius of the weight matrix is a parameter of *memorization*.

less memory more memory

$$0 \xleftrightarrow[\rho(\mathbf{W}^{\mathrm{h}})]{} 1$$

- Activation function: $\tanh(\cdot)$, leaky ratem linear function, Leaky Integrate and Fire (LIF), Radial Basis Function (RBF), etc.
- Other: initialization (Random Matrices Theory) and connectivity design (GT).

Stability of the system

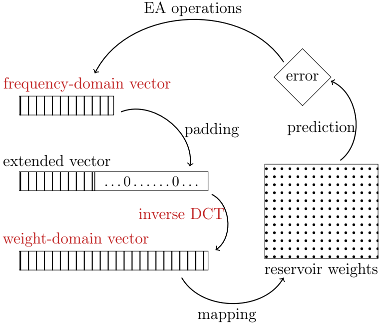
Well identified properties:

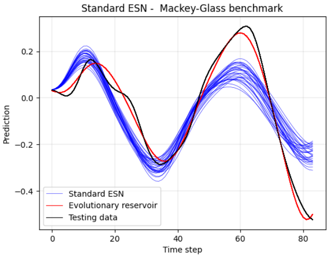
- Stability of $\mathbf{x}(t)$ only depends on the weight matrices (Echo State Property (ESP)). (Independent of $X(0)$).
- The stability behaviour:
  - *Sufficient condition* (max singular value): if the $\sigma(\mathbf{W}^{\mathrm{h}}) < 1$, then the ESP is satisfied.
  - *Necessary condition*: it is necessary that $\rho(\mathbf{W}^{\mathrm{h}}) \leq 1$ in order of holding the ESP.
- Theoretical gap between necessary and sufficient conditions of the Echo State Property.

S. Basterrech. Empirical Analysis of the Necessary and Sufficient Conditions of the Echo State Property. IJCNN'2017.

Recent advances - Neuroevolution

ESN + Fourier compression + Evolutionary Algorithms: EvoESN



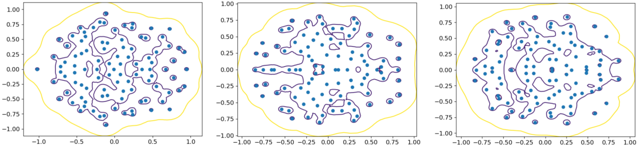


Drawback: control of spectral radius at each iteration of the evolutionary algorithm.

S. Basterrech and G. Rubino. Evolutionary Echo State Network: a neuroevolutionary framework for time series prediction, *Applied Soft-Computing*, Vol. 144, 2023.

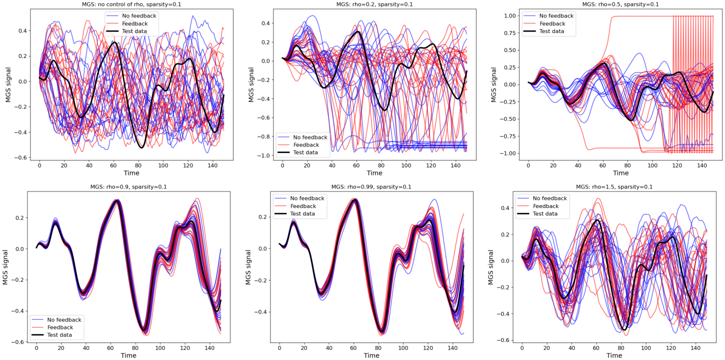
Pseudospectra

- Eigenvalues: $\Lambda(\mathbf{W}^{\mathrm{h}}) = \{z \in \mathbb{C} : det(zI - \mathbf{W}^{\mathrm{h}}) = 0\}$ (points where $(zI - \mathbf{W}^{\mathrm{h}})^{-1}$ is undefined).
- Pseudospectra: $\Lambda_\epsilon(\mathbf{W}^{\mathrm{h}}) = \{z \in \mathbb{C} : \|(zI - \mathbf{W}^{\mathrm{h}})^{-1}\| \geq \epsilon^{-1}\}$.

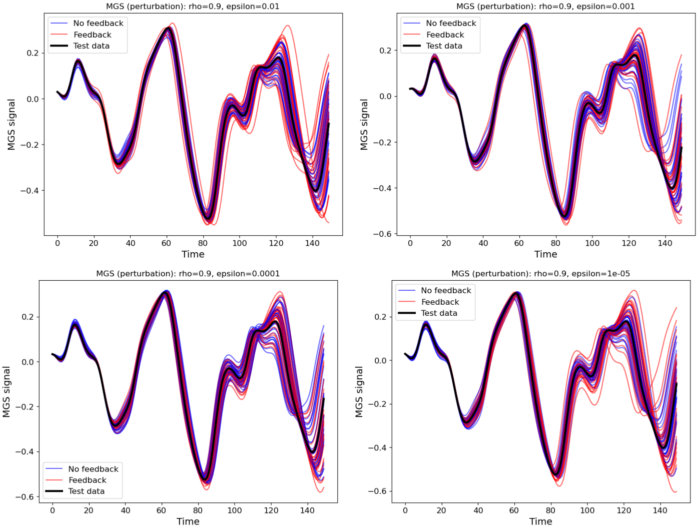


- Pseudospectra, alternative definition based in perturbed matrices:
  $\Lambda_\epsilon(\mathbf{W}^{\mathrm{h}}) = \{z \in \mathbb{C} : z \in \Lambda(\mathbf{W}^{\mathrm{h}} + E), E \text{ with } \|E\| \leq \epsilon\}$.

Stability of ESN - MGS example



Pseudospectra based in random perturbations



Stability of ESNs

(on-going research)

- Experimental analysis of the impact of pseudospectra and spectra in ESNs.
- Focus on ITUC (a region where there are missing theoretical results about stability behavior).
- Advantages of analyzing reservoir perturbations:
  - Control the evolutionary steps in EvoESN.
  - Pseudospectra may provide bounds for $f(\mathbf{W}^{\mathrm{h}})$ (recurrent function of the reservoir state).
  - ESN behaves "optimally" at the edge of chaos; pseudospectra may help in analyzing the edge of chaos.
- Matrix perturbation may also be helpful as a regularization factor.

Closing.....

Thank you!

Contacts:

- e-mail: sebbas@dtudk
- ORCID Id: C-9072-2014