Optimization of ESN reservoirs

Intrinsic Plasticity and Hebbian approaches

Daniele Di Sarli

Computational Neuroscience, Università di Pisa

Initialization of reservoirs

Reservoir state at step n:

$$\mathbf{x}(n) = f(\mathbf{W}_{in}\mathbf{u}(n) + \mathbf{W}_{res}\mathbf{x}(n-1))$$

Can't just "make $\mathbf{W}_{res}$ random" or we may get unstable behavior.

Echo State Property requires $\rho(\mathbf{W}_{res}) < 1$ (i.e. $\mathbf{W}_{res}$ is *contractive*).

- task requires long memory? Set $\rho(\mathbf{W}_{res})$ close to 1
- task requires less memory? Set smaller $\rho(\mathbf{W}_{res})$

In practice: follow experience and brute-force searches.

Can we do better?

Intrinsic Plasticity

Biological neurons can alter their **intrinsic excitability** by modifying their voltage gated channels.

- Allows them to approach an **exponential** firing rate distribution
- Exponential distribution $\implies$ Information maximization

In our case,

$$y = f(x), \quad x \text{ total neuron input, } f \text{ nonlinear}$$

- x has distribution $g_x(x)$
- y has distribution $g_y(y)$, which we want to bring close to $g_{exp}(y) = \frac{1}{\mu} \exp(\frac{-y}{\mu})$

Intrinsic Plasticity Rules

We can derive a **gradient descent** rule for minimizing **KL-divergence** between $g_y(y)$ and $g_{target}(y)$.

Parametrize the activation functions by adding a **gain** a and a **bias** b:

- $y = f(x) = \text{logistic}(ax + b)$
- $y = f(x) = \tanh(ax + b)$

Logistic: (exponential target)

- $\Delta b = \eta \left(1 - \left(2 + \frac{1}{\mu}\right) y + \frac{y^2}{\mu}\right)$
- $\Delta a = \frac{\eta}{a} + \Delta b x$

Tanh: (Gaussian target)

- $\Delta b = -\eta \left(-\frac{\mu}{\sigma^2} + \frac{y}{\sigma^2}(2\sigma^2 + 1 - y^2 + \mu y)\right)$
- $\Delta a = \frac{\eta}{a} + \Delta b x$

Intrinsic Plasticity Rules (2)

- $\Delta b = \eta \left(1 - \left(2 + \frac{1}{\mu}\right) y + \frac{y^2}{\mu}\right)$
- $\Delta a = \frac{\eta}{a} + \Delta b x$

These rules are **strictly local**.

A few **requirements** must be satisfied:

- f monotonically increasing
- f differentiable w.r.t. y
- partial derivatives of $\log \frac{\partial y}{\partial x}$ w.r.t the parameters of f must exist.

Intrinsic Plasticity Rules (3)

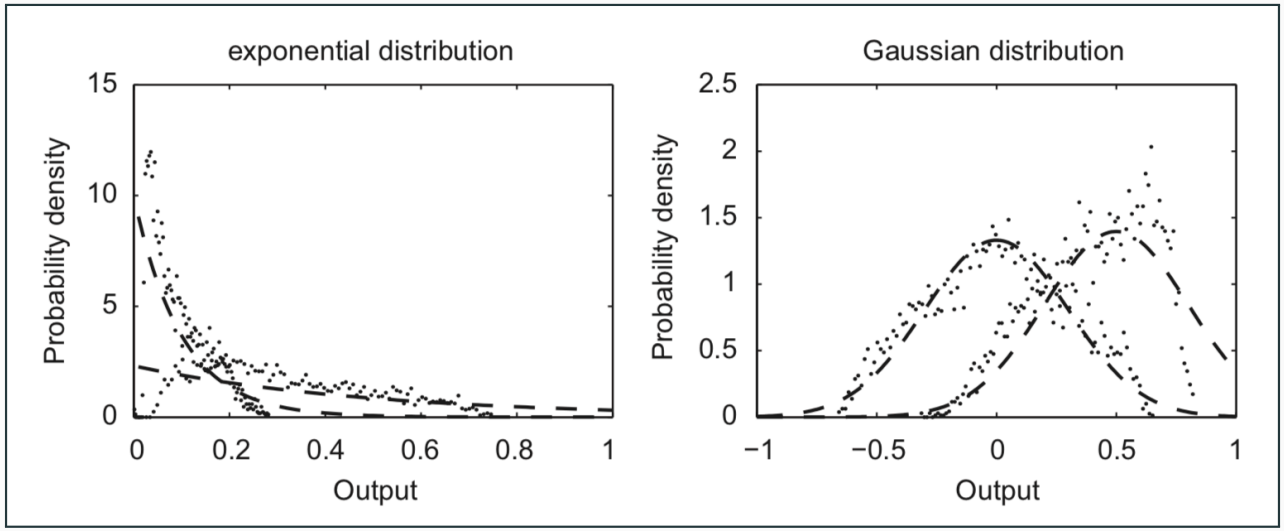


Figure 1: IP rules are able to approximate the desired distribution

It is interesting to see the IP rules under a different perspective:

$$\mathbf{x}(n) = f\left(\hat{\mathbf{W}}_{in}\mathbf{u}(n) + \hat{\mathbf{W}}_{res}\mathbf{x}(n-1) + \mathbf{b}\right)$$

$$\hat{\mathbf{W}}_{in} = \mathrm{diag}(\mathbf{a})\mathbf{W}_{in}$$

$$\hat{\mathbf{W}}_{res} = \mathrm{diag}(\mathbf{a})\mathbf{W}_{res}$$

Hebbian learning

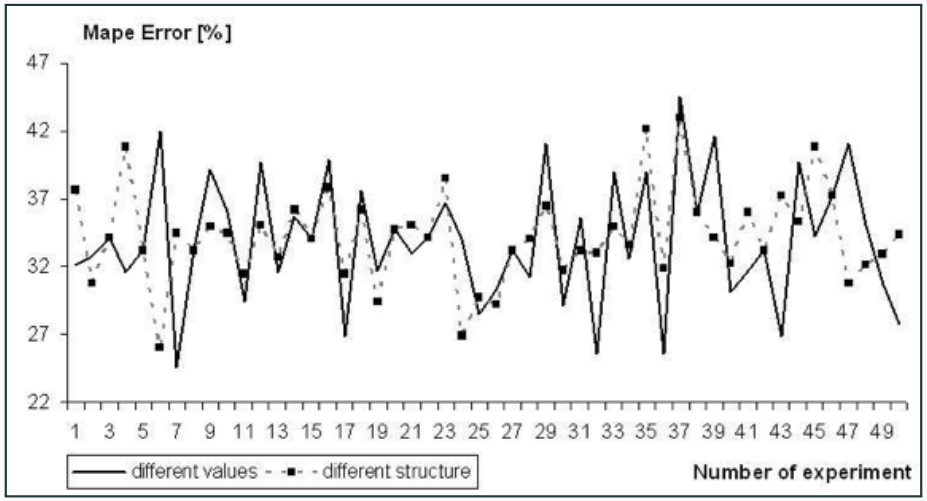


Figure 2: Reservoirs are sensitive to weight values and structure.

Biological inspiration suggests... **Hebbian learning**.

Hebb rule

$$\Delta W_{kj} = \eta\, y_k\, x_j$$

anti-Hebb rule

$$\Delta W_{kj} = -\eta\, y_k\, x_j$$

- x_j: output of presynaptic neuron j
- y_k: output of postsynaptic neuron k

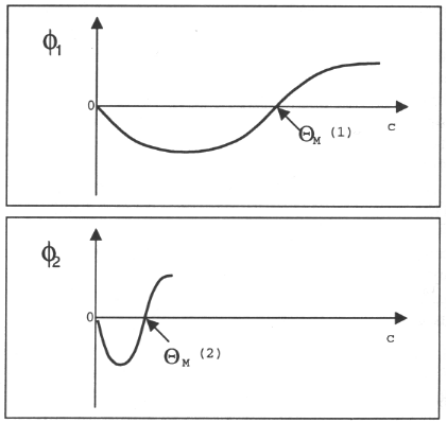
Hebbian-based rules

BCM (Law and Cooper, 1994):

$$\Delta W_{kj}(t) = y_k \left(y_k - \theta_M\right) \frac{x_j}{\theta_M}$$

$$\underbrace{\theta_M = \mathrm{E}[y_k^2]}$$

Sliding threshold



Anti-Oja:

$$\Delta W_{kj}(t) = -\eta \left(y_k x_j - \underbrace{y_k^2 W_{kj}}\right)$$

Forgetting term

Convergence

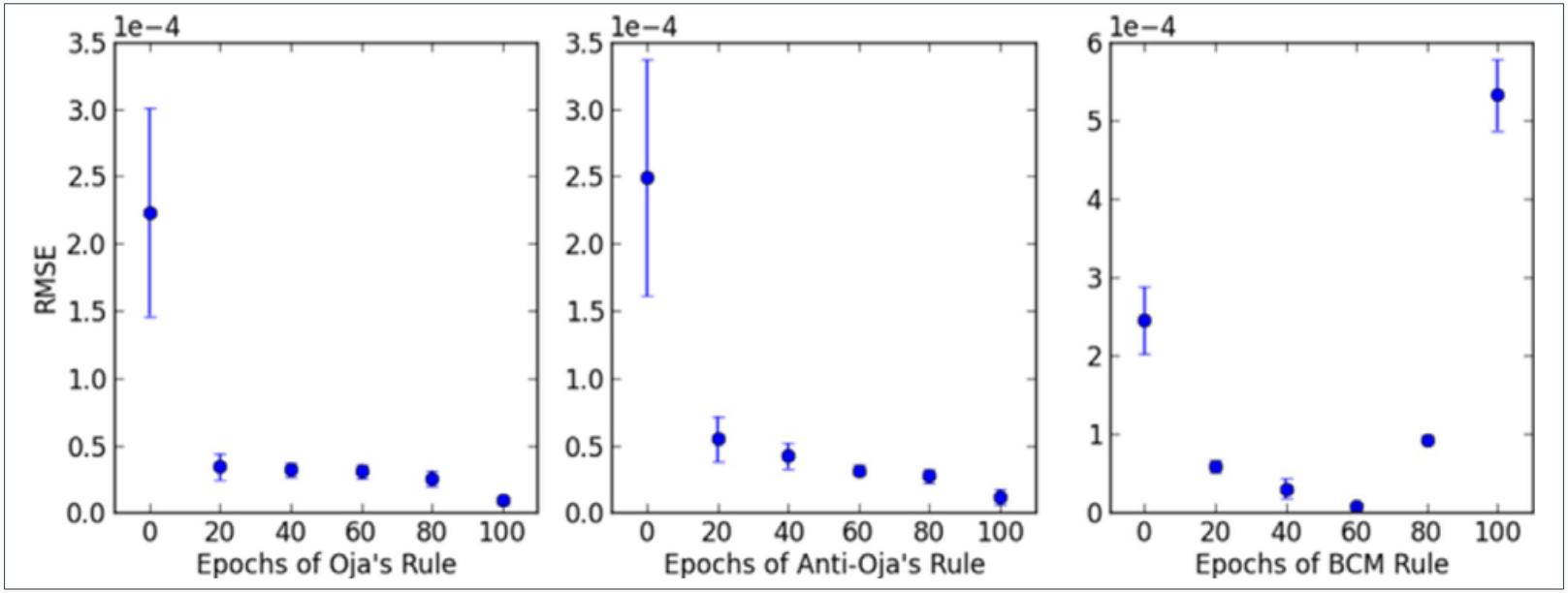


Figure 3: Learning convergence for Mackey Glass prediction task.

- BCM rule causes increase of the spectral radius if applied for too long.
- Performance improvements do not seem to be related to decorrelation effects from anti-Hebbian learning.

Efficacy

Intrinsic Plasticity

Efficacy measured in terms of **average performance** over the parameter choices, and its **standard deviation**.

Tasks:

- Memory capacity (Jaeger)
- NARMA time series
- digit recognition from speech

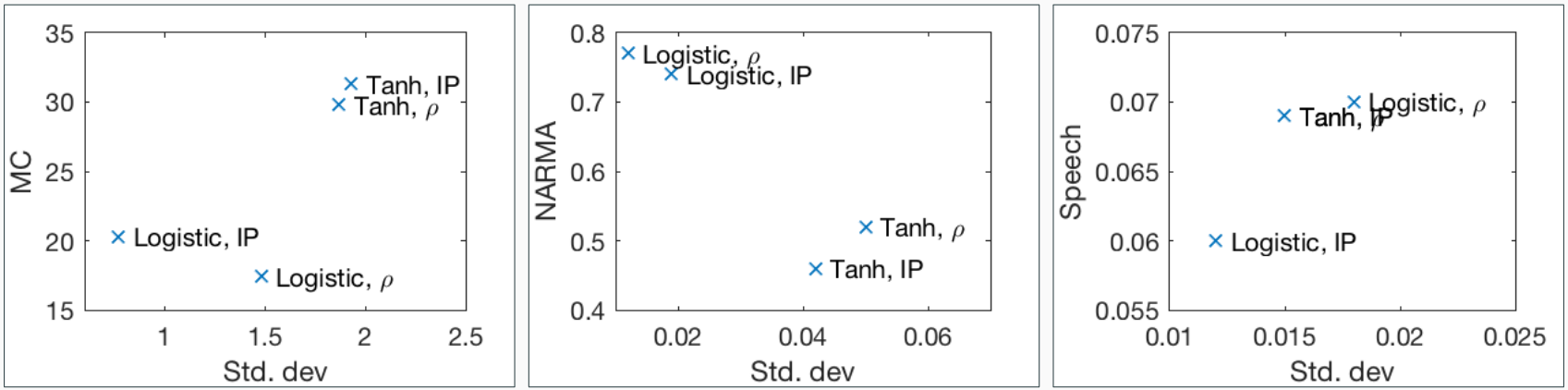


Figure 4: IP rules applied to the three tasks.

Anti-Oja and BCM

Efficacy measured as RMSE over the **Mackey-Glass** time series (and others).

- "Input size": number of history states included in the input
- Spectral radius rescaled to be < 1

Method	Input size			
	2		5	
	$RMSE_{train}$	$RMSE_{test}$	$RMSE_{train}$	$RMSE_{test}$
No plasticity	1.230 E−6	1.695 E−6	9.819 E−7	9.811 E−7
anti-Oja	4.941 E−7	4.941 E−7	2.305 E−7	2.244 E−7
BCM	7.107 E−7	6.987 E−7	1.871 E−7	1.848 E−7

Table 1: RMSE on MG time series for tau = 17 & reservoir size = 1000.

Conclusion

Summary

Unsupervised adaptation is an interesting approach to improve the richness of the reservoir states:

- Biological plausibility
- # unlabeled data $\geq$ # labeled data

It is a tool that must **accompany** a proper model selection for the reservoir, not replace it.

Tradeoff between:

- model selection granularity
- computational complexity

Bibliography


Š. Babinec and J. Pospíchal.
Improving the prediction accuracy of echo state neural networks by anti-oja's learning.
In *International Conference on Artificial Neural Networks*, pages 19–28. Springer, 2007.


J. Chrol-Cannon and Y. Jin.
Computational modeling of neural plasticity for self-organization of neural networks.
BioSystems, 125:43–54, 2014.


D. O. Hebb et al.
The organization of behavior, 1949.


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


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


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


M.-H. Yusoff, J. Chrol-Cannon, and Y. Jin.
Modeling neural plasticity in echo state networks for classification and regression.
Information Sciences, 364:184–196, 2016.