Time-delay differential equations in machine learning

Lyudmila Grigoryeva[1], Julie Henriques[2], Laurent Larger[2], Juan-Pablo Ortega[3,4]

[1]Universität Konstanz, Germany

[2]Université Bourgogne Franche-Comté, France

[3]Centre National de la Recherche Scientifique (CNRS), France

[4]Universität Sankt Gallen, Switzerland

L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. Stochastic time series forecasting using time-delay reservoir computers: performance and universality. *Neural Networks*, 55:59–71, 2014.

L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. Optimal nonlinear information processing capacity in delay-based reservoir computers. *Scientific Reports*, 5(12858):1–11, 2015.

L. Grigoryeva, J. Henriques, L. Larger, J.-P. Ortega, 2016. Nonlinear memory capacity of parallel time-delay reservoir computers in the processing of multidimensional signals. To appear in *Neural Computation*.

L. Grigoryeva, J. Henriques, J.-P. Ortega, 2015. Quantitative evaluation of the performance of discrete-time reservoir computers in the forecasting, filtering, and reconstruction of stochastic stationary signals. Preprint.

L. Grigoryeva, J.-P. Ortega, 2016. Ridge regression with homoscedastic residuals: generalization error with estimated parameters. Preprint.

Outline

1. Reservoir computing: brain-inspired machine learning paradigm
2. Time-Delay Reservoir (TDR) computers:
   - Physical implementation with opto- and electronic systems
   - High-speed and excellent computational performance
   - Architecture of TDR computers
3. Preliminary empirical results:
   - Application of TDR to stochastic nonlinear time series forecasting (multivariate VEC-GARCH models)
   - Parallel reservoir architectures and task-universality
4. Theoretical results on optimal TDR architecture:
   - Unimodality versus bimodality; stability of the TDR
   - VAR(1) model as the TDR approximating model
   - Nonlinear capacity as a quantative measure of performance
5. Further research

Machine learning and brain-inspired neural networks

Machine learning: construction and development of algorithms that can "learn" from the data and are able to adaptively make decisions.

Neural networks: brain-inspired family of statistical models and algorithms that are represented as the collection of interconnected neurons-nodes that have task-adaptive features. Proved to perform in estimation or approximation of functions that are generally unknown (pattern recognition, classification, forecasting).

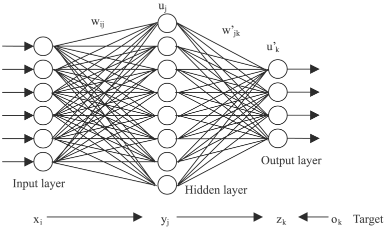


Figure 1: Conventional NN: the weights of the nodes and the activation function have to be chosen at the training stage depending on the task. Disadvantages: convoluted and sometimes ill-defined optimization algorithms for weights determining.

Reservoir computing: brain-inspired machine learning paradigm

- Fundamentally new approach to neural computing [Jae01, JH04, MNM02, VSDS07, LJ09]; defining features of RC: the fading-memory, separation, and approximation properties [LJ09]
- Modification of the traditional RNN in which the architecture and the neuron weights of the network are created in advance (for example randomly) and remain unchanged during the training stage
- The output signal is obtained in the RC with a linear readout layer that is trained using the teacher signal via a ridge (Tikhonov regularized) regression

Physical implementation: reservoir computing (RC) devices

- A major feature of the RC is the possibility of constructing physical realizations of reservoirs instead of simulating them (numerically)
- Chaotic dynamical systems can be used to construct reservoirs that exhibit the RC features: in [ASV+11] using chaotic electronic oscillators or using optoelectronic devices like in [LSB+12]

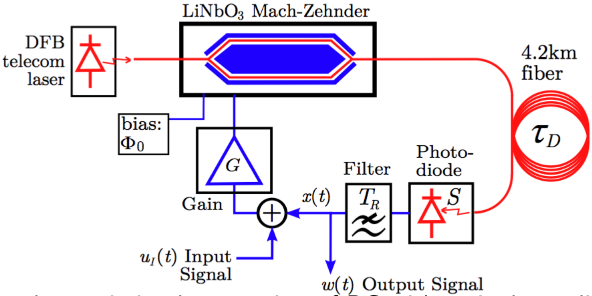


Figure 3: Optoelectronic implementation of RC with a single nonlinear element subject to delayed feedback [LSB+12]

Objectives

- address the reservoir design and working principle problems
- application of RC in the non-deterministic tasks: forecasting of stochastic time series

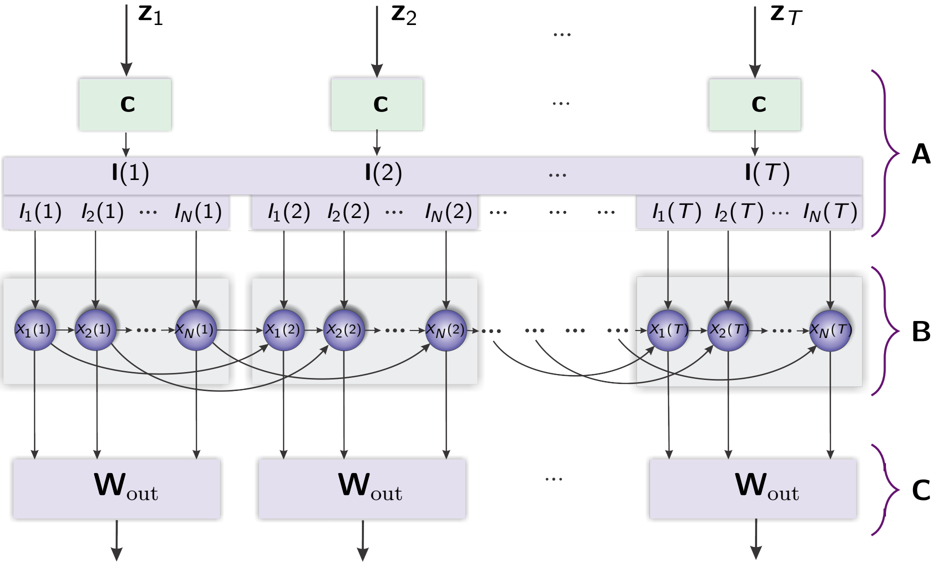


Figure 4: Diagram of architecture of the time-delay reservoir (TDR) and 3 modules of the reservoir computer (RC): the input layer A, the time-delay reservoir B, and the readout layer C.

Input module

Construction of the input layer depends on the computational task of interest and involves the values of the input signal at a given t and the **input mask**; consists of multiplexing the input signal over the delay period and forcing its mean to be zero.

Consider multi-dimensional time series as the input signal: in this case $\mathbf{z}(t) \in \mathbb{R}^n$ and for each t define $\mathbf{I}(t) := C\mathbf{z}(t) \in \mathbb{R}^N$, where $C \in \mathbb{M}_{N,n}$ is the input mask [GHLO14]

Construction of the time-delay reservoir (TDR)

TDRs are based on the "interaction" of the discrete input signal $z(t) \in \mathbb{R}$ with the solution space of a TDDE of the form

$$\dot{x}(t) = -x(t) + f(x(t-\tau), I(t), \boldsymbol{\theta}), \tag{1}$$

where f is a nonlinear smooth function (**nonlinear kernel**), $\boldsymbol{\theta} \in \mathbb{R}^K$ is the parameter vector, $\tau > 0$ is the **delay**, $x(t) \in \mathbb{R}$, and $I(t) \in \mathbb{R}$ is obtained via temporal multiplexing of the input signal $z(t)$ over the delay period; $x \in C^1([-\tau, 0], \mathbb{R})$ needs to be specified prior.

The choice of nonlinear kernel f is determined by the physical implementation; consider two parametric sets of kernels:

- the Mackey-Glass [MG77]: $f(x, I, \boldsymbol{\theta}) = \frac{\eta(x+\gamma I)}{1+(x+\gamma I)^p}$, $\boldsymbol{\theta} = (\eta, \gamma, p)$
- the Ikeda [Ike79]: $f(x, I, \boldsymbol{\theta}) = \eta \sin^2(x + \gamma I + \phi)$, $\boldsymbol{\theta} = (\eta, \gamma, \phi)$

Used in the RC electronic [ASV+11] and optoelectronic [LSB+12] realizations.

Continuous time model of TDR

Consider the regular sampling of solution $x(t)$ of (1) during a given time-delay interval and define $x_i(t)$ the **value** of the ith neuron of the reservoir at time $t\tau$ as

$$x_i(t) := x(t\tau - (N-i)d), \quad i \in \{1, \dots, N\}, \quad t \in \mathbb{Z},$$

where $\tau := dN$, d the **separation between neurons** and we also say that $x_i(t)$ is the ***i*th neuron value of the *t*th layer of the reservoir**.

Discrete time model of TDR

Consider the Euler time-discretization of (1) with integration step $d := \tau / N$:

$$(x(t) - x(t-d))/d = -x(t) + f(x(t-\tau), I(t), \boldsymbol{\theta}). \tag{2}$$

Define **neuron layers** $\mathbf{x}(t)$ and **input layers** $\mathbf{I}(t)$, $t \in \mathbb{Z}$ by setting

$$x_i(t) := x(t\tau - (N-i)d), \quad I_i(t) := I(t\tau - (N-i)d), \quad i \in \{1, \ldots, N\}, \quad t \in \mathbb{Z},$$

where $x_i(t)$ is the **ith neuron value of the tth layer of the reservoir**. Then the solutions of (2) are given by

$$x_i(t) := e^{-\xi} x_{i-1}(t) + (1 - e^{-\xi}) f(x_i(t-1), I_i(t), \boldsymbol{\theta}), \;\; x_0(t) := x_N(t-1), \;\; \xi := \log(1+d),$$

A smooth map $F : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^K \to \mathbb{R}^N$ specifies the neuron values as a recursion via

$$\mathbf{x}(t) = F(\mathbf{x}(t-1), \mathbf{I}(t), \boldsymbol{\theta}), \tag{3}$$

where F is constructed out of the nonlinear kernel map f; F is referred to as the **reservoir map**.

Output module

Let the training be carried out with the input layers $\mathbf{I} := \{\mathbf{I}(1), \ldots, \mathbf{I}(T^*)\}$, that is, for each input layer $\mathbf{I}(t) := (I_1(t), \ldots, I_N(t))$, $t \in \{1, \ldots, T^*\}$, there is a corresponding **teaching signal** $\mathbf{y}(t) \in \mathbb{R}^n$ (in general, $N \gg n$).

Readout W_{out} is given by the solution of the following ridge (or Tikhonov [Tik43]) linear regression problem

$$W_{\mathrm{out}} := \underset{W \in \mathbb{M}_{N,n}}{\arg\min} \left(\sum_{t=1}^{T^*} \| W^\top \cdot \mathbf{x}(t) - \mathbf{y}(t) \|^2 + \lambda \| W \|_{\mathrm{Frob}}^2 \right), \tag{4}$$

whose solution is given by

$$W_{\mathrm{out}} = (XX^T + \lambda \mathbb{I}_N)^{-1} XY, \tag{5}$$

where $X \in \mathbb{M}_{N,T^*}$ is the reservoir output given by $X_{i,j} := x_i(j)$ and $Y \in \mathbb{M}_{T^*,n}$ is the teaching matrix containing the vectors $\mathbf{y}(t)$, $t \in \{1, \ldots, T^*\}$, organized by rows, $\lambda \in \mathbb{R}$ is a regularization parameter (usually obtained via cross-validation).

Stochastic nonlinear time series forecasting with TDR

We propose a TDR based non-parametric approach to forecasting of the stochastic time series which has the following salient advantages:

1. The model selection and estimation stages are incorporated into the training of the TDR with the observed historical data
2. Various non-parametric approaches proved to be efficient in the forecasting of specific time series and are applied in a vast range of forecasting tasks
3. The global reservoir parameters can be optimized in a flexible way to give the best performance with respect to the chosen criteria (in the case of time series forecasting it may be the mean square forecasting error)

Goal

To show the pertinence of using the TDRs in the **nonlinear forecasting of stochastic time series** compared to the standard parametric Box-Jenkins approach. The **nonlinear VEC-GARCH** (generalized autoregressive conditionally heteroscedastic) models proposed by Bollerslev et al [BEW88] are used as data generating process.

Motivation behind the choice of the VEC-GARCH models

The VEC-GARCH family is widely used in financial econometrics as a tool to forecast volatility; captures the specific properties of time series: leptokurticity, volatility clustering, and asymmetric response to volatility shocks. The reasons to choose the VEC-GARCH model as a benchmark include:

1. The model is difficult to calibrate; n-dimensional VEC(1,1) model requires estimating of $n(n+1)(n(n+1)+1)/2$ parameters subjected to specific constraints imposed by the model
2. The explicit expression of the optimal volatility forecast is available, hence the associated error can be computed and used to asses the performance of the TDR
3. The functional dependence between the time series elements that generate the information set and the forecast based on that information set, is nonlinear

General setup

Consider the n-dimensional conditionally heteroscedastic discrete-time process

$$\mathbf{z}_t = H_t^{1/2}\boldsymbol{\epsilon}_\mathbf{t}, \quad \{\boldsymbol{\epsilon}_\mathbf{t}\} \sim \mathrm{IIDN}(\mathbf{0}, \mathbf{I}_n).$$

The VEC-GARCH(1,1) model is determined by

$$\mathbf{h}_t = \mathbf{c} + A\boldsymbol{\eta}_{t-1} + B\mathbf{h}_{t-1}, \tag{6}$$

where $\mathbf{h}_t := \mathrm{vech}(H_t)$, $\boldsymbol{\eta}_t := \mathrm{vech}(\mathbf{z}_t\mathbf{z}_t^T)$, $\mathbf{c} \in \mathbb{R}^N$, and $A, B \in \mathbb{M}_N$ with $N := n(n+1)/2$.

Volatility forecasting

The volatility forecasting task at time T with a forecasting horizon of h time steps consists of providing an estimate $\widehat{H_{T+h}}$ of the conditional covariance matrix H_{T+h} based on the information set $\mathcal{F}_T := \sigma(\mathbf{z}_0, \ldots, \mathbf{z}_T)$. This estimate is produced by minimizing the mean square forecasting error (MSFE) defined as

$$\mathrm{MSFE}(h) := E\left[\left(\mathbf{h}_{T+h} - \widehat{\mathbf{h}_{T+h}}\right)\left(\mathbf{h}_{T+h} - \widehat{\mathbf{h}_{T+h}}\right)^T\right],$$

where $\mathbf{h}_{T+h} := \mathrm{vech}(H_{t+h})$ and $\widehat{\mathbf{h}_{T+h}} := \mathrm{vech}(\widehat{H_{t+h}})$.

The optimal forecast $\widehat{\mathbf{h}_{T+h}}$ for $\mathbf{h}_{T+h}$ is given by:

$$\widehat{\mathbf{h}_{T+h}} := \underset{\widetilde{\mathbf{h}_{T+h|\mathcal{F}_T}}}{\arg\min}\, E\left[\left(\mathbf{h}_{T+h} - \widetilde{\mathbf{h}_{T+h|\mathcal{F}_T}}\right)\left(\mathbf{h}_{T+h} - \widetilde{\mathbf{h}_{T+h|\mathcal{F}_T}}\right)^T\right] = E\left[\mathbf{h}_{T+h} \mid \mathcal{F}_T\right]. \tag{7}$$

The optimal forecast for VEC(1,1) model

The **optimal forecast** $\widehat{\mathbf{h}_{T+h}}$ for the VEC(1,1) model can be computed **explicitly** via the following recursion :

$$\begin{aligned} \widehat{\mathbf{h}_{T+1}} &= \mathbf{h}_{T+1} = \mathbf{c} + A\boldsymbol{\eta}_T + B\mathbf{h}_T, \\ \widehat{\mathbf{h}_{T+2}} &= \mathbf{c} + (A+B)\widehat{\mathbf{h}_{T+1}}, \\ &\vdots \\ \widehat{\mathbf{h}_{T+i}} &= \mathbf{c} + (A+B)\widehat{\mathbf{h}_{T+i-1}}, \\ &\vdots \\ \widehat{\mathbf{h}_{T+h}} &= \mathbf{c} + (A+B)\widehat{\mathbf{h}_{T+h-1}}. \end{aligned} \tag{8}$$

The functional dependence between the forecast $\widehat{\mathbf{h}_{T+h}}$ and the elements $\{\mathbf{z}_0, \ldots, \mathbf{z}_T\}$ that generate the information set $\mathcal{F}_T$ is nonlinear.

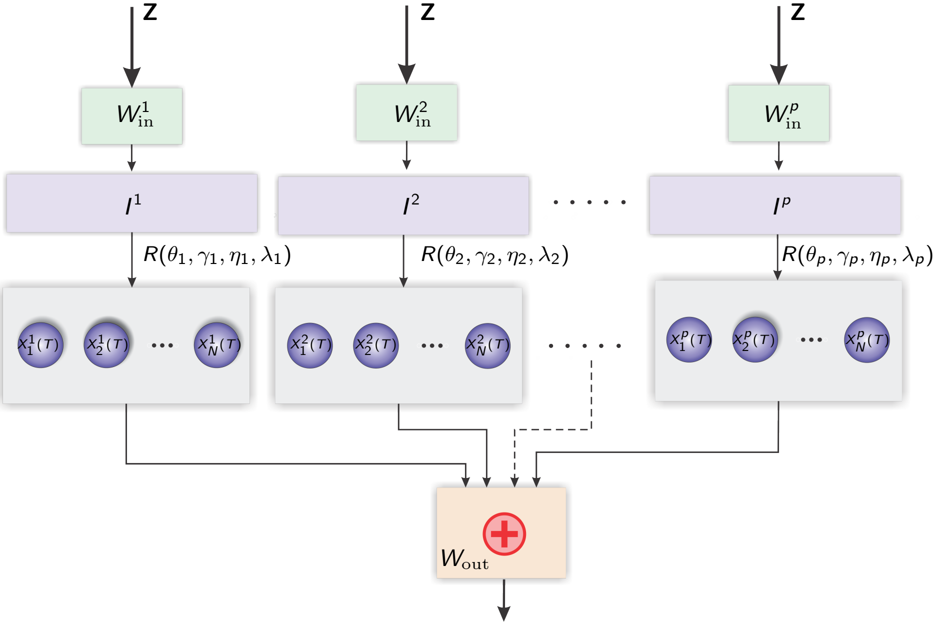
The **MSFE associated to the optimal forecast** can be also computed **explicitly** as we use it as a benchmark to assess the performance of the TDR with the same forecasting task assigned to it.

Parameter optimization of a TDR

- No universal set of optimal parameters (θ, γ, η) that offers top performance of a reservoir for any task assigned to it
- In the case of VEC volatility forecasting the lack of optimality is evidenced when: (i) the forecasting is carried out for different processes (different sets of parameters $\mathbf{c}$, A, and B), (ii) the forecasting horizon changes, that is, different horizons have different optimal reservoir parameters

Two important implications:

- **Numerical cost:** the parameter optimization is carried out via a computational expensive cross validation procedure
- **Parallel reading inefficiency:** in the particular case of the forecasting problem parallel reading can be useful at the time of simultaneously predicting at various horizons out of a single input signal; however, this is only feasible if there is a set of reservoir parameters for which the forecasting performance is acceptable for all the horizons of interest



Z
Z
Z
$W_{\rm in}^1$
$W_{\rm in}^2$
$W_{\rm in}^p$
I^1
I^2
I^p
$R(\theta_1, \gamma_1, \eta_1, \lambda_1)$
$R(\theta_2, \gamma_2, \eta_2, \lambda_2)$
$R(\theta_p, \gamma_p, \eta_p, \lambda_p)$
$x_1^1(T)$ $x_2^1(T)$ ... $x_N^1(T)$
$x_1^2(T)$ $x_2^2(T)$ ... $x_N^2(T)$
$x_1^p(T)$ $x_2^p(T)$ ... $x_N^p(T)$
$W_{\rm out}$

Advantages of parallel reservoirs

Advantages of parallel reservoirs compared to a single optimized reservoir with the same number of neurons

1. **Limited computational effort:** the parallel reservoirs will be constructed by putting together pools of reservoirs with randomly chosen parameter values and by keeping the pool that yields the best performance in an out-of-sample testing step
2. **Better performance for smaller training sample sizes**
3. **Improved universality with respect to changes in the forecasting horizon and in the model specification:** the optimal parameters for the prediction task are not the same neither for different forecasting horizons nor for different data generating processes. This variability is reduced by the use of a parallel array of TDR computers

Empirical results

Four configurations were considered:

(i) **TDR with 400 neurons and grid optimized parameters**

(ii) **TDR with 400 neurons and random optimized parameters**

(iii) **Random optimal parallel array of 40 reservoirs with 10 neurons each**

(iv) **Random optimal parallel array of 80 reservoirs with 5 neurons each**

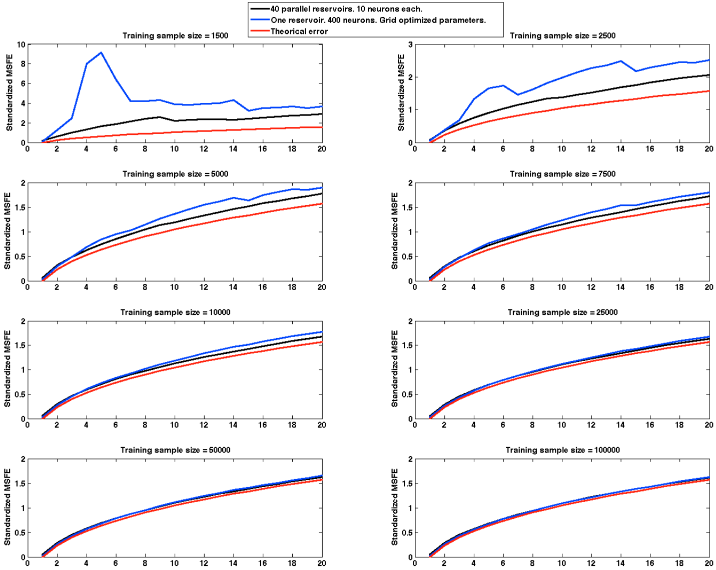


Figure 6: Comparison of the sMSFE committed for different training sample sizes by a single grid optimized TDR with 400 neurons and by a parallel array of 40 reservoirs with 10 neurons each.

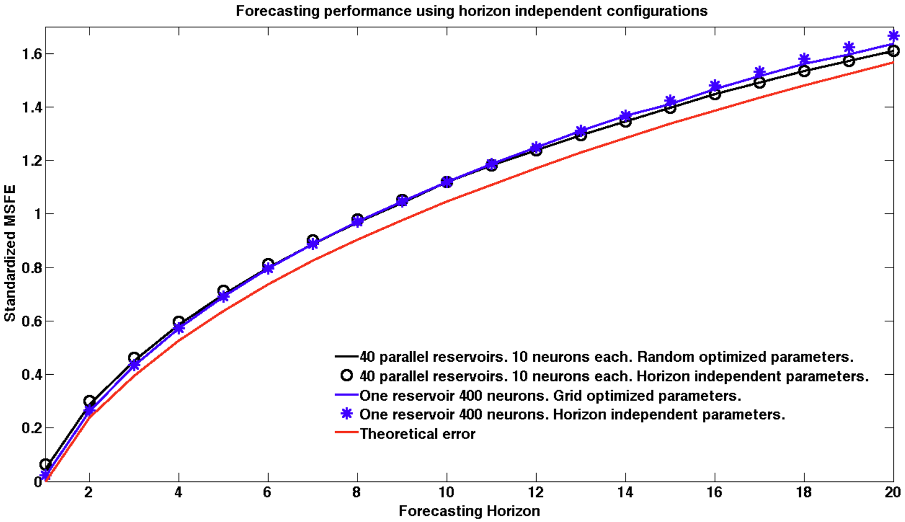


Figure 7: Comparison of the forecasting performances obtained by using horizon adapted parameter configurations and constant parameters (appear more frequently in the tables).

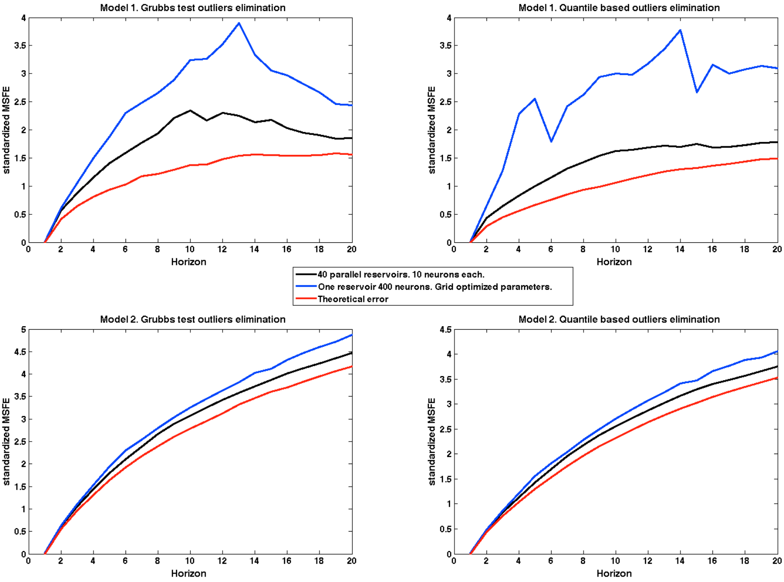


Figure 8: Forecasting performance under model misspecification. In the left hand side outliers are eliminated using the Grubbs test with a significance level of 5%; in the right hand side the quantiles under 0.1% and above 99.9% are eliminated.

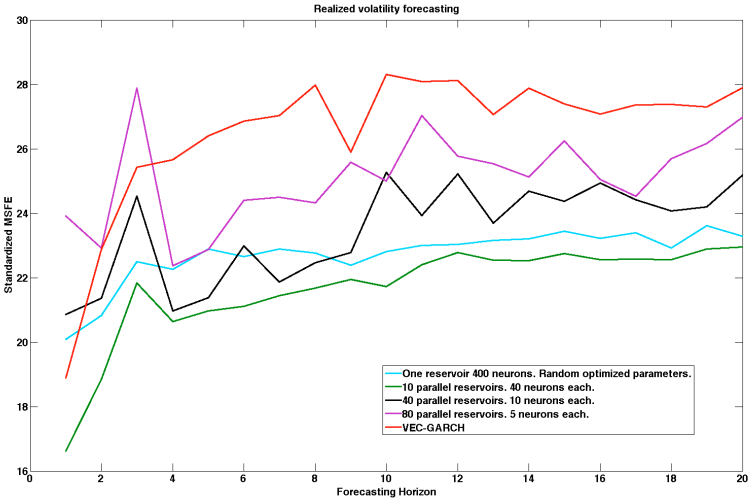


Figure 9: Average realized volatility forecasting performance using RC and VEC(1,1) models estimated via maximum likelihood (MLE). The sMSFE reported is obtained with the estimated parametric models. All the TDRs considered have been generated using the nonlinear Mackey-Glass kernel with $p = 2$.

Main contributions of the empirical work [GHLO14]

- Demonstrate the pertinence of using non-parametric TDR method in the nonlinear forecasting of the multivariate discrete time stochastic time series compared to the standard Box-Jenkins parametric approach (model selection, estimation, diagnostic checking, forecasting)
- Present the evidence of shortfall in task-universality of a single reservoir; given a time-delay reservoir architecture, a set of optimal reservoir parameters $\boldsymbol{\theta}$ for a specific assigned task is not universal
- Use parallel pools of TDRs to overcome the deficiency of the task-universality for an individually operating reservoir
- Application of TDRs to forecasting based on the time series of the real financial market data

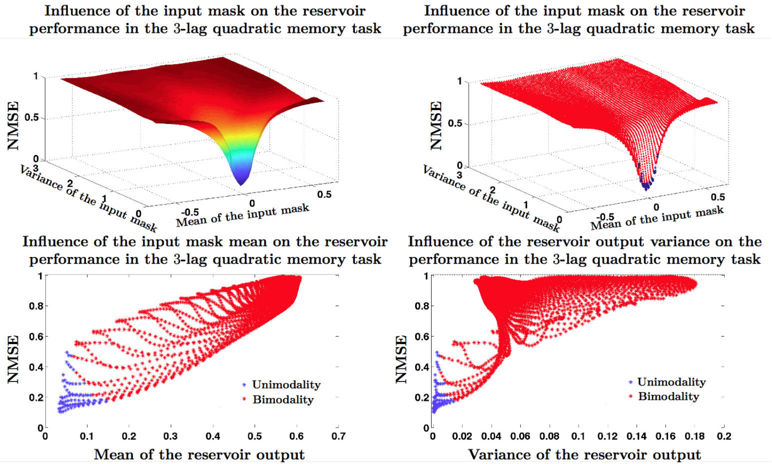


Figure 10. Behavior of the reservoir performance in a quadratic memory task as a function of the $\bar{c}$ and $\mathrm{var}(\mathbf{c})$. The top panels show how the performance degrades very quickly as soon as $\bar{c}$ and $\mathrm{var}(\mathbf{c})$ separate from zero. The bottom panels depict the reservoir performance as a function of the various output means and variances. We have indicated with red markers the cases in which the reservoir visits the stability basin of a contiguous stable equilibrium hence showing how unimodality is associated to optimal performance.

Basic facts

Let $\tau \in \mathbb{R}^+$ be a fixed delay and consider a **time-delay map**

$$X: \quad \begin{array}{ccc} C^1([-\tau,0],\mathbb{R}) \times \mathbb{R} & \longrightarrow & \mathbb{R} \\ (\gamma, t) & \longmapsto & X(\gamma, t). \end{array} \tag{9}$$

Additionally, for any $t \in \mathbb{R}$ define the **shift operator**

$$S_t: \quad \begin{array}{ccc} C^1([-\tau+t,t],\mathbb{R}) & \longrightarrow & C^1([-\tau,0],\mathbb{R}) \\ \gamma & \longmapsto & \gamma \circ \lambda_t, \end{array} \tag{10}$$

where λ_t is the translation operator by $t \in \mathbb{R}$: $\lambda_t(s) := s + t$, for any $s \in \mathbb{R}$.
Let $\gamma \in C^1([-\tau,+\infty),\mathbb{R})$. We say that γ is **a solution of the TDDE** determined by X when

$$\dot{\gamma}(t) = X(S_t \circ \gamma|_{[-\tau+t,t]}, t) \ \text{ for any } \ t \in [0,+\infty). \tag{11}$$

Note that the TDDE

$$\dot{x}(t) = -x(t) + f(x(t-\tau), I(t), \boldsymbol{\theta}), \tag{12}$$

is given by

$$X: \quad \begin{array}{ccc} C^1([-\tau,0],\mathbb{R}) \times \mathbb{R} & \longrightarrow & \mathbb{R} \\ (\gamma, t) & \longmapsto & -\gamma(0) + f(\gamma(-\tau), I(t), \boldsymbol{\theta}). \end{array} \tag{13}$$

Definition

We say that the time-delay map X is locally Lipschitzian on the open set $\Omega \subset C^1([-\tau, 0], \mathbb{R}) \times \mathbb{R}$ if it is Lipschitzian in any compact subset of Ω, that is, for any compact subset Ω_0 of Ω there exists a constant $K \in \mathbb{R}^+$ such that for all (γ_1, t) and (γ_2, t) in Ω_0 one has

$$|X(\gamma_1, t) - X(\gamma_2, t)| < K||\gamma_1 - \gamma_2||_\infty. \tag{14}$$

Theorem (Existence and uniqueness of solutions)

Let X be a continuous and locally Lipschitzian time-delay map in $C^1([-\tau, 0], \mathbb{R}) \times \mathbb{R}$. Then, for any $\phi \in C^1([-\tau, 0], \mathbb{R})$ there exists a unique $\Gamma_\phi \in C^1([-\tau, +\infty), \mathbb{R})$ s.t.

$$\begin{cases} \Gamma_\phi(t) & = \phi(t), & \text{for any } t \in [-\tau, 0] \\ \dot{\Gamma}_\phi(t) & = X(S_t \circ \Gamma_\phi|_{[-\tau+t, t]}, t), & \text{for any } t \in (0, +\infty]. \end{cases} \tag{15}$$

We say that Γ_ϕ is the **solution** *of the TDDE determined by X with initial condition ϕ, or simply the solution through ϕ. The associated* **flow** *is defined as the map*

$$\begin{array}{rccc} F: & [-\tau, +\infty) \times C^1([-\tau, 0], \mathbb{R}) & \longrightarrow & \mathbb{R} \\ & (t, \phi) & \longmapsto & \Gamma_\phi(t) \end{array} \tag{16}$$

and note that $F.(\phi) \in C^1([-\tau, +\infty), \mathbb{R})$.

We now recall also some basic notions of stability of common use in the TDDE context; see [Hal77] and [WHS10] for details.

Let $x_0 \in \mathbb{R}$ and let $\phi_{x_0} \in C^1([-\tau, 0], \mathbb{R})$ be the constant curve at x_0. We say that the point x_0 is an **equilibrium** of the TDDE determined by the time-delay map and with flow F whenever $F_t(\phi_{x_0}) = x_0$, for any $t \in [-\tau, +\infty)$.

The equilibrium x_0 is said to be **stable** (respectively **asymptotically stable**) if for any $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that for any $\phi \in C^1([-\tau, 0], \mathbb{R})$ with $\|\phi - \phi_{x_0}\|_\infty < \delta(\epsilon)$, we have that $|F_t(\phi) - x_0| < \epsilon$, for any $t \in [-\tau, +\infty)$ (respectively $\lim_{t \to \infty} F_t(\phi) = x_0$).

Lyapunov-Krasovskiy stability theorem

Theorem (Lyapunov-Krasovskiy stability theorem)

let $x_0 \in \mathbb{R}$ *be an equilibrium of the time-delay differential equation with flow* $F : [-\tau, +\infty) \times C^1([-\tau, 0], \mathbb{R})) \longrightarrow \mathbb{R}$*. Let* $u, v, w : \overline{\mathbb{R}^+} \longrightarrow \overline{\mathbb{R}^+}$ *be continuous nondecreasing functions such that* $u(0) = v(0) = 0$ *and* $u(t), v(t), w(t) > 0$ *for any* $t \in (0, +\infty)$*. If there exists a continuously differentiable functional* V

$$V : \quad C^1([-\tau, +\infty), \mathbb{R}) \times \mathbb{R} \quad \longrightarrow \mathbb{R} \tag{17}$$

such that for any $\phi \in C^1([-\tau, 0], \mathbb{R}))$ *and any* $t \in [0, +\infty)$ *satisfies that*

(i) $u(|\phi(0)|) \leq V(F.(\phi), t) \leq v(||\phi||_\infty)$,

(ii) $\dot{V}(F.(\phi), t) := \dfrac{d}{dt} V(F.(\phi), t) \leq -w(|\phi(0)|)$,

then x_0 *is asymptotically stable. If* $w(t) \geq 0$ *then* x_0 *is just stable. A functional* V *that satisfies these conditions is called a* **Lyapunov-Krasovskiy functional**.

Stability of the TDR: continuous time model

Use Lyapunov-Krasovskiy stability theorem [Kra63] to establish sufficient conditions for the stability of the equilibria of the TDDE

$$\dot{x}(t) = x(t) + f(x(t-\tau), I(t), \boldsymbol{\theta}). \tag{18}$$

where f is the **nonlinear kernel** function, $\boldsymbol{\theta} \in \mathbb{R}^K$ is the reservoir parameters vector, $\tau > 0$ is the **delay**, $x(t) \in \mathbb{R}$, and $I(t) \in \mathbb{R}$ is obtained via temporal multiplexing over τ of the input signal $z(t)$.

The main tool in the application of that result is the use of a Lyapunov-Krasovskiy functional of the form

$$\begin{array}{rccc} V: & C^1([-\tau, +\infty], \mathbb{R}) \times \mathbb{R} & \longrightarrow & \mathbb{R} \\ & (x_\phi, t) & \longmapsto & \frac{1}{2} x_\phi(t)^2 + m \int_{t-\tau}^{t} x_\phi(s)^2 ds, \end{array} \tag{19}$$

where $m \in \mathbb{R}^+$ and $x_\phi = F.(\phi)$ for some initial curve $\phi \in C^1([-\tau, 0], \mathbb{R})$. See [Kra63], [Hal77] and [WHS10] for extensive discussion.

Theorem (Grigoryeva, Henriques, Larger, Ortega, 2014)

Let x_0 be an equilibrium of the time-delay differential equation (18) in autonomous regime, that is, when $I(t) = 0$, and suppose that there exists $\varepsilon > 0$ and $k_\varepsilon \in \mathbb{R}$ such that one of the following conditions holds

(i) $f(x + x_0, 0, \boldsymbol{\theta}) \leq k_\varepsilon x + x_0$ *for all* $x \in (-\varepsilon, \varepsilon)$

(ii) $\dfrac{f(x + x_0, 0, \boldsymbol{\theta}) - x_0}{x} \leq k_\varepsilon$ *for all* $x \in (-\varepsilon, \varepsilon)$.

If $|k_\varepsilon| < 1$ *then* x_0 *is* **asymptotically stable**. *If* $|k_\varepsilon| \leq 1$ *then* x_0 *is* **stable**.

Corollary (Grigoryeva, Henriques, Larger, Ortega, 2014)

Let x_0 be an equilibrium of the TDDE (18) and suppose that the nonlinear reservoir kernel function f is continuously differentiable at x_0. If $|\partial_x f(x_0, 0, \boldsymbol{\theta})| < 1$ (respectively, $|\partial_x f(x_0, 0, \boldsymbol{\theta})| \leq 1$), then x_0 is asymptotically stable (respectively, stable).

Corollary (Stability of the equilibria of the Mackey-Glass TDDE; Grigoryeva, Henriques, Larger, Ortega, 2014)

Consider the TDDE (18) *in the autonomous regime constructed with the Mackey-Glass kernel with* $p = 2$*, that is,*

$$f(x, 0, \boldsymbol{\theta}) = \frac{\eta x}{1 + x^2}. \tag{20}$$

This TDDE exhibits two families of equilibria depending on the values of η*:*

(i) *The trivial solution* $x_0 = 0$*, for any* $\eta \in \mathbb{R}$*. The equilibrium* $x_0 = 0$ *is asymptotically stable (respectively, stable) if* $|\eta| < 1$ *(respectively,* $|\eta| \leq 1$*).*

(ii) *The non-trivial solutions* $x_0 = \pm\sqrt{\eta - 1}$*, for any* $\eta > 1$*. The equilibria* $x_0 = \pm\sqrt{\eta - 1}$ *are asymptotically stable (respectively, stable) whenever* $1 < \eta < 3$ *(respectively,* $1 < \eta \leq 3$*).*

Corollary (Stability of the equilibria of the Ikeda TDDE; Grigoryeva, Henriques, Larger, Ortega, 2014)

Consider the TDDE (18) *in autonomous regime based on the Ikeda kernel,*

$$f(x, 0, \boldsymbol{\theta}) = \eta \sin^2(x + \phi). \tag{21}$$

The Ikeda nonlinear TDDE exhibits two families of equilibria:

(i) *The trivial solution $x_0 = 0$ for any $\eta \in \mathbb{R}$ and $\phi = \pi n$, $n \in \mathbb{Z}$. The equilibium $x_0 = 0$ is asymptotically stable for any $\eta \in \mathbb{R}$.*

(ii) *The non-trivial equilibria x_0 are obtained as solutions of the equation $x_0 = \eta \sin^2(x_0 + \phi)$, for any $\eta \in \mathbb{R}$ and $\phi \neq \pi n$, $n \in \mathbb{Z}$. These equilibria are asymptotically stable (respectively, stable) if*

$$|\sin(2x_0 + 2\phi)| < \frac{1}{|\eta|} \quad (\textit{respectively, } |\sin(2x_0 + 2\phi)| \leq \frac{1}{|\eta|}). \tag{22}$$

When $|\eta| < 1$ (respectively, $|\eta| \leq 1$), there exists only one non-trivial equilibrium that is always asymptotically stable (respectively, stable).

Stability of the TDR: discrete time approximation

The **discrete time approximation** of the TDR is

$$x_i(t) = e^{-i\xi} x_N(t-1) + (1-e^{-\xi}) \sum_{j=0}^{i-1} e^{-j\xi} f(x_{i-j}(t-1), I_{i-j}(t), \boldsymbol{\theta}), \tag{23}$$

which corresponds to $\mathbf{x}(t) = F(\mathbf{x}(t-1), \mathbf{I}(t), \boldsymbol{\theta})$ that uniquely determines the reservoir map $F : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^K \longrightarrow \mathbb{R}^N$.

Let $x_0 \in \mathbb{R}$ and $\mathbf{x}_0 := x_0 \mathbf{i}_N \in \mathbb{R}^N$. Let $A(\mathbf{x}_0, \boldsymbol{\theta}) := D_{\mathbf{x}} F(\mathbf{x}_0, \mathbf{0}_N, \boldsymbol{\theta})$ be referred to as the **connectivity matrix** of the reservoir at the point $\mathbf{x}_0$:

$$A(\mathbf{x}_0, \boldsymbol{\theta}) = \begin{pmatrix} \Phi & 0 & \dots & 0 & e^{-\xi} \\ e^{-\xi}\Phi & \Phi & \dots & 0 & e^{-2\xi} \\ e^{-2\xi}\Phi & e^{-\xi}\Phi & \dots & 0 & e^{-3\xi} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e^{-(N-1)\xi}\Phi & e^{-(N-2)\xi}\Phi & \dots & e^{-\xi}\Phi & \Phi + e^{-N\xi} \end{pmatrix}, \tag{24}$$

where $\Phi := (1-e^{-\xi})\partial_x f(x_0, 0, \boldsymbol{\theta})$ and $\partial_x f(x_0, 0, \boldsymbol{\theta})$ is the first derivative of the nonlinear kernel f with respect to the first argument and computed at the point $(x_0, 0, \boldsymbol{\theta})$.

Proposition (Grigoryeva, Henriques, Larger, Ortega, 2014)

The point $x_0 \in \mathbb{R}$ is an equilibrium of the time-delay differential equation (18) *in autonomous regime, that is when $I(t) = 0$, if and only if the vector $\mathbf{x}_0 := x_0 \mathbf{i}_N$ is a fixed point of the N-dimensional discretized nonlinear time-delay reservoir*

$$\dot{\mathbf{x}}(t) = F(\mathbf{x}(t-1), \mathbf{I}(t), \boldsymbol{\theta}) \tag{25}$$

in autonomous regime, that is, when $\mathbf{I}(t) = \mathbf{0}_N$.

Theorem (Grigoryeva, Henriques, Larger, Ortega, 2014)

Let $\mathbf{x}_0 = x_0 \mathbf{i}_N$ be a fixed point of the N-dimensional recursion $\mathbf{x}(t) = F(\mathbf{x}(t-1), \mathbf{I}(t), \boldsymbol{\theta})$ in autonomous regime. Then, $\mathbf{x}_0 \in \mathbb{R}^N$ is asymptotically stable (respectively stable) if $|\partial_x f(x_0, 0, \boldsymbol{\theta})| < 1$ (respectively, $|\partial_x f(x_0, 0, \boldsymbol{\theta})| \leq 1$).

Optimal performance: stability and unimodality

Conclusions: *Optimal TDR performance is attained when the TDR operates in a unimodal regime around an asymptotically stable state. We find common stability conditions for the continuous and discrete time systems.*

Approximating model and nonlinear memory capacity

(1) We construct an approximation of the TDR via its partial linearization at the equilibrium point with respect to the delayed self feedback term and respecting the nonlinearity of the input injection.

The approximating model

Consider a stable equilibrium $x_0 \in \mathbb{R}$ of the autonomous system associated to (1) or, equivalently, a stable fixed point $\mathbf{x}_0 := (x_0, \ldots, x_0)^\top \in \mathbb{R}^N$ of (3). We construct the approximation of (3) by using its linearization at $\mathbf{x}_0$ with respect to the delayed self-feedback and its Rth-order Taylor expansion with respect to its dependence on the signal injection:

$$\mathbf{x}(t) = F(\mathbf{x}_0, \mathbf{0}_N, \boldsymbol{\theta}) + A(\mathbf{x}_0, \boldsymbol{\theta})(\mathbf{x}(t-1) - \mathbf{x}_0) + \boldsymbol{\varepsilon}(t), \tag{26}$$

where $A(\mathbf{x}_0, \boldsymbol{\theta}) := D_{\mathbf{x}} F(\mathbf{x}_0, \mathbf{0}_N, \boldsymbol{\theta})$ and $\boldsymbol{\varepsilon}(t)$ is given by:

$$\boldsymbol{\varepsilon}(t) = (1 - e^{-\xi}) \left(q_R\left(z(t), c_1\right), \ldots, q_R\left(z(t), c_1, \ldots, c_N\right)\right)^\top,$$

with

$$q_R\left(z(t), c_1, \ldots, c_r\right) := \sum_{i=1}^{R} \frac{z(t)^i}{i!} (\partial_I^{(i)} f)(x_0, 0, \boldsymbol{\theta}) \sum_{j=1}^{r} e^{-(r-j)\xi} c_j^i,$$

and $(\partial_I^{(i)} f)(x_0, 0, \boldsymbol{\theta})$ the ith order partial derivative of the nonlinear kernel f with respect to $I(t)$ evaluated at $(x_0, 0, \boldsymbol{\theta})$.

(2) For statistically independent input signals the approximation (26) allows us to visualize the TDR as a N-dimensional vector autoregressive stochastic process of order one (VAR(1), [L05]).

Let the input signal be $\{z(t)\}_{t\in\mathbb{Z}} \sim \mathrm{IID}(0,\sigma_z^2)$, then $\{\mathbf{I}(t)\}_{t\in\mathbb{Z}} \sim \mathrm{IID}(\mathbf{0}_N,\Sigma_I)$, with $\Sigma_I := \sigma_z^2\mathbf{c}^\top\mathbf{c}$, and $\{\varepsilon(t)\}_{t\in\mathbb{Z}} \sim \mathrm{IID}(\boldsymbol{\mu}_\varepsilon,\Sigma_\varepsilon)$ with

$$\boldsymbol{\mu}_\varepsilon = (1-e^{-\xi})\left(q_R\left(\mu_z,c_1\right),\ldots,q_R\left(\mu_z,c_1,\ldots,c_N\right)\right)^\top,$$

where $\mu_z^i := \mathrm{E}\left[z(t)^i\right]$ and $\Sigma_\varepsilon := \mathrm{E}\left[(\boldsymbol{\varepsilon}(t)-\boldsymbol{\mu}_\varepsilon)(\boldsymbol{\varepsilon}(t)-\boldsymbol{\mu}_\varepsilon)^\top\right] \in \mathbb{S}_N$ with the entries given by:

$$\begin{aligned}(\Sigma_\varepsilon)_{ij} =& (1-e^{-\xi})^2((q_R(\cdot,c_1,\ldots,c_i)\cdot q_R(\cdot,c_1,\ldots,c_j))(\mu_z)\\ &-q_R(\mu_z,c_1,\ldots,c_i)q_R(\mu_z,c_1,\ldots,c_j)),\quad i,j=1,\ldots,N.\end{aligned}$$

The process (26) is a VAR(1) model

$$\mathbf{x}(t)-\boldsymbol{\mu}_x = A(\mathbf{x}_0,\boldsymbol{\theta})(\mathbf{x}(t-1)-\boldsymbol{\mu}_x)+(\boldsymbol{\varepsilon}(t)-\boldsymbol{\mu}_\varepsilon) \tag{27}$$

with $\boldsymbol{\mu}_x = (I_N - A(\mathbf{x}_0,\boldsymbol{\theta}))^{-1}(F(\mathbf{x}_0,\mathbf{0}_N,\boldsymbol{\theta}) - A(\mathbf{x}_0,\boldsymbol{\theta})\mathbf{x}_0+\boldsymbol{\mu}_\varepsilon)$ and an autocovariance function $\Gamma(k) := \mathrm{E}\left[\left(\mathbf{x}(t)-\boldsymbol{\mu}_x\right)\left(\mathbf{x}(t-k)-\boldsymbol{\mu}_x\right)^\top\right]$, $k\in\mathbb{Z}$, recursively determined by the Yule-Walker equations [LÖ5]:

$$\begin{aligned}\mathrm{vec}(\Gamma(0)) &= (\mathbb{I}_{N^2} - A(\mathbf{x}_0,\boldsymbol{\theta})\otimes A(\mathbf{x}_0,\boldsymbol{\theta}))^{-1}\,\mathrm{vec}(\Sigma_\varepsilon),\\ \Gamma(k) &= A(\mathbf{x}_0,\boldsymbol{\theta})\Gamma(k-1),\ \ \Gamma(-k)=\Gamma(k)^\top.\end{aligned}$$

The nonlinear memory capacity estimations

(3) The approximation (26) allows us to write the nonlinear capacities of the TDR as the function of the intrinsic architecture parameters $\boldsymbol{\theta}$ and the input mask $\mathbf{c}$.

The nonlinear memory capacity estimations

A h-**lag memory task** is determined by a function $H : \mathbb{R}^{h+1} \to \mathbb{R}$ (in general nonlinear) that is used to generate $y(t) := H(z(t), z(t-1), \ldots, z(t-h)) \in \mathbb{R}$ out of the reservoir input $\{z(t)\}_{t\in\mathbb{Z}}$.

Recall, that the optimal linear readout $\mathbf{W}_{\mathrm{out}}$ adapted to the memory task H is given by the solution of a ridge (or Tikhonov [Tik43]) linear regression problem

$$(\mathbf{W}_{\mathrm{out}}, a_{\mathrm{out}}) := \underset{\mathbf{W}\in\mathbb{R}^N, a\in\mathbb{R}}{\arg\min} \left(\mathrm{E}\left[(\mathbf{W}^\top \cdot \mathbf{x}(t) + a - y(t))^2 \right] + \lambda \|\mathbf{W}\|^2 \right). \tag{28}$$

Using the fact that $\{\mathbf{x}(t)\}_{t\in\mathbb{Z}}$ is the unique stationary solution of VAR(1) approximating system (27) for the TDR (27) obtain

$$\begin{aligned} \mathbf{W}_{\mathrm{out}} &= (\Gamma(0) + \lambda \mathbb{I}_N)^{-1} \mathrm{Cov}(y(t), \mathbf{x}(t)), && (29) \\ a_{\mathrm{out}} &= \mathrm{E}\left[y(t)\right] - \mathbf{W}_{\mathrm{out}}^\top \boldsymbol{\mu}_x, && (30) \end{aligned}$$

where $\boldsymbol{\mu}_x$, $\Gamma(0) \in \mathbb{S}_N$ are provided in (27), and $\mathrm{Cov}(y(t), \mathbf{x}(t))$ is a vector in $\mathbb{R}^N$ that has to be determined for every specific memory task H.

The error committed by the reservoir when using the optimal readout is

$$\begin{aligned}\mathrm{MSE}_H = \mathrm{var}\,(y(t)) - \mathrm{Cov}(y(t), \mathbf{x}(t))^\top (\Gamma(0) + \lambda \mathbb{I}_N)^{-1} (\Gamma(0) + 2\lambda \mathbb{I}_N) \\ \times (\Gamma(0) + \lambda \mathbb{I}_N)^{-1} \mathrm{Cov}(y(t), \mathbf{x}(t)).\end{aligned}$$

Using the VAR(1) approximating model (27) of RC, the corresponding *H*-**memory capacity** is

$$\begin{aligned} C_H(\boldsymbol{\theta}, \mathbf{c}, \lambda) =& \mathrm{Cov}(y(t), \mathbf{x}(t))^\top (\Gamma(0) + \lambda \mathbb{I}_N)^{-1} (\Gamma(0) + 2\lambda \mathbb{I}_N) & (31) \\ & \times (\Gamma(0) + \lambda \mathbb{I}_N)^{-1} \mathrm{Cov}(y(t), \mathbf{x}(t))/\mathrm{var}(y(t)). & (32) \end{aligned}$$

Additionally,

$$0 \leq C_H(\boldsymbol{\theta}, \mathbf{c}, \lambda) \leq 1.$$

Once a specific reservoir and task H have been fixed, the capacity function $C_H(\boldsymbol{\theta}, \mathbf{c}, \lambda)$ can be explicitly written down and it can hence be used to find reservoir parameters $\boldsymbol{\theta}_{\mathrm{opt}}$ and an input mask $\mathbf{c}_{\mathrm{opt}}$ that maximize it, by solving the optimization problem

$$(\boldsymbol{\theta}_{\mathrm{opt}}, \mathbf{c}_{\mathrm{opt}}) := \underset{\boldsymbol{\theta} \in \mathbb{R}^K, \mathbf{c} \in \mathbb{R}^N}{\arg\max}\; C_H(\boldsymbol{\theta}, \mathbf{c}, \lambda). \tag{33}$$

Optimal nonlinear capacity

The *h*-lag quadratic memory task. Take a quadratic task function of the form $H(\mathbf{z}^h(t)) := \mathbf{z}^h(t)^\top Q\mathbf{z}^h(t)$, for some symmetric $h+1$-dimensional matrix Q. In this case $\mathrm{var}(y(t)) = (\mu_z^4 - \sigma_z^4)\sum_{i=1}^{h+1} Q_{ii}^2 + 4\sigma_z^4 \sum_{i=1}^{h+1}\sum_{j>i}^{h+1} Q_{ij}^2$, and

$$\mathrm{Cov}(y(t), x_i(t)) = (1-e^{-\xi})\sum_{j=1}^{h+1}\sum_{r=1}^{N} Q_{jj}(A^{j-1})_{ir} \\ \times (s_R(\mu_z, c_1, \ldots, c_r) - \sigma_z^2 q_R(\mu_z, c_1, \ldots, c_r)),$$

where the polynomial s_R on the variable x is defined as $s_R(x, c_1, \ldots, c_r) := x^2 \cdot q_R(x, c_1, \ldots, c_r)$.

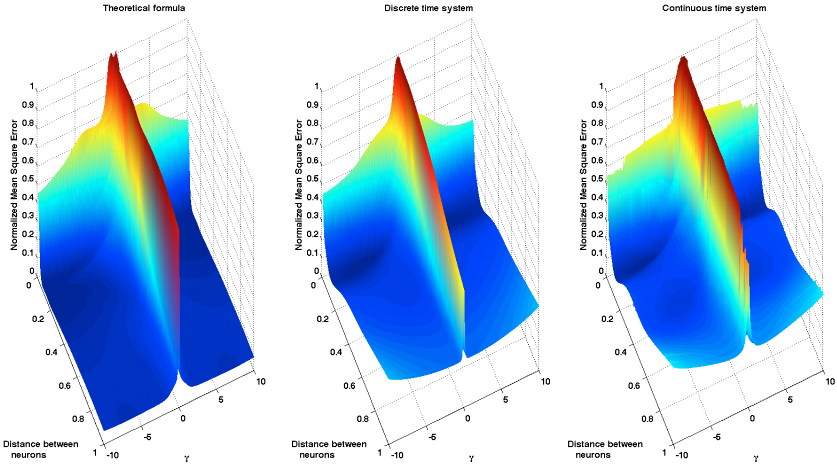


Figure 11. Error exhibited by a TDR computer with a Mackey-Glass kernel in a 3-lag quadratic memory task as a function of the separation between neurons d and the parameter γ, respectively. The points in the surfaces of the middle and right panels are the result of Monte Carlo evaluations of the NMSE exhibited by the discrete and continuous time TDRs, respectively. The left panel was constructed modeling the reservoir with an approximating VAR(1) model.

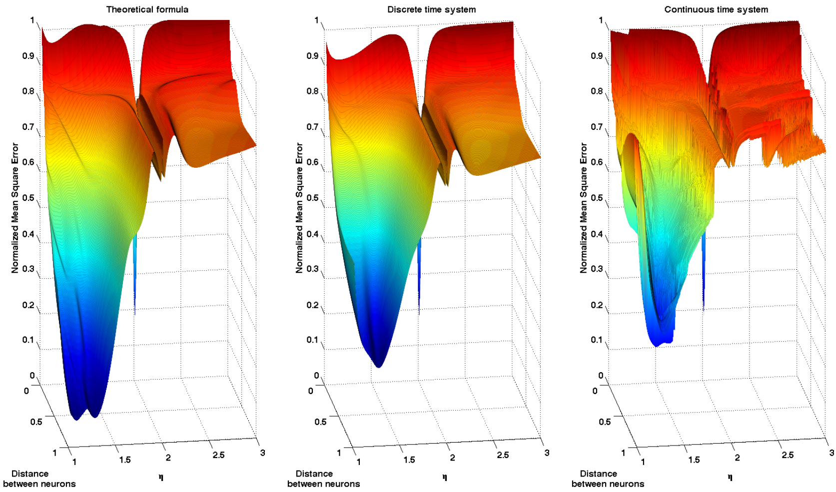


Figure 12. Error exhibited by a TDR computer with a Mackey-Glass kernel in a 6-lag quadratic memory task as a function of the separation between neurons d and the parameter η. The points in the surfaces of the middle and right panels are the result of Monte Carlo evaluations of the NMSE exhibited by the discrete and continuous time TDRs, respectively. The left panel was constructed modeling the reservoir with an approximating VAR(1) model.

Conclusions: *The quality of the approximation* (26) *at the time of evaluating the memory capacities of the original system is excellent and the resulting function (nonlinear capacity) can be hence used for RC optimization purposes regarding the intrinsic TDR architecture parameters* $\boldsymbol{\theta}$ *and the input mask* $\mathbf{c}$.

Perspectives

1. Modeling of the reservoir computing working principle and the design of optimal architectures
   - Extension to non-independent and multivariate signals
   - Theoretical treatment of classification problems
   - Modeling parallel reservoir computers [GHLO14] and their properties
   - Use of the reservoir model to establish the reservoir computing defining features
2. Technological implementation of optimal reservoir architectures
3. Applications to classification tasks for biomedical signals (like Hi-Res EEG)
4. Real-time information processing with reservoir computing

References I

L. Appeltant, M. C. Soriano, G. Van der Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, and I. Fischer.
Information processing using a single dynamical node as complex system.
Nature Communications, 2:468, January 2011.

Lyudmila Grigoryeva, Julie Henriques, Laurent Larger, and Juan-Pablo Ortega.
Stochastic time series forecasting using time-delay reservoir computers: performance and universality.
Neural Networks, 55:59–71, 2014.

Lyudmila Grigoryeva, Julie Henriques, Laurent Larger, and Juan-Pablo Ortega.
Optimal nonlinear information processing capacity in delay-based reservoir computers.
Scientific Reports, 5(12858):1–11, 2015.

Jack Hale.
Theory of Functional Differential Equations.
Springer-Verlag, 1977.

Kensuke Ikeda.
Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system.
Optics Communications, 30(2):257–261, August 1979.

Herbert Jaeger.
The 'echo state' approach to analysing and training recurrent neural networks.
German National Research Center for Information Technology, 2001.

Herbert Jaeger and Harald Haas.
Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication.
Science, 304(5667):78–80, 2004.

References II

N. N. Krasovskiy.
Stability of Motion.
Stanford University Press, 1963.

Helmut Lütkepohl.
New Introduction to Multiple Time Series Analysis.
Springer-Verlag, Berlin, 2005.

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

L. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutierrez, L. Pesquera, C. R. Mirasso, and I. Fischer.
Photonic information processing beyond Turing: an optoelectronic implementation of reservoir computing.
Optics Express, 20(3):3241, January 2012.

M. C. Mackey and L. Glass.
Oscillation and chaos in physiological control systems.
Science, 197:287–289, 1977.

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

A. N. Tikhonov.
On the stability of inverse problems.
Dokl. Akad. Nauk SSSR, 39(5):195–198, 1943.

References III

D. Verstraeten, B. Schrauwen, M. D'Haene, and D. Stroobandt.
An experimental unification of reservoir computing methods.
Neural Networks, 20:391–403, 2007.

Min Wu, Yong He, and Jin-Hua She.
Stability Analysis and Robust Control of Time-Delay Systems.
Springer, 2010.