

# Nonlinear memory capacity of parallel time-delay reservoir computers in the processing of multidimensional signals

Lyudmila Grigoryeva<sup>1</sup>, Julie Henriques<sup>2</sup>, Laurent Larger<sup>3</sup>, and Juan-Pablo Ortega<sup>4,\*</sup>

## Abstract

This paper addresses the reservoir design problem in the context of delay-based reservoir computers for multidimensional input signals, parallel architectures, and real-time multitasking. First, an approximating reservoir model is presented in those frameworks that provides an explicit functional link between the reservoir parameters and architecture and its performance in the execution of a specific task. Second, the inference properties of the ridge regression estimator in the multivariate context is used to assess the impact of finite sample training on the decrease of the reservoir capacity. Finally, an empirical study is conducted that shows the adequacy of the theoretical results with the empirical performances exhibited by various reservoir architectures in the execution of several nonlinear tasks with multidimensional inputs. Our results confirm the robustness properties of the parallel reservoir architecture with respect to task misspecification and parameter choice that had already been documented in the literature.

**Key Words:** Reservoir computing, echo state networks, liquid state machines, time-delay reservoir, parallel computing, memory capacity, multidimensional signals processing, Big Data.

## 1 Introduction

The recent and fast development of numerous massive data acquisition technologies results in a considerable growth of the data volumes that are stored and that need to be processed in the context of many human activities. The variability, complexity, and volume of this information have motivated the appearance of the generic term *Big Data*, which is mainly used to refer to datasets whose features make the traditional data processing approaches inadequate. This relatively new concept calls for the development of specialized tools for data preprocessing, analysis, transferring, and visualization, as well as for novel data mining and machine learning techniques in order to tackle specific computational tasks.

In this context, there is a recent but already well established paradigm for neural computation known by the name of **reservoir computing (RC)** [Jaeg 01, Jaeg 04, Maas 02, Maas 11, Croo 07, Vers 07,

<sup>1</sup>Department of Mathematics and Statistics. Universität Konstanz. Box 146. D-78457 Konstanz. Germany. Lyudmila.Grigoryeva@uni-konstanz.de

<sup>2</sup>CHRU Besançon. 2 place Saint Jacques. F-25000 Besançon. jhenriques@chu-besancon.fr

<sup>3</sup>FEMTO-ST, UMR CNRS 6174, Optics Department, Université de Franche-Comté, UFR des Sciences et Techniques. 15, Avenue des Montboucons. F-25000 Besançon cedex. France. Laurent.Larger@univ-fcomte.fr

<sup>4</sup>Corresponding author. Centre National de la Recherche Scientifique, Laboratoire de Mathématiques de Besançon, UMR CNRS 6623, Université de Franche-Comté, UFR des Sciences et Techniques. 16, route de Gray. F-25030 Besançon cedex. France. Juan-Pablo.Ortega@univ-fcomte.fr

Luko 09] (also referred to as *Echo State Networks* and *Liquid State Machines*), that has already shown a significant potential in successfully confronting some of the challenges that we just described.

This brain-inspired machine learning methodology exhibits several competitive advantages with respect to more traditional approaches. First, the supervised learning scheme associated to it is extremely simple. Second, some implementations of the RC paradigm are based on the computational capacities of certain dynamical systems [Crut 10] that open the door to physical realizations that have already been built using dedicated hardware (see, for instance, [Jaeg 07, Atiy 00, Appe 11, Roda 11, Larg 12, Paqu 12]) and that, recently, have shown unprecedented information processing speeds [Brun 13]. Our work takes place in the context of this specific type of RCs and, more explicitly, in the so called **time-delay reservoirs (TDRs)** that use the sampling of the solutions of time-delay differential equations in the construction of the RC.

Despite the outstanding empirical performances of TDRs described in the above listed references and the convenience of the learning scheme associated to them, a well-known important drawback is that these devices show a certain lack of structural task universality. More specifically, each task presented to a TDR requires that the TDR parameters and, more generally, its architecture are tuned in order to achieve optimal performance or, equivalently, small deviations from the optimal parameter values can seriously degrade the reservoir performance. The optimal parameters have been traditionally found by trial and error or by running costly numerical scannings for each task. More recently, in [Grig 15a] we introduced a method to overcome this difficulty by providing a functional link between the RC parameters and its performance with respect to a given task and that can be used to accurately determine the optimal reservoir architecture by solving a well structured optimization problem; this feature simplifies enormously the implementation effort and sheds new light on the mechanisms that govern this information processing technique.

This paper builds on the techniques introduced in [Grig 15a] and extends those results in the following directions:

- (i) The memory capacity formulas in [Grig 15a] are generalized to **multidimensional input signals** and we provide capacity estimations for the simultaneous execution of several memory tasks. This feature, sometimes referred to as **real-time multitasking** [Maas 11] is usually presented as one of the most prominent computational advantages of RC.
- (ii) We provide memory capacity estimations for **parallel arrays of reservoir computers**. This reservoir architecture has been introduced in [Orti 12, Grig 14] and has been empirically shown to exhibit improved robustness properties with respect to the dependence of the optimal reservoir parameters on the task presented to the device and also with respect to task misspecification.
- (iii) We carry out an in-depth study of the ridge regression estimator in the multivariate context in order to assess **the impact of finite sample training on the decrease of reservoir capacity**. More specifically, when the teaching signal used to train the RC has finite size, the faulty estimation of the RC readout layer (see Section 2.2) introduces an error that is cumulated with the characteristic error associated to the RC scheme and that we explicitly quantify.
- (iv) We conduct an empirical study that shows the adequacy of our theoretical results with the empirical performances exhibited by TDRs in the execution of various nonlinear tasks with multidimensional inputs. Additionally, we confirmed using the approximating model, the robustness properties of the parallel reservoir architecture with respect to task misspecification and parameter choice that had already been documented in [Grig 14].

The paper is organized as follows: Section 2 recalls the general setup for time-delay reservoir computing, as well as the notions of characteristic error and memory capacity in the multitasking setup.

Section 3 constitutes the core of the paper and addresses the points (i) through (iii) listed in the previous paragraphs. The empirical study described in point (iv) is contained in Section 4.

**Acknowledgments:** We acknowledge partial financial support of the Région de Franche-Comté (Convention 2013C-5493), the European project PHOCUS (FP7 Grant No. 240763), the ANR “BIPHOPROC” project (ANR-14-OHRI-0002-02), and Deployment S.L. LG acknowledges financial support from the Faculty for the Future Program of the Schlumberger Foundation.

## 2 Notation and preliminaries

In this section we introduce the notation that we use in the paper, we briefly recall the general setup for time-delay reservoirs (TDRs), and provide various preliminary concepts that are needed in the following sections.

### 2.1 Notation

Column vectors are denoted by bold lower or upper case symbol like  $\mathbf{v}$  or  $\mathbf{V}$ . We write  $\mathbf{v}^\top$  to indicate the transpose of  $\mathbf{v}$ . Given a vector  $\mathbf{v} \in \mathbb{R}^n$ , we denote its entries by  $v_i$ , with  $i \in \{1, \dots, n\}$ ; we also write  $\mathbf{v} = (v_i)_{i \in \{1, \dots, n\}}$ . The symbols  $\mathbf{1}_n$  and  $\mathbf{0}_n$  stand for the vectors of length  $n$  consisting of ones and zeros, respectively. We denote by  $\mathbb{M}_{n,m}$  the space of real  $n \times m$  matrices with  $m, n \in \mathbb{N}$ . When  $n = m$ , we use the symbol  $\mathbb{M}_n$  to refer to the space of square matrices of order  $n$ . Given a matrix  $A \in \mathbb{M}_{n,m}$ , we denote its components by  $A_{ij}$  and we write  $A = (A_{ij})$ , with  $i \in \{1, \dots, n\}$ ,  $j \in \{1, \dots, m\}$ . We write  $\mathbb{I}_n$  and  $\mathbb{O}_n$  to denote the identity matrix and the zero matrix of dimension  $n$ , respectively. We use  $\mathbb{S}_n$  to indicate the subspace  $\mathbb{S}_n \subset \mathbb{M}_n$  of symmetric matrices, that is,  $\mathbb{S}_n = \{A \in \mathbb{M}_n \mid A^\top = A\}$ . Given a matrix  $A \in \mathbb{M}_{n,m}$ , we denote by  $\text{vec}$  the operator that transforms  $A$  into a vector of length  $nm$  by stacking all its columns, namely,

$$\text{vec} : \mathbb{M}_{n,m} \longrightarrow \mathbb{R}^{nm}, \quad \text{vec}(A) = (A_{11}, \dots, A_{n1}, \dots, A_{1m}, \dots, A_{nm})^\top, \quad A \in \mathbb{M}_{n,m}.$$

When  $A$  is symmetric, we denote by  $\text{vech}$  the operator that stacks the elements on and below the main diagonal of  $A$  into a vector of length  $N := \frac{1}{2}n(n+1)$ , that is,

$$\text{vech} : \mathbb{S}_n \longrightarrow \mathbb{R}^N, \quad \text{vech}(A) = (A_{11}, \dots, A_{n1}, A_{22}, \dots, A_{n2}, \dots, A_{nn})^\top, \quad A \in \mathbb{S}_n.$$

Let  $N := \frac{1}{2}n(n+1)$ . We denote by  $L_n \in \mathbb{M}_{N,n^2}$  and by  $D_n \in \mathbb{M}_{n^2,N}$  the elimination and the duplication matrices [Lutk 05], respectively. These matrices satisfy that:

$$\text{vech}(A) = L_n \text{vec}(A), \quad \text{and} \quad \text{vec}(A) = D_n \text{vech}(A). \quad (2.1)$$

Consider  $A \in \mathbb{S}_n$ ,  $\mathbf{v} = \text{vech}(A) \in \mathbb{R}^N$ , and  $S = \{(i, j) \mid i, j \in \{1, \dots, n\}, i \geq j\}$ . Let  $\sigma : S \longrightarrow \{1, \dots, N\}$  be the operator that assigns to the position of the entry  $(i, j)$ ,  $i \geq j$ , of the matrix  $A$  the position of the corresponding element of  $\mathbf{v}$  in the  $\text{vech}$  representation. We refer to the inverse of this operator as  $\sigma^{-1} : \{1, \dots, N\} \longrightarrow S$ . The symbol  $\|A\|_{\text{Frob}}$  denotes the Frobenius norm of  $A \in \mathbb{M}_{m,n}$  defined as  $\|A\|_{\text{Frob}}^2 := \text{trace}(A^\top A)$  [Meye 00]. Finally, the symbols  $\mathbb{E}[\cdot]$ ,  $\text{var}(\cdot)$ , and  $\text{Cov}(\cdot, \cdot)$  denote the mathematical expectation, the variance, and the covariance, respectively.

### 2.2 The general setup for time-delay reservoir (TDR) computing

**The functional time-delay differential equations used for TDR computing.** The time-delay reservoirs studied in this paper are constructed by sampling the solutions of time-delay differential equations of the form

$$\dot{x}(t) = -x(t) + f(x(t - \tau), I(t), \boldsymbol{\theta}), \quad (2.2)$$

where  $f$  is a nonlinear smooth function that will be referred to as **nonlinear kernel**,  $\boldsymbol{\theta} \in \mathbb{R}^K$  is a vector that contains the parameters of the nonlinear kernel,  $\tau > 0$  is the **delay**,  $x(t) \in \mathbb{R}$ , and  $I(t) \in \mathbb{R}$  is an external forcing that makes (2.2) non-autonomous and that in our construction will be used as an inlet into the system for the signal that needs to be processed. We emphasize that the solution space of equation (2.2) is infinite dimensional since an entire function  $x \in C^1([-\tau, 0], \mathbb{R})$  needs to be specified in order to initialize it. The nonlinear kernel  $f$  is chosen based on the concrete physical implementation of the computing system that is envisioned. We consider two specific parametric sets of kernels that have already been explored in the literature, namely:

(i) The **Mackey-Glass** [Mack 77] nonlinear kernel:

$$f(x, I, \boldsymbol{\theta}) = \frac{\eta(x + \gamma I)}{1 + (x + \gamma I)^p}, \quad \boldsymbol{\theta} := (\gamma, \eta, p) \in \mathbb{R}^3, \quad (2.3)$$

which is used in electronics-based RC implementations [Appe 11].

(ii) The **Ikeda** [Ikeda 79] nonlinear kernel

$$f(x, I, \boldsymbol{\theta}) = \eta \sin^2(x + \gamma I + \phi), \quad \boldsymbol{\theta} := (\eta, \gamma, \phi) \in \mathbb{R}^3, \quad (2.4)$$

associated to the optical RC implementations [Larg 12].

For these specific choices of nonlinear kernel, the parameters  $\gamma$  and  $\eta$  are usually referred to as the **input** and **feedback gains**, respectively.

**Continuous and discrete-time approaches to multidimensional TDR computing.** We briefly recall the design of a TDR using the solutions of (2.2). The following constructions are discussed in detail in [Grig 15a]. TDRs are based on the sampling of the solutions of (2.2) when driven by an input forcing obtained out of the signal that needs to be processed. More specifically, let  $\mathbf{z}(t) \in \mathbb{R}^n$ ,  $t \in \mathbb{Z}$ , be an  $n$ -dimensional discrete-time **input signal**. This signal is, first, time and dimensionally multiplexed over a delay period by using an **input mask**  $C \in \mathbb{M}_{N,n}$  and by setting  $\mathbf{I}(t) := C\mathbf{z}(t)$ ,  $t \in \mathbb{Z}$ , where  $N$  is a design parameter called the **number of neurons** of each **reservoir layer**. The resulting discrete-time  $N$ -dimensional signal  $\mathbf{I}(t) \in \mathbb{R}^N$  is called **input forcing**.

We now consider two different constructions of the TDR depending on the way in which the solutions of the time-delay differential equation (2.2) are handled. The **continuous time TDR** is constructed as a collection of **neuron values**  $x_i(t)$  organized in **layers**  $\mathbf{x}(t) \in \mathbb{R}^N$  of  $N \in \mathbb{N}$  of **virtual neurons** each, parameterized by  $t \in \mathbb{Z}$ . The value  $x_i(t)$ , referred to as the  **$i$ th neuron value** of the  $t$ th layer  $\mathbf{x}(t)$  of the reservoir, is obtained by sampling a solution  $x(t)$  of (2.2) by setting

$$x_i(t) := x(t\tau - (N - i)d), \quad i \in \{1, \dots, N\}, \quad t \in \mathbb{Z}, \quad (2.5)$$

where  $d := \tau/N$  is referred to as the **separation between neurons**. The solution  $x(t)$  has been obtained by using an external forcing  $I(s)$  in (2.2) constructed out of the input forcing  $\mathbf{I}(t)$  as follows: given  $s \in \mathbb{R}$ , let  $t \in \mathbb{Z}$  and  $i \in \{2, \dots, N\}$  be the unique values such that  $s \in (t\tau - (N - i - 1)d, t\tau - (N - i)d]$  and that we use to define the external forcing as  $I(s) := (\mathbf{I}(t))_i$ .

The **discrete-time TDR** is constructed via the Euler time-discretization of (2.2) with an integration step of  $d := \tau/N$ . In this case, the neuron values are determined by the following recursions:

$$x_i(t) := e^{-\xi} x_{i-1}(t) + (1 - e^{-\xi}) f(x_i(t-1), (\mathbf{I}(t))_i, \boldsymbol{\theta}), \quad (2.6)$$

with  $x_0(t) := x_N(t-1)$ ,  $\xi := \log(1 + d)$ , and  $i \in \{1, \dots, N\}$ . In this case, the recursions (2.6) uniquely determine a smooth map  $F : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^K \rightarrow \mathbb{R}^N$  referred to as the **reservoir map** that specifies

the neuron values of a given  $t$ th layer as a recursion on the neuron values of the preceding layer  $t - 1$  via an expression of the form

$$\mathbf{x}(t) = F(\mathbf{x}(t - 1), \mathbf{I}(t), \boldsymbol{\theta}). \quad (2.7)$$

**The TDR memory capacity for real-time multitasking.** In this paper we study the performance of TDRs at the time of simultaneously performing several memory tasks (see Figure 1). This means that we will evaluate the ability of the TDR to reproduce a prescribed multidimensional nonlinear function of the input signal

$$H : \begin{array}{ccc} \mathbb{R}^{(h+1)n} & \longrightarrow & \mathbb{R}^q \\ \text{vec}(\mathbf{z}(t), \dots, \mathbf{z}(t - h)) & \longmapsto & \mathbf{y}(t), \end{array} \quad (2.8)$$

that we will call  $q$ -dimensional  $h$ -lag memory task for the  $n$ -dimensional input signal  $\{\mathbf{z}(t)\}_{t \in \mathbb{Z}}$ .

In the RC context, this task is performed by using a finite size realization of the input signal  $\{\mathbf{z}(-h + 1), \dots, \mathbf{z}(T)\}$  that is used to construct a  $q$ -dimensional **teaching signal**  $\{\mathbf{y}(1), \dots, \mathbf{y}(T)\}$  by setting  $\mathbf{y}(t) := H(\text{vec}(\mathbf{z}(t), \dots, \mathbf{z}(t - h)))$ . The teaching signal is subsequently used to determine a pair  $(W_{\text{out}}, \mathbf{a}_{\text{out}}) \in \mathbb{M}_{N,q} \times \mathbb{R}^q$  that performs the memory task as an affine combination of the reservoir outputs. The optimal pair  $(W_{\text{out}}, \mathbf{a}_{\text{out}})$  is obtained with a ridge regression that minimizes the regularized square error, that is,

$$(W_{\text{out}}, \mathbf{a}_{\text{out}}) = \arg \min_{W \in \mathbb{M}_{N,q}, \mathbf{a} \in \mathbb{R}^q} (\text{trace}(\mathbb{E}[(W^\top \cdot \mathbf{x}(t) + \mathbf{a} - \mathbf{y}(t))^\top (W^\top \cdot \mathbf{x}(t) + \mathbf{a} - \mathbf{y}(t))]) + \lambda \|W\|_{\text{Frob}}^2) \quad (2.9)$$

$$=: \arg \min_{W \in \mathbb{M}_{N,q}, \mathbf{a} \in \mathbb{R}^q} (\text{MSE}(W, \mathbf{a}) + \lambda \|W\|_{\text{Frob}}^2). \quad (2.10)$$

The optimal pair  $(W_{\text{out}}, \mathbf{a}_{\text{out}})$  that solves the ridge regression problem (2.10) is referred to as the **readout layer**. The ridge regularization parameter  $\lambda \in \mathbb{R}$  is usually tuned during the training phase via cross-validation. The explicit solution of the optimization problem (2.10) (see [Grig 15a] for the details) is given by

$$W_{\text{out}} = (\Gamma(0) + \lambda \mathbb{I}_N)^{-1} \text{Cov}(\mathbf{x}(t), \mathbf{y}(t)), \quad (2.11)$$

$$\mathbf{a}_{\text{out}} = \boldsymbol{\mu}_y - W_{\text{out}}^\top \boldsymbol{\mu}_x, \quad (2.12)$$

where  $\boldsymbol{\mu}_x := \mathbb{E}[\mathbf{x}(t)] \in \mathbb{R}^N$ ,  $\boldsymbol{\mu}_y := \mathbb{E}[\mathbf{y}(t)] \in \mathbb{R}^q$ ,  $\Gamma(0) := \text{Cov}(\mathbf{x}(t), \mathbf{x}(t)) \in \mathbb{S}_N$ , and  $\text{Cov}(\mathbf{x}(t), \mathbf{y}(t)) \in \mathbb{M}_{N,q}$ . Stationarity hypotheses are assumed on the teaching signal and the reservoir output so that the first and second order moments that we just listed are time-independent. The error committed by the reservoir when accomplishing the task  $H$  with the optimal readout will be referred to as its **characteristic error** and is given by the expression

$$\text{MSE}(W_{\text{out}}, \mathbf{a}_{\text{out}}) = \text{trace}(\text{Cov}(\mathbf{y}(t), \mathbf{y}(t)) - W_{\text{out}}^\top (\Gamma(0) + 2\lambda \mathbb{I}_N) W_{\text{out}}), \quad (2.13)$$

that can be encoded under the form of a **memory capacity**  $C_H(\boldsymbol{\theta}, C, \lambda)$  with values between zero and one that depends on the task  $H$  that is being tackled, the input mask  $C$ , the reservoir parameters  $\boldsymbol{\theta}$  and the regularization constant  $\lambda$ :

$$C_H(\boldsymbol{\theta}, C, \lambda) := 1 - \frac{\text{MSE}(W_{\text{out}}, \mathbf{a}_{\text{out}})}{\text{trace}(\text{Cov}(\mathbf{y}(t), \mathbf{y}(t)))} = \frac{\text{trace}(W_{\text{out}}^\top (\Gamma(0) + 2\lambda \mathbb{I}_N) W_{\text{out}})}{\text{trace}(\text{Cov}(\mathbf{y}(t), \mathbf{y}(t)))}, \quad (2.14)$$

where  $W_{\text{out}}$  is provided by the solution in (2.11). In order to evaluate (2.14) for a specific memory task, the expressions of  $\Gamma(0)$ ,  $\text{Cov}(\mathbf{x}(t), \mathbf{y}(t))$ , and  $\text{Cov}(\mathbf{y}(t), \mathbf{y}(t))$  need to be computed. The matrix

$\Gamma(0)$  depends exclusively on the input signal and the reservoir architecture but  $\text{Cov}(\mathbf{y}(t), \mathbf{y}(t))$  and  $\text{Cov}(\mathbf{x}(t), \mathbf{y}(t))$  are related to the specific memory task  $H$  at hand. The computation of (2.14) is in general very complicated and that is why in [Grig 15a] we introduced a simplified reservoir model that allowed us to efficiently evaluate it for one-dimensional statistically independent input signals and memory tasks. The extension of this theoretical tool to a multidimensional setup and to parallel architectures is one of the main goals of this paper.

Finally, there are situations in which the moments  $\mu_x$ ,  $\mu_y$ ,  $\Gamma(0)$ , and  $\text{Cov}(\mathbf{x}(t), \mathbf{y}(t))$ , necessary to compute the readout layer ( $W_{\text{out}}, \mathbf{a}_{\text{out}}$ ) using the equations (2.11)-(2.12), are obtained directly out of finite sample realizations of the teaching signal and of the reservoir output. The use in that context of finite sample empirical estimators carries in its wake an additional error that adds up to the characteristic error (2.13) and that we study later on in Section 3.4.

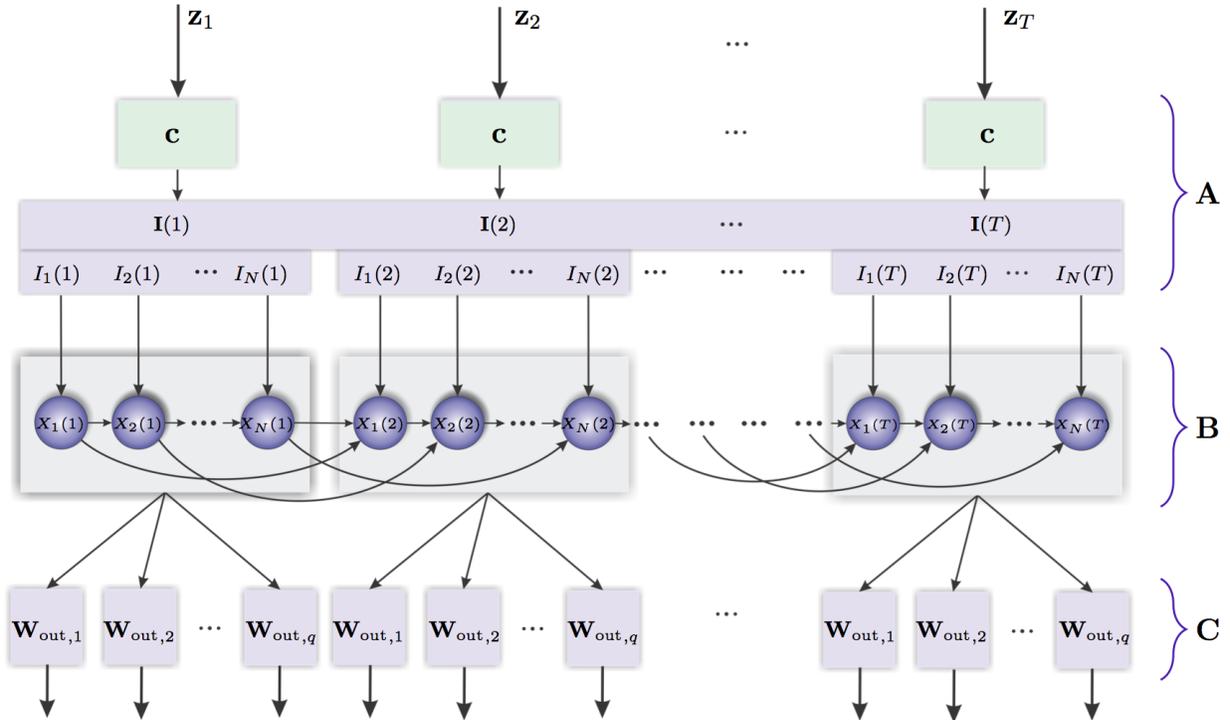


Figure 1: Diagram representing the architecture of a TDR reservoir with a multitask readout. The module A is the input layer, B is a neural diagram representing the discrete-time reservoir dynamics implied by equation (2.6), and C is the multitask readout layer in which each column of the matrix  $W_{\text{out}}$  accomplishes a different task based on the same reservoir output.

### 3 Reservoir memory capacities for parallel architectures and for multidimensional input signals and tasks

In this section we generalize the reservoir model proposed in [Grig 15a] in order to accommodate the treatment of multidimensional input signals and the computation of memory capacities associated to the execution of several simultaneous memory tasks, as well as the performance evaluation of parallel

reservoir architectures. The main virtue of the reservoir model is that it allows for the explicit computation of the different elements that constitute the capacity formula (2.14) making hence accessible its evaluation.

In the last section we study the dependence of the reservoir performance on the length of the teaching signal and the regularization regression parameter. In particular we formulate asymptotic expressions that provide an estimation of the added error that is committed in memory tasks when the training is incomplete due to the finiteness of the training sample.

The reservoir model introduced in [Grig 15a] is based on the observation that optimal reservoir performance is frequently attained when the reservoir is functioning in a neighborhood of an asymptotically stable equilibrium of the autonomous system associated to (2.2). This feature suggests the possibility of approximating the reservoir by its partial linearization at that stable fixed point with respect to the delayed self feedback but keeping the nonlinearity at the level of the input signal injection. This observation motivated in [Grig 15a] an in-depth study of the stability properties of the equilibria  $x_0$  of the time-delay differential equation (2.2) and of the corresponding fixed points  $\mathbf{x}_0 = x_0 \mathbf{i}_N \in \mathbb{R}^N$  of the discrete-time approximation (2.7), both considered in the autonomous regime, that is, when  $I(t) = 0$  in (2.2) and  $\mathbf{I}(t) = \mathbf{0}_N$  in (2.7), respectively. In particular, it was shown that (see Corollary D.5 and Theorem D.10 in the Supplementary Material in [Grig 15a]) that  $|\partial_x f(x_0, 0, \boldsymbol{\theta})| < 1$  is a sufficient condition for the asymptotic stability of  $x_0 \in \mathbb{R}$  and  $\mathbf{x}_0 = x_0 \mathbf{i}_N \in \mathbb{R}^N$  in the continuous and discrete-time cases, respectively, which given a particular kernel  $f$  allows for the identification of specific regions in parameter space in which stability is guaranteed (see Corollaries D.6 and D.7 of the Supplementary Material in [Grig 15a] for the Mackey-Glass and Ikeda kernel cases).

### 3.1 The reservoir model for multidimensional input signals

Consider a discrete-time TDR described by a reservoir map  $F : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^K \rightarrow \mathbb{R}^N$  as in (2.7). Let  $\mathbf{x}_0 \in \mathbb{R}^N$  be a stable fixed point of the autonomous systems associated to (2.7), that is,  $F(\mathbf{x}_0, \mathbf{0}_N, \boldsymbol{\theta}) = \mathbf{x}_0$ . In order to write down the approximate reservoir model as in [Grig 15a] we start by approximating (2.7) by its partial linearization at  $\mathbf{x}_0$  with respect to the delayed self feedback and by the  $R$ th-order Taylor series expansion on the input forcing  $\mathbf{I}(t) \in \mathbb{R}^N$ . We obtain the following expression:

$$\mathbf{x}(t) = F(\mathbf{x}_0, \mathbf{0}_N, \boldsymbol{\theta}) + A(\mathbf{x}_0, \boldsymbol{\theta})(\mathbf{x}(t-1) - \mathbf{x}_0) + \boldsymbol{\varepsilon}(t), \quad (3.1)$$

where  $F(\mathbf{x}_0, \mathbf{0}_N, \boldsymbol{\theta})$  is the reservoir map evaluated at the point  $(\mathbf{x}_0, \mathbf{0}_N, \boldsymbol{\theta})$  and  $A(\mathbf{x}_0, \boldsymbol{\theta}) := D_{\mathbf{x}}F(\mathbf{x}_0, \mathbf{0}_N, \boldsymbol{\theta})$  is the first derivative of  $F$  with respect to its first argument, computed at the point  $(\mathbf{x}_0, \mathbf{0}_N, \boldsymbol{\theta})$ . The vector  $\boldsymbol{\varepsilon}(t) \in \mathbb{R}^N$ ,  $t \in \mathbb{Z}$ , in (3.1) is obtained out of the Taylor series expansion of  $F(\mathbf{x}(t), \mathbf{I}(t), \boldsymbol{\theta})$  in (2.7) on  $\mathbf{I}(t)$  up to some fixed order  $R \in \mathbb{N}$ . For each  $r \in \{1, \dots, N\}$  its  $r$ th component can be written as

$$\varepsilon_r(t) = (1 - e^{-\xi}) \sum_{i=1}^R \frac{1}{i!} (\partial_I^{(i)} f)(x_0, 0, \boldsymbol{\theta}) \sum_{j=1}^r e^{-(r-j)\xi} I_j(t)^i, \quad (3.2)$$

where  $(\partial_I^{(i)} f)(x_0, 0, \boldsymbol{\theta})$  is the  $i$ th order partial derivative of the nonlinear reservoir kernel map  $f$  in (2.2) with respect to the second argument  $I(t)$  computed at the point  $(x_0, 0, \boldsymbol{\theta})$ . Finally,  $A(\mathbf{x}_0, \boldsymbol{\theta})$  is called the **connectivity matrix** of the reservoir at the point  $\mathbf{x}_0$  and has the following explicit form

$$A(\mathbf{x}_0, \boldsymbol{\theta}) = \begin{pmatrix} \Phi & 0 & \dots & 0 & e^{-\xi} \\ e^{-\xi}\Phi & \Phi & \dots & 0 & e^{-2\xi} \\ e^{-2\xi}\Phi & e^{-\xi}\Phi & \dots & 0 & e^{-3\xi} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e^{-(N-1)\xi}\Phi & e^{-(N-2)\xi}\Phi & \dots & e^{-\xi}\Phi & \Phi + e^{-N\xi} \end{pmatrix}, \quad (3.3)$$

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  $f$  in (2.2) with respect to the first argument and computed at the point  $(x_0, 0, \boldsymbol{\theta})$ .

Suppose now that the input signal is a collection of  $n$ -dimensional independent and identically distributed random variables  $\{\mathbf{z}(t)\}_{t \in \mathbb{Z}} \sim \text{IID}(\mathbf{0}_n, \Sigma_z)$ ,  $\Sigma_z \in \mathbb{S}_n^+$ , and that we take as input mask the matrix  $C \in \mathbb{M}_{N,n}$ . Since for each  $t \in \mathbb{Z}$  the input forcing  $\mathbf{I}(t) \in \mathbb{R}^N$  is constructed via the assignment  $\mathbf{I}(t) := C\mathbf{z}(t)$  we have that  $\{\mathbf{I}(t)\}_{t \in \mathbb{Z}} \sim \text{IID}(\mathbf{0}_N, \Sigma_I)$ , with  $\Sigma_I := C\Sigma_z C^\top$ . It follows then that  $I_j(t) = \sum_{k=1}^n C_{jk} z_k(t)$  which substituted in (3.2) yields that

$$\boldsymbol{\varepsilon}(t) = (1 - e^{-\xi}) \begin{pmatrix} V_R(\mathbf{z}(t), \{C_{1,\cdot}\}, x_0, \boldsymbol{\theta}) \\ V_R(\mathbf{z}(t), \{C_{j,\cdot}\}_{j \in \{1,2\}}, x_0, \boldsymbol{\theta}) \\ \vdots \\ V_R(\mathbf{z}(t), \{C_{j,\cdot}\}_{j \in \{1,\dots,N\}}, x_0, \boldsymbol{\theta}) \end{pmatrix}, \quad (3.4)$$

with the polynomials

$$V_R(\mathbf{z}(t), \{C_{j,\cdot}\}_{j \in \{1,\dots,r\}}, x_0, \boldsymbol{\theta}) := \sum_{i=1}^R (\partial_I^{(i)} f)(x_0, 0, \boldsymbol{\theta}) \sum_{j=1}^r e^{-(r-j)\xi} \sum_{k_1 + \dots + k_n = i} \frac{1}{k_1! \dots k_n!} \prod_{s=1}^n C_{js}^{k_s} \cdot \prod_{s=1}^n z_s(t)^{k_s}. \quad (3.5)$$

The symbol  $\{C_{j,\cdot}\}$ ,  $j \in \{1, \dots, N\}$ , denotes the set of all the entries in the  $j$ th row of the input mask matrix  $C$ . The assumption  $\{\mathbf{z}(t)\}_{t \in \mathbb{Z}} \sim \text{IID}(\mathbf{0}_n, \Sigma_z)$  implies that  $\{\boldsymbol{\varepsilon}(t)\}_{t \in \mathbb{Z}}$  is also a family of  $N$ -dimensional independent and identically distributed random variables with mean  $\boldsymbol{\mu}_\varepsilon$  and covariance matrix  $\Sigma_\varepsilon$  given by

$$(\boldsymbol{\mu}_\varepsilon)_r = (1 - e^{-\xi}) \sum_{i=1}^R (\partial_I^{(i)} f)(x_0, 0, \boldsymbol{\theta}) \sum_{j=1}^r e^{-(r-j)\xi} \sum_{k_1 + \dots + k_n = i} \frac{1}{k_1! \dots k_n!} \prod_{s=1}^n C_{js}^{k_s} \cdot \mu_{k_1, \dots, k_n}(\mathbf{z}), \quad (3.6)$$

where

$$\mu_{k_1, \dots, k_n}(\mathbf{z}(t)) := \mathbb{E} \left[ \prod_{s=1}^n z_s(t)^{k_s} \right] \quad (3.7)$$

denotes a higher-order moment of  $\mathbf{z}(t) \in \mathbb{R}^n$  whose existence we assume for values  $k_1, \dots, k_n$  such that  $k_1 + \dots + k_n \leq 2R$ . Additionally,  $\Sigma_\varepsilon := \mathbb{E}[(\boldsymbol{\varepsilon}(t) - \boldsymbol{\mu}_\varepsilon)(\boldsymbol{\varepsilon}(t) - \boldsymbol{\mu}_\varepsilon)^\top]$  has entries determined by the relation:

$$(\Sigma_\varepsilon)_{rs} = (1 - e^{-\xi})^2 \mathbb{E} \left[ V_R(\mathbf{z}(t), \{C_{j,\cdot}\}_{j \in \{1,\dots,r\}}, x_0, \boldsymbol{\theta}) \cdot V_R(\mathbf{z}(t), \{C_{j,\cdot}\}_{j \in \{1,\dots,s\}}, x_0, \boldsymbol{\theta}) \right] - (\boldsymbol{\mu}_\varepsilon)_r (\boldsymbol{\mu}_\varepsilon)_s, \quad r, s \in \{1, \dots, N\}, \quad (3.8)$$

where the first summand is computed by first multiplying the polynomials  $V_R(\mathbf{z}, \{C_{j,\cdot}\}_{j \in \{1,\dots,r\}}, x_0, \boldsymbol{\theta})$  and  $V_R(\mathbf{z}, \{C_{j,\cdot}\}_{j \in \{1,\dots,s\}}, x_0, \boldsymbol{\theta})$  on the variable  $\mathbf{z} \in \mathbb{R}^n$  and subsequently evaluating the resulting polynomial according to the following convention: any monomial of the form  $az_1^{k_1} \dots z_n^{k_n}$  is replaced by  $a\mu_{k_1, \dots, k_n}(\mathbf{z})$ .

A particular case in which the higher-order moments (3.7) can be readily computed is when  $\{\mathbf{z}(t)\}_{t \in \mathbb{Z}} \sim \text{IN}(\mathbf{0}_n, \Sigma_z)$ , that is,  $\{\mathbf{z}(t)\}_{t \in \mathbb{Z}}$  follows an  $n$ -dimensional multivariate normal distribution. Indeed, following [Holm 88, Tria 03], let  $k_1, \dots, k_n \in \mathbb{N}$  be  $n$  nonzero natural numbers such that  $K := k_1 + k_2 + \dots + k_n$  and let  $\mathbf{z}^{\mathcal{K}} := (z_1 \mathbf{i}_{k_1}^\top, z_2 \mathbf{i}_{k_2}^\top, \dots, z_n \mathbf{i}_{k_n}^\top)^\top \in \mathbb{R}^K$ . The vector  $\mathbf{z}^{\mathcal{K}} \in \mathbb{R}^K$  is Gaussian with zero mean

and covariance matrix  $\Sigma_z^K \in \mathbb{S}_K$  given by  $(\Sigma_z^K)_{ij} = \text{Cov}(z_i^K, z_j^K)$ , for any  $i, j \in \{1, \dots, K\}$ , that is,  $\mathbf{z}^K \sim \text{IN}(\mathbf{0}_K, \Sigma_z^K)$ . Then, using Theorem 1 in [Tri03], we can write that

$$\mu_{k_1, \dots, k_n}(\mathbf{z}) = \begin{cases} Hf(\Sigma_z^K), & \text{when } K = 2l, \quad l \in \mathbb{N}, \\ 0, & \text{otherwise,} \end{cases} \quad (3.9)$$

where the symbol  $Hf(\Sigma_z^K)$  denotes the hafnian of the covariance matrix  $\Sigma_z^K$  of order  $2l$ ,  $l \in \mathbb{N}$ , defined by

$$Hf(\Sigma_z^K) := \sum_{I, J} (\Sigma_z^K)_{i'_1 j'_1} (\Sigma_z^K)_{i'_2 j'_2} \cdots (\Sigma_z^K)_{i'_l j'_l}, \quad (3.10)$$

where the sum is running over all the possible decompositions of  $\{1, 2, \dots, 2l = K\}$  into disjoint subsets  $I, J$  of the form  $I = \{i'_1, \dots, i'_l\}$ ,  $J = \{j'_1, \dots, j'_l\}$ , such that  $i'_1 < \dots < i'_l$ ,  $j'_1 < \dots < j'_l$ , and  $i'_w < j'_w$ , for each  $w \in \{1, \dots, l\}$ .

We now proceed as in [Grig15a] and consider (3.1) as a VAR(1) model [Lutk05] driven by the independent noise  $\{\varepsilon(t)\}_{t \in \mathbb{Z}}$ . If we assume that the nonlinear kernel  $f$  satisfies the stability condition  $|\partial_x f(x_0, 0, \boldsymbol{\theta})| < 1$ , then the proof of Theorem D.10 in [Grig15a] shows that the spectral radius  $\rho(A(\mathbf{x}_0, \boldsymbol{\theta})) < 1$ , which implies in turn that (3.1) has a unique causal and second order stationary solution [Lutk05, Proposition 2.1]  $\{\mathbf{x}(t)\}_{t \in \mathbb{Z}}$  with time-independent mean

$$\boldsymbol{\mu}_x = (I_N - A(\mathbf{x}_0, \boldsymbol{\theta}))^{-1} (F(\mathbf{x}_0, \mathbf{0}_N, \boldsymbol{\theta}) - A(\mathbf{x}_0, \boldsymbol{\theta})\mathbf{x}_0 + \boldsymbol{\mu}_\varepsilon). \quad (3.11)$$

The model (3.1) can hence be rewritten in mean-adjusted form as

$$\mathbf{x}(t) - \boldsymbol{\mu}_x = A(\mathbf{x}_0, \boldsymbol{\theta})(\mathbf{x}(t-1) - \boldsymbol{\mu}_x) + (\varepsilon(t) - \boldsymbol{\mu}_\varepsilon). \quad (3.12)$$

Additionally, the autocovariance function  $\Gamma(k) := \mathbb{E} \left[ (\mathbf{x}(t) - \boldsymbol{\mu}_x) (\mathbf{x}(t-k) - \boldsymbol{\mu}_x)^\top \right]$  at lag  $k \in \mathbb{Z}$  is determined by the Yule-Walker equations [Lutk05], which have the following solutions in vectorized form:

$$\text{vech}(\Gamma(0)) = (\mathbb{I}_{N'} - L_N(A(\mathbf{x}_0, \boldsymbol{\theta}) \otimes A(\mathbf{x}_0, \boldsymbol{\theta})))^{-1} \text{vech}(\Sigma_\varepsilon), \quad (3.13)$$

$$\Gamma(k) = A(\mathbf{x}_0, \boldsymbol{\theta})\Gamma(k-1) \quad \text{with} \quad \Gamma(-k) = \Gamma(k)^\top, \quad (3.14)$$

where  $N' := \frac{1}{2}N(N+1)$  and  $L_N \in \mathbb{M}_{N', N^2}$ ,  $D_N \in \mathbb{M}_{N^2, N'}$  are the elimination and the duplication matrices, respectively. We recall that the autocovariance function  $\Gamma(0)$  is one of the key components of the capacity formula (2.14) that we intend to explicitly evaluate.

### 3.2 The reservoir model for parallel TDRs with multidimensional input signals

In this section we generalize the reservoir model that we just developed to the parallel time-delay reservoir architecture that introduced in [Orti12, Grig14]. This reservoir design has shown very satisfactory robustness properties with respect to model misspecification and parameter choice. The basic idea on which this approach is built consists of presenting the input signal to a parallel array of reservoirs, each of them running with different parameter values. The concatenation of the outputs of these reservoirs is then used to construct a single readout layer via a ridge regression. Figure 2 provides a diagram representing the parallel reservoir computing architecture.

**Discrete-time description of the parallel TDRs.** Consider a parallel array of  $p$  time-delay reservoirs as in Figure 2. For each  $j \in \{1, \dots, p\}$ , the  $j$ th time-delay reservoir is based on a time-delay

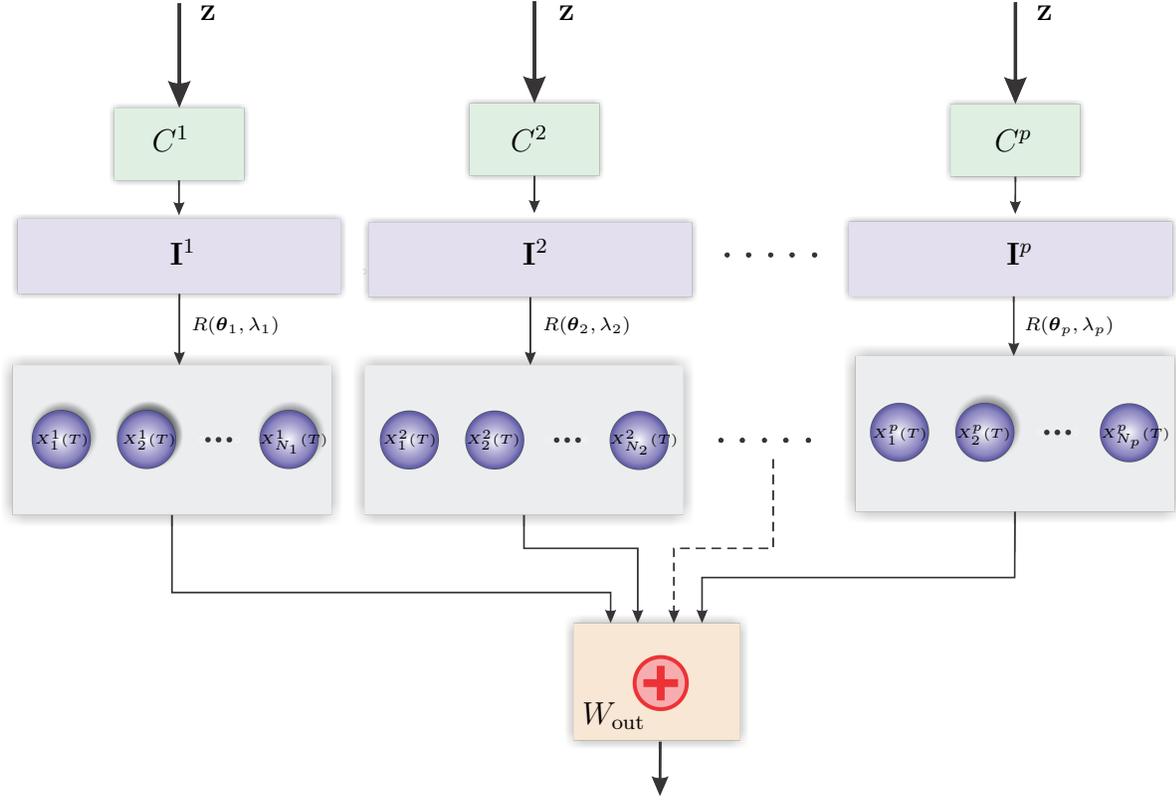


Figure 2: Diagram representing the architecture of a parallel reservoir computer.

differential equation like (2.2), has an associated nonlinear kernel  $f^{(j)}$  that depends on the parameters vector  $\boldsymbol{\theta}^{(j)} \in \mathbb{R}^{K_j}$  and the time-delay  $\tau_j > 0$ , namely

$$\dot{x}(t) = -x(t) + f^{(j)}(x(t - \tau_j), I(t), \boldsymbol{\theta}^{(j)}). \quad (3.15)$$

Let  $N_j \in \mathbb{N}$  be the number of the virtual neurons of the  $j$ th reservoir and let  $d_j := \tau_j/N_j$  be the corresponding separation between neurons. Let  $N^* := \sum_{j=1}^p N_j$  and  $K^* := \sum_{j=1}^p K_j$  be the total number of virtual neurons and the total number of parameters of the parallel array, respectively. The discrete-time description of the parallel array of  $p$  TDRs with total number of neurons  $N^*$  is obtained by Euler time-discretizing each of the differential equations (3.15) with integration step  $d_j$  and by organizing the solutions in neural layers described by the following recursions

$$x_i^{(j)}(t) := e^{-\xi^{(j)}} x_{i-1}^{(j)}(t) + (1 - e^{-\xi^{(j)}}) f^{(j)}(x_i^{(j)}(t-1), (\mathbf{I}^{(j)}(t))_i, \boldsymbol{\theta}^{(j)}), \quad i \in \{1, \dots, N_j\}, j \in \{1, \dots, p\} \quad (3.16)$$

with  $\xi^{(j)} := \log(1 + d_j)$  and using the convention  $x_0^{(j)}(t) := x_{N_j}^{(j)}(t-1)$ . The different  $p$  input forcings  $\mathbf{I}^{(j)}(t) \in \mathbb{R}^{N_j}$ ,  $j \in \{1, \dots, p\}$ , are created out of the  $n$ -dimensional input signal  $\{\mathbf{z}(t)\}_{t \in \mathbb{Z}}$  by using  $p$  input masks  $C^{(j)} \in \mathbb{M}_{N_j, n}$  and by setting  $\mathbf{I}^{(j)}(t) := C^{(j)} \mathbf{z}(t)$ .

Consider now  $\mathbf{X}(t) = (\mathbf{x}^{(1)}(t), \dots, \mathbf{x}^{(p)}(t)) \in \mathbb{R}^{N^*}$ , where  $\mathbf{x}^{(j)}(t) \in \mathbb{R}^{N_j}$ ,  $j \in \{1, \dots, p\}$ , is the neuron layer at time  $t$  corresponding to the  $j$ th individual TDR. As in the case of the individually operating time-delay reservoir in Section 3.1, the recursions (3.16) uniquely determine reservoir maps  $F^{(j)} : \mathbb{R}^{N_j} \times \mathbb{R}^{N_j} \times \mathbb{R}^{K_j} \rightarrow \mathbb{R}^{N_j}$ , constructed out of the associated nonlinear kernels  $f^{(j)}$ ,  $j \in \{1, \dots, p\}$

that can be put together to determine the map

$$\begin{pmatrix} \mathbf{x}^{(1)}(t) \\ \mathbf{x}^{(2)}(t) \\ \vdots \\ \mathbf{x}^{(p)}(t) \end{pmatrix} = \begin{pmatrix} F^{(1)}(\mathbf{x}^{(1)}(t-1), \mathbf{I}^{(1)}(t), \boldsymbol{\theta}^{(1)}) \\ F^{(2)}(\mathbf{x}^{(2)}(t-1), \mathbf{I}^{(2)}(t), \boldsymbol{\theta}^{(2)}) \\ \vdots \\ F^{(p)}(\mathbf{x}^{(p)}(t-1), \mathbf{I}^{(p)}(t), \boldsymbol{\theta}^{(p)}) \end{pmatrix}. \quad (3.17)$$

that can be rewritten as

$$\mathbf{X}(t) = F(\mathbf{X}(t-1), \mathbf{I}(t), \boldsymbol{\Theta}), \quad (3.18)$$

where  $F: \mathbb{R}^{N^*} \times \mathbb{R}^{N^*} \times \mathbb{R}^{K^*} \rightarrow \mathbb{R}^{N^*}$  is referred to as the **parallel reservoir map**,  $\mathbf{I}(t) := (\mathbf{I}^{(1)}(t), \dots, \mathbf{I}^{(p)}(t))$ , and  $\boldsymbol{\Theta} := (\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(p)}) \in \mathbb{R}^{K^*}$  is the vector containing all the parameters of the parallel array of TDRs. Parallel TDRs based on the recursion (3.18) are referred to in the sequel as **discrete-time parallel TDRs**.

We now generalize to the parallel context the reservoir model that we introduced in Section 3.1. We start by choosing  $p$  stable equilibria  $x_0^{(j)}$  of the dynamical systems (3.15) or, equivalently,  $p$  fixed points of the form  $\mathbf{x}_0^{(j)} := x_0^{(j)} \mathbf{i}_{N_j}$  of each of the reservoir maps in (3.17). These fixed points determine a fixed point  $\mathbf{X}_0 := (\mathbf{x}_0^{(1)}, \dots, \mathbf{x}_0^{(p)}) \in \mathbb{R}^{N^*}$  of the parallel array in (3.18). Now, as in Section 3.1 we partially linearize (3.18) at  $\mathbf{X}_0$  and use a higher  $R$ th-order Taylor series expansion on the forcing  $\mathbf{I}(t) \in \mathbb{R}^{N^*}$ . Analogously to the single reservoir case, we obtain that

$$\mathbf{X}(t) = F(\mathbf{X}_0, \mathbf{0}_{N^*}, \boldsymbol{\Theta}) + A(\mathbf{X}_0, \boldsymbol{\Theta})(\mathbf{X}(t-1) - \mathbf{X}_0) + \boldsymbol{\varepsilon}^{(\mathbf{X}_0, \boldsymbol{\Theta})}(t), \quad (3.19)$$

where  $A(\mathbf{X}_0, \boldsymbol{\Theta}) := D_{\mathbf{X}}F(\mathbf{X}_0, \mathbf{0}_{N^*}, \boldsymbol{\Theta})$  is the **parallel reservoir connectivity matrix**, which is the first derivative of  $F$  with respect to its first argument and evaluated at the point  $(\mathbf{X}_0, \boldsymbol{\Theta})$ , and

$$\boldsymbol{\varepsilon}(t)^{(\mathbf{X}_0, \boldsymbol{\Theta})} := \begin{pmatrix} \boldsymbol{\varepsilon}(t)^{(\mathbf{x}_0^{(1)}, \boldsymbol{\theta}^{(1)})} \\ \boldsymbol{\varepsilon}(t)^{(\mathbf{x}_0^{(2)}, \boldsymbol{\theta}^{(2)})} \\ \vdots \\ \boldsymbol{\varepsilon}(t)^{(\mathbf{x}_0^{(p)}, \boldsymbol{\theta}^{(p)})} \end{pmatrix} \in \mathbb{R}^{N^*} \quad (3.20)$$

with

$$\boldsymbol{\varepsilon}(t)^{(\mathbf{x}_0^{(j)}, \boldsymbol{\theta}^{(j)})} := (1 - e^{-\xi^{(j)}}) \begin{pmatrix} V_R^{(j)}(\mathbf{z}(t), \{C_{1,\cdot}^{(j)}\}, x_0^{(j)}, \boldsymbol{\theta}^{(j)}) \\ V_R^{(j)}(\mathbf{z}(t), \{C_{i,\cdot}^{(j)}\}_{i \in \{1,2\}}, x_0^{(j)}, \boldsymbol{\theta}^{(j)}) \\ \vdots \\ V_R^{(j)}(\mathbf{z}(t), \{C_{i,\cdot}^{(j)}\}_{i \in \{1, \dots, N_j\}}, x_0^{(j)}, \boldsymbol{\theta}^{(j)}) \end{pmatrix} \in \mathbb{R}^{N_j}, \quad j \in \{1, \dots, p\}, \quad (3.21)$$

where the polynomials  $V_R^{(j)}$  are defined as in (3.5). The assumption that the input signal  $\{\mathbf{z}(t)\}_{t \in \mathbb{Z}}$  is a family of  $n$ -dimensional independent and identically distributed random variables implies that the same property holds for the family  $\{\boldsymbol{\varepsilon}^{(\mathbf{X}_0, \boldsymbol{\Theta})}(t)\}_{t \in \mathbb{Z}}$  of  $N^*$ -dimensional random variables in (3.19), namely, that  $\{\boldsymbol{\varepsilon}^{(\mathbf{X}_0, \boldsymbol{\Theta})}(t)\}_{t \in \mathbb{Z}} \sim \text{IID}(\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}^{(\mathbf{X}_0, \boldsymbol{\Theta})}, \Sigma_{\boldsymbol{\varepsilon}}^{(\mathbf{X}_0, \boldsymbol{\Theta})})$ . The mean  $\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}^{(\mathbf{X}_0, \boldsymbol{\Theta})}$  can be written as

$$\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}^{(\mathbf{X}_0, \boldsymbol{\Theta})} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{\varepsilon}}^{(\mathbf{x}_0^{(1)}, \boldsymbol{\theta}^{(1)})} \\ \boldsymbol{\mu}_{\boldsymbol{\varepsilon}}^{(\mathbf{x}_0^{(2)}, \boldsymbol{\theta}^{(2)})} \\ \vdots \\ \boldsymbol{\mu}_{\boldsymbol{\varepsilon}}^{(\mathbf{x}_0^{(p)}, \boldsymbol{\theta}^{(p)})} \end{pmatrix}, \quad (3.22)$$

where  $\boldsymbol{\mu}_\varepsilon^{(\mathbf{x}_0^{(j)}, \boldsymbol{\theta}^{(j)})} := E \left[ \boldsymbol{\varepsilon}(t)(\mathbf{x}_0^{(j)}, \boldsymbol{\theta}^{(j)}) \right] \in \mathbb{R}^{N_j}$ ,  $j \in \{1, \dots, p\}$ , whose components are determined by an expression of the form (3.6), that is,

$$\begin{aligned} \left( \boldsymbol{\mu}_\varepsilon^{(\mathbf{x}_0^{(j)}, \boldsymbol{\theta}^{(j)})} \right)_r &= (1 - e^{-\xi^{(j)}}) \sum_{i=1}^R (\partial_I^{(i)} f^{(j)})(x_0^{(j)}, 0, \boldsymbol{\theta}^{(j)}) \sum_{j'=1}^r e^{-(r-j')\xi^{(j)}} \\ &\times \sum_{k_1 + \dots + k_n = i} \frac{1}{k_1! \dots k_n!} \prod_{s=1}^n (C_{j's}^{(j)})^{k_s} \cdot \mu_{k_1, \dots, k_n}(\mathbf{z}), \quad r \in \{1, \dots, N_j\}. \end{aligned} \quad (3.23)$$

Additionally, the covariance matrix  $\Sigma_\varepsilon^{(\mathbf{X}_0, \boldsymbol{\Theta})} := E \left[ (\boldsymbol{\varepsilon}^{(\mathbf{X}_0, \boldsymbol{\Theta})}(t) - \boldsymbol{\mu}_\varepsilon^{(\mathbf{X}_0, \boldsymbol{\Theta})})(\boldsymbol{\varepsilon}^{(\mathbf{X}_0, \boldsymbol{\Theta})}(t) - \boldsymbol{\mu}_\varepsilon^{(\mathbf{X}_0, \boldsymbol{\Theta})})^\top \right]$  can be written as

$$\Sigma_\varepsilon^{(\mathbf{X}_0, \boldsymbol{\Theta})} = \begin{pmatrix} \Sigma_\varepsilon^{(\mathbf{x}_0^{(1)}, \boldsymbol{\theta}^{(1)}), (\mathbf{x}_0^{(1)}, \boldsymbol{\theta}^{(1)})} & \dots & \Sigma_\varepsilon^{(\mathbf{x}_0^{(1)}, \boldsymbol{\theta}^{(1)}), (\mathbf{x}_0^{(p)}, \boldsymbol{\theta}^{(p)})} \\ \vdots & \ddots & \vdots \\ \Sigma_\varepsilon^{(\mathbf{x}_0^{(p)}, \boldsymbol{\theta}^{(p)}), (\mathbf{x}_0^{(1)}, \boldsymbol{\theta}^{(1)})} & \dots & \Sigma_\varepsilon^{(\mathbf{x}_0^{(p)}, \boldsymbol{\theta}^{(p)}), (\mathbf{x}_0^{(p)}, \boldsymbol{\theta}^{(p)})} \end{pmatrix}, \quad (3.24)$$

where each block  $\Sigma_\varepsilon^{(\mathbf{x}_0^{(i)}, \boldsymbol{\theta}^{(i)}), (\mathbf{x}_0^{(j)}, \boldsymbol{\theta}^{(j)})} \in \mathbb{M}_{N_i, N_j}$ ,  $i, j \in \{1, \dots, p\}$  represents the covariance between the innovation components that drive the  $i$ th and the  $j$ th time-delay reservoirs, respectively, and which has entries determined by:

$$\begin{aligned} (\Sigma_\varepsilon^{(\mathbf{x}_0^{(i)}, \boldsymbol{\theta}^{(i)}), (\mathbf{x}_0^{(j)}, \boldsymbol{\theta}^{(j)})})_{rs} &= (1 - e^{-\xi^{(i)}})(1 - e^{-\xi^{(j)}}) E \left[ V_R^{(i)} \left( \mathbf{z}(t), \left\{ C_{j', \cdot}^{(i)} \right\}_{j' \in \{1, \dots, r\}}, x_0^{(i)}, \boldsymbol{\theta}^{(i)} \right) \right. \\ &\quad \times \left. V_R^{(j)} \left( \mathbf{z}(t), \left\{ C_{j', \cdot}^{(j)} \right\}_{j' \in \{1, \dots, s\}}, x_0^{(j)}, \boldsymbol{\theta}^{(j)} \right) \right] \\ &\quad - (\boldsymbol{\mu}_\varepsilon^{(\mathbf{x}_0^{(i)}, \boldsymbol{\theta}^{(i)})})_r (\boldsymbol{\mu}_\varepsilon^{(\mathbf{x}_0^{(j)}, \boldsymbol{\theta}^{(j)})})_s, \quad r \in \{1, \dots, N_i\}, \quad s \in \{1, \dots, N_j\}, \end{aligned} \quad (3.25)$$

where the first summand is computed using the same approach that we described in expression (3.8).

The connectivity matrix can be easily written in terms of the connectivity matrices of each of the reservoirs that make up the parallel pool as

$$A(\mathbf{X}_0, \boldsymbol{\Theta}) := D_{\mathbf{X}} F(\mathbf{X}_0, \mathbf{0}_{N^*}, \boldsymbol{\Theta}) = \begin{pmatrix} A^{(1)}(\mathbf{x}_0^{(1)}, \boldsymbol{\theta}^{(1)}) & \mathbb{O}_{N_1, N_2} & \dots & \mathbb{O}_{N_1, N_p} \\ \mathbb{O}_{N_2, N_1} & A^{(2)}(\mathbf{x}_0^{(2)}, \boldsymbol{\theta}^{(2)}) & \dots & \mathbb{O}_{N_2, N_p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{O}_{N_p, N_1} & \mathbb{O}_{N_p, N_2} & \dots & A^{(p)}(\mathbf{x}_0^{(p)}, \boldsymbol{\theta}^{(p)}) \end{pmatrix}, \quad (3.26)$$

where for each  $j \in \{1, \dots, p\}$  the matrix  $A^{(j)}(\mathbf{x}_0^{(j)}, \boldsymbol{\theta}^{(j)}) := D_{\mathbf{X}} F^{(j)}(\mathbf{x}_0^{(j)}, \mathbf{0}_{N_j}, \boldsymbol{\theta}^{(j)})$  is the connectivity matrix of the  $j$ th TDR, determined as the first derivative of the corresponding  $j$ th reservoir map  $F^{(j)}$  with respect to its first argument, evaluated at the point  $(\mathbf{x}_0^{(j)}, \mathbf{0}_{N_j}, \boldsymbol{\theta}^{(j)})$ . Each of those individual connectivity matrices  $A^{(j)}(\mathbf{x}_0^{(j)}, \boldsymbol{\theta}^{(j)})$  has the explicit form provided in (3.3). Moreover, if for each of the individual equilibria  $x_0^{(j)}$  used in the construction we require the stability condition  $|\partial_x f^{(j)}(x_0^{(j)}, 0, \boldsymbol{\theta})| < 1$ , then by the proof of Theorem D.10 in [Grig 15a] we have that the spectral radii  $\rho(A^{(j)}(\mathbf{x}_0^{(j)}, \boldsymbol{\theta}^{(j)})) < 1$  and, consequently

$$\rho(A(\mathbf{X}_0, \boldsymbol{\Theta})) < 1. \quad (3.27)$$

In these conditions, (3.19) determines a VAR(1) model driven by the noise  $\{\boldsymbol{\varepsilon}^{(\mathbf{X}_0, \boldsymbol{\Theta})}(t)\}_{t \in \mathbb{Z}} \sim \text{IID}(\boldsymbol{\mu}_\varepsilon^{(\mathbf{X}_0, \boldsymbol{\Theta})}, \Sigma_\varepsilon^{(\mathbf{X}_0, \boldsymbol{\Theta})})$  and that has a unique causal and second order stationary solution  $\{\mathbf{X}(t)\}_{t \in \mathbb{Z}}$

with time-independent mean

$$\boldsymbol{\mu}_X^{(\mathbf{X}_0, \boldsymbol{\Theta})} = \begin{pmatrix} \boldsymbol{\mu}_x^{(\mathbf{x}_0^{(1)}, \boldsymbol{\theta}^{(1)})} \\ \boldsymbol{\mu}_x^{(\mathbf{x}_0^{(2)}, \boldsymbol{\theta}^{(2)})} \\ \vdots \\ \boldsymbol{\mu}_x^{(\mathbf{x}_0^{(p)}, \boldsymbol{\theta}^{(p)})} \end{pmatrix}, \quad (3.28)$$

with

$$\boldsymbol{\mu}_x^{(\mathbf{x}_0^{(j)}, \boldsymbol{\theta}^{(j)})} = (\mathbb{I}_{N_j} - A^{(j)}(\mathbf{x}_0^{(j)}, \boldsymbol{\theta}^{(j)}))^{-1} (F^{(j)}(\mathbf{x}_0^{(j)}, \mathbf{0}_{N_j}, \boldsymbol{\theta}^{(j)}) - A^{(j)}(\mathbf{x}_0^{(j)}, \boldsymbol{\theta}^{(j)}) \mathbf{x}_0^{(j)} + \boldsymbol{\mu}_\varepsilon^{(\mathbf{x}_0^{(j)}, \boldsymbol{\theta}^{(j)})}). \quad (3.29)$$

This allows us to write the parallel reservoir model (3.19) in mean-adjusted form as

$$\mathbf{X}(t) - \boldsymbol{\mu}_X^{(\mathbf{X}_0, \boldsymbol{\Theta})} = A(\mathbf{X}_0, \boldsymbol{\Theta})(\mathbf{X}(t-1) - \boldsymbol{\mu}_X^{(\mathbf{X}_0, \boldsymbol{\Theta})}) + (\varepsilon(t)^{(\mathbf{X}_0, \boldsymbol{\Theta})} - \boldsymbol{\mu}_\varepsilon^{(\mathbf{X}_0, \boldsymbol{\Theta})}). \quad (3.30)$$

Finally, the autocovariance function  $\Gamma(k) := \mathbb{E}[(\mathbf{X}(t) - \boldsymbol{\mu}_X^{(\mathbf{X}_0, \boldsymbol{\Theta})})(\mathbf{X}(t-k) - \boldsymbol{\mu}_X^{(\mathbf{X}_0, \boldsymbol{\Theta})})^\top]$  of  $\{\mathbf{X}(t)\}_{t \in \mathbb{Z}}$  at lag  $k \in \mathbb{Z}$  is determined by the Yule-Walker equations [Lutk 05] whose solutions in vectorized form are:

$$\text{vech}(\Gamma(0)) = (\mathbb{I}_{N^{*'}} - L_{N^*}(A(\mathbf{X}_0, \boldsymbol{\Theta}) \otimes A(\mathbf{X}_0, \boldsymbol{\Theta})) D_{N^*})^{-1} \text{vech}(\Sigma_\varepsilon^{(\mathbf{X}_0, \boldsymbol{\Theta})}), \quad (3.31)$$

$$\Gamma(k) = A(\mathbf{X}_0, \boldsymbol{\Theta}) \Gamma(k-1) \quad \text{with} \quad \Gamma(-k) = \Gamma(k)^\top, \quad (3.32)$$

where  $N^{*'} := \frac{1}{2} N^*(N^* + 1)$  and  $L_{N^*} \in \mathbb{M}_{N^{*'}, N^{*2}}$ ,  $D_{N^*} \in \mathbb{M}_{N^{*2}, N^{*'}}$  are the elimination and the duplication matrices, respectively.

### 3.3 Memory capacity estimations for multidimensional memory tasks

In what follows we explain how to use the reservoir models (3.12) and (3.30) presented in the previous two sections as well as their dynamical features in order to explicitly compute the reservoir capacities (2.14) associated to different memory tasks. As we explained in (2.8), a memory task is determined by a map

$$H : \begin{array}{ccc} \mathbb{R}^{(h+1)n} & \longrightarrow & \mathbb{R}^q \\ \text{vec}(\mathbf{z}(t), \dots, \mathbf{z}(t-h)) & \mapsto & \mathbf{y}(t), \end{array} \quad (3.33)$$

with  $q, h \in \mathbb{N}$ , that is made out of  $q$  different real valued functions of the input signal,  $h$  time steps into the past. The reservoir memory capacity  $C_H(\boldsymbol{\theta}, C, \lambda)$  associated to  $H$  measures the ability of the reservoir with parameters  $\boldsymbol{\theta}$  to recover that function after being trained using a teaching signal.

In the next paragraphs we place ourselves in the context of a parallel array of  $p$  time delay reservoirs as in Figure 2, with collective nonlinear kernel parameters  $\boldsymbol{\Theta} \in \mathbb{R}^{K^*}$  and operating in the neighbourhood of a stable fixed point  $\mathbf{X}_0 \in \mathbb{R}^{N^*}$ , with  $N^*$  and  $K^*$  the total number of neurons and the total number of parameters of the array, respectively. In order to estimate in this context the formula (2.14), we first use the parallel TDR model (3.30) in order to determine the autocovariance  $\Gamma(0) \in \mathbb{S}_{N^*}$  out of the solutions (3.31) of the Yule-Walker equation. The input signal  $\{\mathbf{z}(t)\}_{t \in \mathbb{Z}}$  is assumed to be a family of  $n$ -dimensional independent and identically distributed random variables with mean zero and covariance matrix  $\Sigma_z \in \mathbb{S}_n^+$ , that is  $\{\mathbf{z}(t)\}_{t \in \mathbb{Z}} \sim \text{IID}(\mathbf{0}_n, \Sigma_z)$ . In these conditions, once a specific memory task of interest has been determined, all that is needed in order to complete the evaluation of the capacity formula (2.14) is the expressions for  $\text{Cov}(\mathbf{X}(t), \mathbf{y}(t))$  and  $\text{trace}(\text{Cov}(\mathbf{y}(t), \mathbf{y}(t)))$  that we now explicitly derive for the particular cases of the linear and quadratic memory tasks, respectively.

**Linear memory task.** Consider the linear  $q$ -dimensional  $h$ -lag memory task function  $H : \mathbb{R}^{(h+1)n} \longrightarrow \mathbb{R}^q$  determined by the assignment  $H(\mathbf{z}^h(t)) := L^\top \mathbf{z}^h(t) =: \mathbf{y}(t)$ , where  $\mathbf{z}^h(t) = \text{vec}(\mathbf{z}(t), \mathbf{z}(t-1), \dots, \mathbf{z}(t-h))$

$h)) \in \mathbb{R}^{(h+1)n}$  and  $L \in \mathbb{M}_{(h+1)n, q}$ . We now compute in this case  $\text{Cov}(\mathbf{X}(t), \mathbf{y}(t))$  and  $\text{trace}(\text{Cov}(\mathbf{y}(t), \mathbf{y}(t))) = \text{Cov}(\mathbf{y}(t)^\top, \mathbf{y}(t)^\top)$ , that are required for the memory capacity evaluation.

(i) We start with  $\text{Cov}(\mathbf{y}(t)^\top, \mathbf{y}(t)^\top)$  and write

$$\begin{aligned} \text{Cov}(\mathbf{y}(t)^\top, \mathbf{y}(t)^\top) &= \mathbb{E}[\mathbf{y}(t)^\top \mathbf{y}(t)] - \mathbb{E}[\mathbf{y}(t)^\top] \mathbb{E}[\mathbf{y}(t)] \\ &= \mathbb{E}[\mathbf{z}^h(t)^\top L L^\top \mathbf{z}^h(t)] = \text{trace}(L L^\top \mathbb{E}[\mathbf{z}^h(t) \mathbf{z}^h(t)^\top]) = \text{trace}(L L^\top \Sigma_{z^h}), \end{aligned} \quad (3.34)$$

where the covariance matrix  $\Sigma_{z^h} \in \mathbb{S}_{(h+1)n \times (h+1)n}$  has the form

$$\Sigma_{z^h} = \begin{pmatrix} \Sigma_z & \cdots & \mathbb{O}_n \\ \vdots & \ddots & \vdots \\ \mathbb{O}_n & \cdots & \Sigma_z \end{pmatrix}.$$

(ii) We now compute  $\text{Cov}(\mathbf{X}(t), \mathbf{y}(t))$ . As we already pointed out, the stability condition on the fixed point  $\mathbf{X}_0$  implies that the unique stationary solution of the VAR(1) model (3.30) admits a MA( $\infty$ ) representation of the form:

$$\mathbf{X}(t) - \boldsymbol{\mu}_X^{(\mathbf{X}_0, \Theta)} = \sum_{j=0}^{\infty} \Psi_j \boldsymbol{\rho}(t-j), \quad (3.35)$$

with  $\Psi_j \in \mathbb{M}_{N^*}$  and  $\boldsymbol{\rho}(t) := \boldsymbol{\varepsilon}(t)^{(\mathbf{X}_0, \Theta)} - \boldsymbol{\mu}_\varepsilon^{(\mathbf{X}_0, \Theta)}$ . Then, for any  $i \in \{1, \dots, N^*\}$  and  $j \in \{1, \dots, q\}$  we write

$$\begin{aligned} \text{Cov}(X_i(t), y_j(t)) &= \sum_{k=0}^{\infty} \text{Cov}((\Psi_k \boldsymbol{\rho}(t-k))_i, (L^\top \cdot \mathbf{z}^h(t))_j) = \sum_{k=0}^{\infty} \sum_{u=1}^{N^*} \sum_{v=1}^{(h+1)n} (\Psi_k)_{iu} L_{vj} \mathbb{E}[\rho_u(t-k) (\mathbf{z}^h(t))_v] \\ &= \sum_{k=0}^{\infty} \sum_{u=1}^{N^*} \sum_{v=1}^n \sum_{s=1}^{h+1} (\Psi_k)_{iu} L_{v \cdot s, j} \mathbb{E}[(\varepsilon_u(t-k)^{(\mathbf{X}_0, \Theta)} - (\boldsymbol{\mu}_\varepsilon^{(\mathbf{X}_0, \Theta)})_u) z_v(t-s+1)] \\ &= \sum_{u=1}^{N^*} \sum_{v=1}^n \sum_{s=0}^h (\Psi_s)_{iu} L_{v \cdot s, j} \mathbb{E}[\varepsilon_u(t-s)^{(\mathbf{X}_0, \Theta)} z_v(t-s)], \end{aligned} \quad (3.36)$$

where the vector  $\boldsymbol{\varepsilon}(t)^{(\mathbf{X}_0, \Theta)}$  is provided in (3.20)-(3.21) and the expectations  $\mathbb{E}[\varepsilon_u(t-s)^{(\mathbf{X}_0, \Theta)} z_v(t-s)]$  are computed by multiplying the  $V_R$  polynomial corresponding to  $\varepsilon_u(t)^{(\mathbf{X}_0, \Theta)}$  by the monomial  $z_v(t)$ ; the resulting polynomial is the evaluated on the higher order moments of  $\{\mathbf{z}(t)\}_{t \in \mathbb{Z}}$  using the same rule that we stated after (3.8).

**Quadratic memory task.** Consider now the quadratic  $q$ -dimensional  $h$ -lag memory task function  $H : \mathbb{R}^{(h+1)n} \rightarrow \mathbb{R}^q$  determined by the assignment  $H(\mathbf{z}^h(t)) := Q \cdot \text{vech}(\mathbf{z}^h(t) \cdot \mathbf{z}^h(t)^\top) =: \mathbf{y}(t)$ , where  $\mathbf{z}^h(t) = \text{vec}(\mathbf{z}(t), \mathbf{z}(t-1), \dots, \mathbf{z}(t-h)) \in \mathbb{R}^{(h+1)n}$ ,  $Q \in \mathbb{M}_{q, q^*}$  and  $q^* := \frac{1}{2}(h+1)n((h+1)n+1)$ . We now provide explicit expressions for  $\text{Cov}(\mathbf{X}(t), \mathbf{y}(t))$  and  $\text{Cov}(\mathbf{y}(t)^\top, \mathbf{y}(t)^\top)$  in this case, that are required in order to evaluate the corresponding memory capacity.

(i) We start with  $\text{Cov}(\mathbf{y}(t)^\top, \mathbf{y}(t)^\top)$ . Let  $M^h := \mathbf{z}^h(t) \mathbf{z}^h(t)^\top$  and write

$$\begin{aligned} \text{Cov}(\mathbf{y}(t)^\top, \mathbf{y}(t)^\top) &= \mathbb{E}[\mathbf{y}(t)^\top \mathbf{y}(t)] - \mathbb{E}[\mathbf{y}(t)^\top] \mathbb{E}[\mathbf{y}(t)] \\ &= \mathbb{E}[(\text{vech}(\mathbf{z}^h(t) \mathbf{z}^h(t)^\top))^\top Q^\top Q (\text{vech}(\mathbf{z}^h(t) \mathbf{z}^h(t)^\top))] - \mathbb{E}[\mathbf{y}(t)^\top] \mathbb{E}[\mathbf{y}(t)] \\ &= \text{trace}(Q^\top Q \cdot \mathbb{E}[\text{vech}(M^h) (\text{vech}(M^h))^\top]) - \mathbb{E}[\mathbf{y}(t)^\top] \mathbb{E}[\mathbf{y}(t)]. \end{aligned} \quad (3.37)$$

Notice now that for any  $i, j \in \{1, \dots, (h+1)n\}$  there exist  $l_i, l_j \in \{0, \dots, h\}$  and  $m_i, m_j \in \{0, \dots, n\}$  such that  $i = l_i p + m_i$ ,  $j = l_j p + m_j$  and hence

$$M_{ij}^h = (\mathbf{z}^h(t) \mathbf{z}^h(t)^\top)_{ij} = z_{m_i}(t - l_i) z_{m_j}(t - l_j). \quad (3.38)$$

Consequently, for any  $i, j \in \{1, \dots, (h+1)n\}$

$$\begin{aligned} \mathbb{E}[\text{vech}(M^h) \text{vech}(M^h)_{ij}^\top] &= \mathbb{E}[\text{vech}(M^h)_i \text{vech}(M^h)_j] = \mathbb{E}[M_{\sigma^{-1}(i)}^h M_{\sigma^{-1}(j)}^h] \\ &= \mathbb{E}[z_{m_{r(i)}}(t - l_{r(i)}) z_{m_{s(i)}}(t - l_{s(i)}) z_{m_{u(j)}}(t - l_{u(j)}) z_{m_{v(j)}}(t - l_{v(j)})], \end{aligned} \quad (3.39)$$

where the operator  $\sigma^{-1}$  assigns to the index of the position of an element in  $\text{vech}(M^h)$  the two indices corresponding to its position in the matrix  $M^h \in \mathbb{S}_{q^*}$ . In this expression  $(r(i), s(i)) := \sigma^{-1}(i)$ ,  $(u(j), v(j)) := \sigma^{-1}(j)$  and  $r(i) = l_{r(i)} p + m_{r(i)}$ ,  $s(i) = l_{s(i)} p + m_{s(i)}$ ,  $u(j) = l_{u(j)} p + m_{u(j)}$ ,  $v(j) = l_{v(j)} p + m_{v(j)}$  with  $l_{r(i)}, l_{s(i)}, l_{u(j)}, l_{v(j)} \in \{0, \dots, h\}$  and  $m_{r(i)}, m_{s(i)}, m_{u(j)}, m_{v(j)} \in \{1, \dots, p\}$ . Additionally, using this notation the following relation holds true

$$\mathbb{E}[y_k(t)] = \sum_{i=1}^{q^*} Q_{kj} \mathbb{E}[\text{vech}(M^h)_i] = \sum_{j=1}^{q^*} Q_{kj} \mathbb{E}[z_{m_{r(i)}}(t - l_{r(i)}) z_{m_{s(i)}}(t - l_{s(i)})]. \quad (3.40)$$

We hence now can derive the expression for (3.37) as

$$\begin{aligned} \text{Cov}(\mathbf{y}(t)^\top, \mathbf{y}(t)^\top) &= \sum_{i=1}^{q^*} \sum_{k=1}^q Q_{ki} \left( \sum_{j=1}^{q^*} Q_{kj} \cdot \mathbb{E}[z_{m_{r(i)}}(t - l_{r(i)}) z_{m_{s(i)}}(t - l_{s(i)}) z_{m_{u(j)}}(t - l_{u(j)}) z_{m_{v(j)}}(t - l_{v(j)})] \right. \\ &\quad \left. - Q_{ki} \mathbb{E}[z_{m_{r(i)}}(t - l_{r(i)}) z_{m_{s(i)}}(t - l_{s(i)})]^2 \right), \end{aligned} \quad (3.41)$$

where we used the same notation as in (3.39) and (3.40).

(ii)  $\text{Cov}(\mathbf{X}(t), \mathbf{y}(t))$ : for any  $i \in \{1, \dots, N^*\}$ ,  $j \in \{1, \dots, q\}$  we have

$$\begin{aligned} \text{Cov}(X_i(t), y_j(t)) &= \sum_{k=0}^{\infty} \text{Cov}((\Psi_k \boldsymbol{\rho}(t - k))_i, (Q \cdot \text{vech}(M^h))_j) \\ &= \sum_{k=0}^{\infty} \sum_{u=1}^{N^*} \sum_{v=1}^{q^*} (\Psi_k)_{iu} Q_{jv} \mathbb{E}[\rho_u(t - k) (\text{vech}(M^h))_v] \\ &= \sum_{k=0}^{\infty} \sum_{u=1}^{N^*} \sum_{v=1}^{q^*} (\Psi_k)_{iu} Q_{jv} \mathbb{E}[(\varepsilon_u(t - k)^{(\mathbf{X}_0, \boldsymbol{\Theta})} - (\boldsymbol{\mu}_\varepsilon^{(\mathbf{X}_0, \boldsymbol{\Theta})})_u) (\text{vech}(M^h))_v] \\ &= \sum_{k=0}^{\infty} \sum_{u=1}^{N^*} \sum_{v=1}^{q^*} (\Psi_k)_{iu} Q_{jv} \{ \mathbb{E}[(\varepsilon_u(t - k)^{(\mathbf{X}_0, \boldsymbol{\Theta})} - (\boldsymbol{\mu}_\varepsilon^{(\mathbf{X}_0, \boldsymbol{\Theta})})_u) z_{m_{r(v)}}(t - l_{m_{r(v)}}) z_{m_{s(v)}}(t - l_{m_{s(v)}})] \} \\ &= \sum_{k=0}^h \sum_{u=1}^{N^*} \sum_{v=1}^{q^*} (\Psi_k)_{iu} Q_{jv} \{ \mathbb{E}[\varepsilon_u(t - k)^{(\mathbf{X}_0, \boldsymbol{\Theta})} \cdot z_{m_{r(v)}}(t - l_{m_{r(v)}}) z_{m_{s(v)}}(t - l_{m_{s(v)}})] \\ &\quad - (\boldsymbol{\mu}_\varepsilon^{(\mathbf{X}_0, \boldsymbol{\Theta})})_u \mathbb{E}[z_{m_{r(v)}}(t - l_{m_{r(v)}}) z_{m_{s(v)}}(t - l_{m_{s(v)}})] \}, \end{aligned} \quad (3.42)$$

where the vector  $\boldsymbol{\varepsilon}(t)^{(\mathbf{X}_0, \boldsymbol{\Theta})}$  is provided in (3.20)-(3.21) and where we used the same notation as in (3.39) and (3.40).

### 3.4 The impact of the teaching signal size in the reservoir performance

In the preceding sections we evaluated the reservoir characteristic error or, equivalently, its capacity, in terms of various second order moments of the input signal and the reservoir output that in turn can be explicitly written in terms of the reservoir parameters. There are situations in which those moments are obtained directly out of finite sample realizations of the teaching signal and of the reservoir output using empirical estimators. That approach introduces an estimation error in the readout layer ( $W_{\text{out}}, \mathbf{a}_{\text{out}}$ ) that adds to the characteristic error (2.13). The quantification of that error is the main goal of this section. Since this error depends on the value of the regularizing constant  $\lambda$ , the pair  $(W_\lambda, \mathbf{a}_\lambda)$  will denote in what follows the optimal readout layer ( $W_{\text{out}}, \mathbf{a}_{\text{out}}$ ) given by (2.11)-(2.12) for a fixed value of the parameter  $\lambda$ .

**Properties of the ridge estimator.** Consider now  $\{\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(T)\}$  and  $\{\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(T)\}$  samples of size  $T$  of the reservoir output  $\{\mathbf{x}(t)\}_{t \in \mathbb{N}}$  and the teaching signal  $\{\mathbf{y}(t)\}_{t \in \mathbb{N}}$  processes. We concatenate horizontally these observations and we obtain the matrices  $X := (\mathbf{x}(1)|\mathbf{x}(2)|\dots|\mathbf{x}(T)) \in \mathbb{M}_{N,T}$  and  $Y := (\mathbf{y}(1)|\mathbf{y}(2)|\dots|\mathbf{y}(T)) \in \mathbb{M}_{q,T}$ . We now quantify the cost in terms of memory capacity or, equivalently, increased error, of using in the RC not the optimal readout layer  $(W_\lambda, \mathbf{a}_\lambda)$  given by (2.11)-(2.12) but an estimation  $(\widehat{W}_\lambda, \widehat{\mathbf{a}}_\lambda)$  of it based on estimators of the different moments present in those expressions using  $X$  and  $Y$ . More specifically, for a fixed  $\lambda$  and samples  $X$  and  $Y$ , we produce an estimation  $(\widehat{W}_\lambda, \widehat{\mathbf{a}}_\lambda)$  of  $(W_\lambda, \mathbf{a}_\lambda)$  by using in (2.11)-(2.12) the empirical estimators

$$\widehat{\boldsymbol{\mu}}_x := \frac{1}{T} X \mathbf{i}_T, \quad \widehat{\boldsymbol{\mu}}_y := \frac{1}{T} Y \mathbf{i}_T, \quad \widehat{\Gamma}(0) := \frac{1}{T} X X^\top - \widehat{\boldsymbol{\mu}}_x \widehat{\boldsymbol{\mu}}_x^\top = \frac{1}{T} X A X^\top, \quad (3.43)$$

$$\widehat{\text{Cov}}(\mathbf{y}(t), \mathbf{y}(t)) = \frac{1}{T} Y Y^\top - \widehat{\boldsymbol{\mu}}_y \widehat{\boldsymbol{\mu}}_y^\top = \frac{1}{T} Y A Y^\top, \quad (3.44)$$

$$\widehat{\text{Cov}}(\mathbf{x}(t), \mathbf{y}(t)) = \frac{1}{T} X Y^\top - \widehat{\boldsymbol{\mu}}_x \widehat{\boldsymbol{\mu}}_y^\top = \frac{1}{T} X A Y^\top, \quad (3.45)$$

where  $A := \mathbb{I}_T - \frac{1}{T} \mathbf{i}_T \mathbf{i}_T^\top$ . These expressions substituted in (2.11)-(2.12) yield

$$\widehat{W}_\lambda = (\widehat{\Gamma}(0) + \lambda \mathbb{I}_N)^{-1} \widehat{\text{Cov}}(\mathbf{x}(t), \mathbf{y}(t)) = (X A X^\top + \lambda T \mathbb{I}_N)^{-1} X A Y^\top, \quad (3.46)$$

$$\widehat{\mathbf{a}}_\lambda = \widehat{\boldsymbol{\mu}}_y - \widehat{W}_\lambda^\top \widehat{\boldsymbol{\mu}}_x = \frac{1}{T} (Y - \widehat{W}_\lambda^\top X) \mathbf{i}_T, \quad (3.47)$$

and determine a finite sample ridge regression estimator. The error associated to its use can be read out of its statistical properties that have been studied in the literature under various hypotheses. In the sequel we follow [Grig 15b], where  $(\widehat{W}_\lambda, \widehat{\mathbf{a}}_\lambda)$  is considered as a ridge estimator of the regression model

$$\mathbf{y}(t) = \mathbf{a} + W^\top \mathbf{x}(t) + \boldsymbol{\varepsilon}(t), \quad (3.48)$$

with  $\mathbf{a} \in \mathbb{R}^q$  a constant vector-intercept,  $W \in \mathbb{M}_{N,q}$ , and where the different random variables in (3.48) are assumed to satisfy the following hypotheses:

- (H1) The error terms  $\{\boldsymbol{\varepsilon}(t)\}_{t \in \mathbb{Z}}$  constitute a family of  $q$ -dimensional independent normally distributed vectors with mean  $\mathbf{0}_q$  and covariance matrix  $\Sigma_\varepsilon^q \in \mathbb{S}_q$ , that is,  $\{\boldsymbol{\varepsilon}(t)\}_{t \in \mathbb{Z}} \sim \text{IN}(\mathbf{0}_q, \Sigma_\varepsilon^q)$ .
- (H2)  $\{\mathbf{x}(t)\}_{t \in \mathbb{Z}}$  is a stochastic process such that  $\mathbf{x}(t) \in \mathbb{R}^N$  is a random vector independent of  $\boldsymbol{\varepsilon}(s) \in \mathbb{R}^q$  for any  $t, s \in \mathbb{Z}$ .
- (H3) The processes  $\{\mathbf{x}(t)\}_{t \in \mathbb{Z}}$ ,  $\{\mathbf{y}(t)\}_{t \in \mathbb{Z}}$ , and the joint process  $\{(\mathbf{x}(t), \mathbf{y}(t))\}_{t \in \mathbb{Z}}$  are second-order ergodic. This implies that the estimators in (3.43)-(3.45) converge to the corresponding moments when  $T \rightarrow \infty$  and also that these are second-order stationary (their means and autocovariances are time-independent).

Under those hypotheses, it can be shown that the estimator  $(\widehat{W}_\lambda, \widehat{\mathbf{a}}_\lambda)$  has the properties described in the following result whose proof can be found in [Grig 15b]. In the statement we use the notation  $Z \sim \text{MN}_{m,n}(M_Z, U_Z, V_Z)$  to indicate that  $Z \in \mathbb{M}_{m,n}$  is a matrix random variable distributed according to the matrix normal distribution with mean matrix  $M_Z \in \mathbb{M}_{m,n}$  and scale matrices  $U_Z \in \mathbb{S}_m$ ,  $V_Z \in \mathbb{S}_n$ . Details on these distributions can be found in [Gupt 00], and references therein.

**Proposition 3.1** *Consider the regression problem in (3.48) subjected to the hypotheses (H1)-(H3) and let  $\widehat{W}_\lambda \in \mathbb{M}_{N,q}$  and  $\widehat{\mathbf{a}}_\lambda \in \mathbb{R}^q$  be the ridge estimators given by (3.46) and (3.47), respectively, based on the samples of length  $T$  contained in the matrices  $X \in \mathbb{M}_{N,T}$  and  $Y \in \mathbb{M}_{q,T}$ . Then,*

- (i) *The distribution of the ridge estimator  $\widehat{W}_\lambda|X$  of  $W_\lambda$  conditioned by  $X$  is given by*

$$(\widehat{W}_\lambda - W_\lambda)|X \sim \text{MN}_{N,q}(-\lambda TRW_\lambda, \Sigma_W^r, \Sigma_W^c), \quad (3.49)$$

*or, equivalently,*

$$\text{vec}(\widehat{W}_\lambda - W_\lambda)|X \sim \text{N}(-\lambda T \text{vec}(RW_\lambda), \Sigma_W^c \otimes \Sigma_W^r), \quad (3.50)$$

*where the symbol  $\otimes$  stands for the Kronecker product and where the row and column covariance matrices  $\Sigma_W^r \in \mathbb{S}_N$  and  $\Sigma_W^c \in \mathbb{S}_T$ , respectively, are given by*

$$\Sigma_W^r = (\mathbb{I}_N - \lambda TR)R \quad \text{and} \quad \Sigma_W^c = \Sigma_\varepsilon^q, \quad (3.51)$$

*where  $R := (XAX^\top + \lambda T\mathbb{I}_N)^{-1}$  and  $A := \mathbb{I}_T - \frac{1}{T}\mathbf{i}_T\mathbf{i}_T^\top$ .*

- (ii) *The distribution of the ridge estimator  $\widehat{\mathbf{a}}_\lambda|X$  of  $\mathbf{a}_\lambda$  conditioned by  $X$  is given by*

$$(\widehat{\mathbf{a}}_\lambda - \mathbf{a}_\lambda)|X \sim \text{N}(\lambda W_\lambda^\top R X \mathbf{i}_T, \Sigma_a), \quad (3.52)$$

*where the covariance matrix  $\Sigma_a \in \mathbb{S}_q$  is given by*

$$\Sigma_a := \frac{1}{T} \left( 1 + \frac{1}{T} \text{trace}((\mathbb{I}_N - \lambda TR) R X \mathbf{i}_T \mathbf{i}_T^\top X^\top) \right) \Sigma_\varepsilon^q, \quad (3.53)$$

*and where  $R$  is defined as in (i).*

- (iii) *The covariance  $\Sigma_{W_a} \in \mathbb{M}_{Nq,q}$  between the conditioned ridge estimators  $\text{vec}(\widehat{W}_\lambda|X) \in \mathbb{R}^{Nq}$  and  $\widehat{\mathbf{a}}_\lambda|X \in \mathbb{R}^q$  is given by*

$$\Sigma_{W_a} := \frac{1}{T} (\text{vec}((\mathbb{I}_N - \lambda TR)W_\lambda) \mathbf{i}_T^\top X^\top ((\mathbb{I}_N - \lambda TR)W_\lambda) - \Sigma_\varepsilon^q \otimes ((\mathbb{I}_N - \lambda TR) R X \mathbf{i}_T)). \quad (3.54)$$

**The regression error with estimated regression parameters.** The reservoir error and the corresponding memory capacity in relations (2.13) and (2.14), respectively, were computed assuming that the optimal ridge parameters  $(W_\lambda, \mathbf{a}_\lambda)$  are known. This error, that is exclusively associated to the ability of the reservoir to perform or not the memory task in question, will be referred to as the **characteristic error** and will be denoted as

$$\begin{aligned} \text{MSE}_{\text{char},\lambda} &:= \text{trace} \left( \mathbb{E} \left[ (W_\lambda^\top \cdot \mathbf{x}(t) + \mathbf{a}_\lambda - \mathbf{y}(t))(W_\lambda^\top \cdot \mathbf{x}(t) + \mathbf{a}_\lambda - \mathbf{y}(t))^\top \right] \right) \\ &= \text{trace}(\text{Cov}(\mathbf{y}(t), \mathbf{y}(t)) - \text{Cov}(\mathbf{x}(t), \mathbf{y}(t))^\top (\Gamma(0) + \lambda \mathbb{I}_N)^{-1} (\Gamma(0) + 2\lambda \mathbb{I}_N) (\Gamma(0) + \lambda \mathbb{I}_N)^{-1} \text{Cov}(\mathbf{x}(t), \mathbf{y}(t))). \end{aligned} \quad (3.55)$$

When the reservoir is put to work using instead a readout layer  $(\widehat{W}_\lambda, \widehat{\mathbf{a}}_\lambda)$  that has been estimated using finite sample realizations of the processes  $\{\mathbf{x}(t)\}_{t \in \mathbb{Z}}$  and  $\{\mathbf{y}(t)\}_{t \in \mathbb{Z}}$ , the estimation error piles up with the characteristic one. The resulting error will be called **total error** and denoted  $\text{MSE}_{\text{total}, \lambda}$ . It can be computed using the distribution properties of the ridge estimator spelled out in Proposition 3.1 under the assumption that the samples  $(X, Y)$  that have been used to obtain the estimate  $(\widehat{W}_\lambda, \widehat{\mathbf{a}}_\lambda)$  of the output layer and those to evaluate the RC performance are independent. Indeed, under those hypotheses, the total reservoir error  $\text{MSE}_{\text{total}, \lambda} | X$  conditional on  $X$  is

$$\text{MSE}_{\text{total}, \lambda} | X := \text{trace} \left( \mathbb{E} \left[ ((W_\lambda + M_\lambda)^\top \cdot \mathbf{x}(t) + \mathbf{a}_\lambda + \mathbf{v}_\lambda - \mathbf{y}(t)) ((W_\lambda + M_\lambda)^\top \cdot \mathbf{x}(t) + \mathbf{a}_\lambda + \mathbf{v}_\lambda - \mathbf{y}(t))^\top | X \right] \right),$$

where the conditional random variables  $(M_\lambda, \mathbf{v}_\lambda) | X \in \mathbb{M}_{N, q} \times \mathbb{R}^q$  are independent from the processes  $\{\mathbf{x}(t)\}_{t \in \mathbb{Z}}$  and  $\{\mathbf{y}(t)\}_{t \in \mathbb{Z}}$  and have the same distribution properties than those spelled out in Proposition 3.1 for  $(\widehat{W}_\lambda - W_\lambda, \widehat{\mathbf{a}}_\lambda - \mathbf{a}_\lambda) | X$ . The main result in [Grig 15b] shows that, under those hypotheses, the total error conditional on  $X$  is given by

$$\begin{aligned} \text{MSE}_{\text{total}, \lambda} | X &= \text{MSE}_{\text{char}, \lambda} + \frac{1}{T} \text{trace}(\Sigma_\varepsilon^q) + \text{trace}(\Sigma_\varepsilon^q) \text{trace}((\mathbb{I}_N - \lambda T \mathcal{R}) \mathcal{R} Q) \\ &+ \lambda^2 T^2 \text{trace}(W_\lambda^\top \mathcal{R} Q \mathcal{R} W_\lambda) + 2 \lambda^2 T \text{trace}(W_\lambda^\top \mathcal{R} W_\lambda), \end{aligned} \quad (3.56)$$

where

$$Q := \Gamma(0) + \boldsymbol{\mu}_x \boldsymbol{\mu}_x^\top + \frac{1}{T^2} X \mathbf{i}_T \mathbf{i}_T^\top X^\top - \frac{2}{T} \boldsymbol{\mu}_x \mathbf{i}_T^\top X^\top \quad (3.57)$$

and  $\text{MSE}_{\text{char}, \lambda}$  is given by (3.55),  $W_\lambda = (\Gamma(0) + \lambda \mathbb{I}_N)^{-1} \text{Cov}(\mathbf{x}(t), \mathbf{y}(t))$ , and where

$$R = (T\Gamma(0) + \lambda T \mathbb{I}_N)^{-1}, \quad \mathcal{R} = (X A X^\top + \lambda T \mathbb{I}_N)^{-1}, \quad \text{and} \quad A = \mathbb{I}_T - \frac{1}{T} \mathbf{i}_T \mathbf{i}_T^\top. \quad (3.58)$$

In empirical applications, the moments  $\boldsymbol{\mu}_x$  and  $\Gamma(0)$  are in practice estimated with the same sample  $X, Y$  used to evaluate the reservoir performance. Therefore, if we replace  $\boldsymbol{\mu}_x$ ,  $\Gamma(0)$ ,  $\text{Cov}(\mathbf{y}(t), \mathbf{y}(t))$ , and  $\text{Cov}(\mathbf{x}(t), \mathbf{y}(t))$  in (3.56) by  $\frac{1}{T} X \mathbf{i}_T$ ,  $\frac{1}{T} X A X^\top$ ,  $\frac{1}{T} Y A Y^\top$  and  $\frac{1}{T} X A Y^\top$ , respectively, we obtain the approximated but simpler expression  $\text{MSE}_{\text{total}, \lambda}^{\text{approx}}(\lambda) | X$  for the total error:

$$\begin{aligned} \text{MSE}_{\text{total}, \lambda}^{\text{approx}} | X &= \text{MSE}_{\text{char}, \lambda}^{\text{approx}} \\ &+ \frac{1}{T} \text{trace}(\Sigma_\varepsilon^q) \left[ 1 + \text{trace}((\mathbb{I}_N - \lambda T \mathcal{R}) \mathcal{R} X A X^\top) \right] + \lambda^2 T \text{trace}(W_\lambda^\top \mathcal{R} (3 \mathbb{I}_N - \lambda T \mathcal{R}) W_\lambda), \end{aligned} \quad (3.59)$$

with

$$\text{MSE}_{\text{char}, \lambda}^{\text{approx}} = \frac{1}{T} \text{trace}(Y A Y^\top - Y A X^\top (\mathbb{I}_N + \lambda T \mathcal{R}) \mathcal{R} X A Y^\top). \quad (3.60)$$

This approximation can be taken one step further by replacing in (3.59)  $W_\lambda = (\Gamma(0) + \lambda \mathbb{I}_N)^{-1} \text{Cov}(\mathbf{x}(t), \mathbf{y}(t))$  by its estimator  $\widehat{W}_\lambda = (X A X^\top + \lambda T \mathbb{I}_N)^{-1} X A Y^\top$  and  $\Sigma_\varepsilon^q$  by its natural estimator in terms of  $X$  and  $Y$  for which

$$\text{trace}(\widehat{\Sigma}_\varepsilon^q) = \text{MSE}_{\text{char}, \lambda}^{\text{approx}}.$$

With these two additional substitutions, the equality (3.59) yields the following approximated expression for the total error:

$$\begin{aligned} \text{MSE}_{\text{total}, \lambda}^{\text{approx}} | X &= \text{MSE}_{\text{char}, \lambda}^{\text{approx}} + \\ &+ \frac{1}{T} \text{MSE}_{\text{char}, \lambda}^{\text{approx}} \left[ 1 + \text{trace}((\mathbb{I}_N - \lambda T \mathcal{R}) \mathcal{R} X A X^\top) \right] + \lambda^2 T \text{trace}(Y A X^\top \mathcal{R}^2 (3 \mathbb{I}_N - \lambda T \mathcal{R}) \mathcal{R} X A Y^\top). \end{aligned}$$

## 4 Empirical study

The goal of this section is twofold. First, we will assess the ability of the reservoir model in the multidimensional setup introduced in Section 3.1 to produce in that context good estimates of the memory capacity of the original reservoir, both in the continuous and the discrete-time setups. Our study shows that there is a good match between the performances of the model and of the actual reservoir and hence shows that the explicit expression of the reservoir capacity coming from the model can be used to find, for a given multidimensional task with multidimensional input signal, parameters and input masks for the original system that optimize its performance.

Second, we will use the parallel reservoir model in Section 3.2 and the memory capacity formulas that can be explicitly written with it, in order to verify various robustness properties of this reservoir architecture that were already documented in [Grig 14]. These properties have to do mainly with task misspecification and sensitivity to the parameter choice.

### Evaluation of the TDR performance in the processing of multidimensional signals.

In the first empirical experiment we present a quadratic memory task to an individually operating TDR. More explicitly, we inject in the reservoir a three dimensional independent input signal  $\{\mathbf{z}(t)\}_{t \in \mathbb{Z}}$ ,  $\mathbf{z}(t) \in \mathbb{R}^3$  with mean zero and covariance matrix  $\Sigma_z$ , and we study its ability to reconstruct the signal  $y(t) = \sum_{h=0}^3 \sum_{i,j=1}^3 z_i(t-h)z_j(t-h) \in \mathbb{R}$ . In the terminology introduced in Section 3.3, this exercise amounts to a quadratic task characterized by the vector  $Q \in \mathbb{R}^{78}$  defined by  $Q := (\text{vec}(Q^*))^\top D_{12}$ , with  $D_{12}$  the duplication matrix in dimension twelve and  $Q^* \in \mathbb{S}_{12}$  a block diagonal matrix with four matrices  $\mathbf{i}_3 \mathbf{i}_3^\top$  as its diagonal blocks. Indeed, if  $\mathbf{z}^3(t) := (\mathbf{z}(t), \mathbf{z}(t-1), \mathbf{z}(t-2), \mathbf{z}(t-3))$ , it is clear that

$$\begin{aligned} y(t) &= \mathbf{z}^3(t)^\top Q^* \mathbf{z}^3(t) = \text{trace}(\mathbf{z}^3(t)^\top Q^* \mathbf{z}^3(t)) = (\text{vec}(Q^{*\top}))^\top (\mathbf{z}^3(t) \otimes \mathbb{I}_{12}) \text{vec}(\mathbf{z}^3(t)) \\ &= (\text{vec}(Q^*))^\top \text{vec}(\mathbf{z}^3(t) \mathbf{z}^3(t)^\top) = (\text{vec}(Q^*))^\top D_{12} \text{vech}(\mathbf{z}^3(t) \mathbf{z}^3(t)^\top) = Q \cdot \text{vech}(\mathbf{z}^3(t) \mathbf{z}^3(t)^\top). \end{aligned}$$

In order to tackle this multidimensional task we use two twenty neuron TDRs constructed using the Mackey-Glass (2.3) and the Ikeda (2.4) nonlinear kernels. In Figures 3 and 4 we depict the error surfaces exhibited by both RCs in discrete and continuous time as a function of the distance between neurons and the feedback gain  $\eta$  and using fixed input gains  $\gamma$  whose values are indicated in the legends. Those error surfaces are computed using Monte Carlo simulations. At the same time we compute the error surfaces produced by the corresponding reservoir model and by evaluating the explicit capacity formula that can be written down in that case. The resulting figures exhibit a remarkable similarity that had already been observed for scalar input signals in [Grig 15a]. More importantly, the figures show the ability of the theoretical formula based on the reservoir model to locate the regions in parameter space for which the reservoir performance is optimal.

### Robustness properties of the parallel reservoir architecture.

We now study using the parallel reservoir model introduced in Section 3.2 the robustness properties of the parallel architecture with respect to parameter choice and misspecification task.

(i) **Parallel TDR configurations and robustness with respect to the choice of reservoir parameters.** An interesting feature of parallel TDR architectures that was observed in [Grig 14] is that optimal performance has a reduced sensitivity with respect to the choice of reservoir parameters when compared to that of individually operating reservoirs. In order to provide additional evidence of this fact, we have constructed parallel pools of 2, 5, 10, and 20 parallel Mackey-Glass-based TDRs and we present to them a nine-lag quadratic memory task  $y(t) = \sum_{i=0}^9 z(t-i)^2$  that, in the terminology introduced in Section 3.3 corresponds to the quadratic task characterized by the vector  $Q = (\text{vec}(\mathbb{I}_{10}))^\top D_{10}$ , with  $D_{10}$  the duplication matrix in dimension ten. For each of these parallel TDR architectures, as well as for an individually operating TDR, we will vary the number of the constituting neurons from 20 to

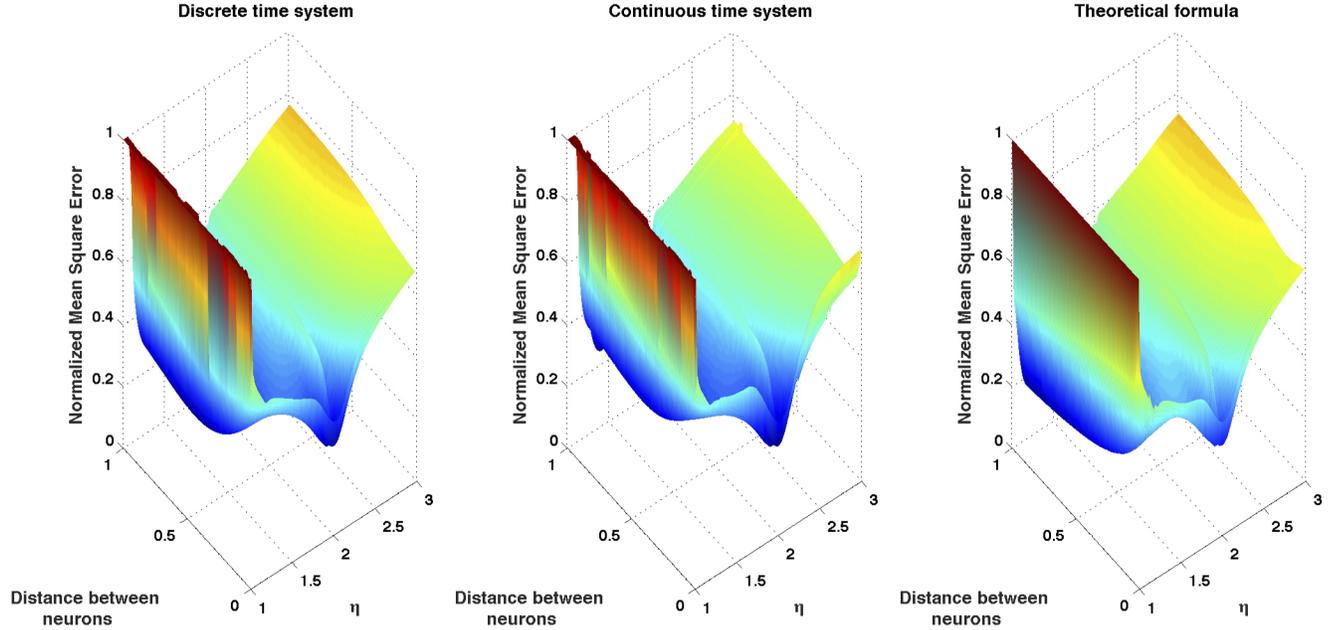


Figure 3: Normalized mean square error surfaces exhibited by an individually operating TDR constructed using a nonlinear Mackey-Glass kernel (2.3) performing a quadratic 3-lags memory task on a 3-dimensional independent mean zero input signal with covariance matrix  $\Sigma_z$  given by  $\text{vech}(\Sigma_z) = (0.0016, 0.0012, 0.0008, 0.0017, 0.0002, 0.0018)$ . In these figures the input gain  $\gamma = 0.6163$  is kept constant. The values of the input mask  $C \in \mathbb{M}_{20,3}$  are chosen randomly using a uniform distribution in the interval  $[-1; 1]$ . The left and middle panels show the error surfaces exhibited by the discrete and continuous time reservoirs computed via Monte Carlo simulations. The right panel shows the error produced by the reservoir model and computed evaluating the explicit capacity formula that can be written down in that case.

100. For each of these resulting configurations we randomly draw 1000 sets of input masks with entries uniformly distributed in the interval  $[-3, 3]$  and reservoir parameters and distance between neurons  $d$ , also uniformly distributed in the intervals  $\eta \in [1, 3]$ ,  $\gamma \in [-3, 3]$ , and  $d \in (0, 1)$ .

Figure 5 provides the box plots corresponding to the distributions of normalized mean squared errors obtained with each configuration by making the input masks and reservoir parameter values randomly vary, all of them computed using the capacity formulas associated to the parallel reservoir model (3.30). This figure provides striking evidence of the facts that first, the parallel architecture performs on average better and second, that this performance is not sensitive to the choice of reservoir parameters and input mask.

**(ii) Parallel TDR configurations and robustness with respect to memory task misspecification.** The goal of the next experience is to see how, given a specific memory task and given a parallel array of TDRs or an individually operating one that have been optimized for that particular task, the performance of the different configurations degrades when the task is modified but not the reservoir parameters. We use again parallel pools of 1, 2, 5, 10, and 20 Mackey-Glass-based TDRs with neurons ranging from 10 to 100. For each of those configurations, we choose parameters that optimize their performance with respect to the 3-lag quadratic memory task specified by the matrix  $Q = (\text{vec}(\mathbb{I}_4))^T D_4$  and with respect to a one-dimensional independent input signal with mean zero and variance 0.0001. Once the optimal parameters for each configuration have been found using the nonlinear capacity function

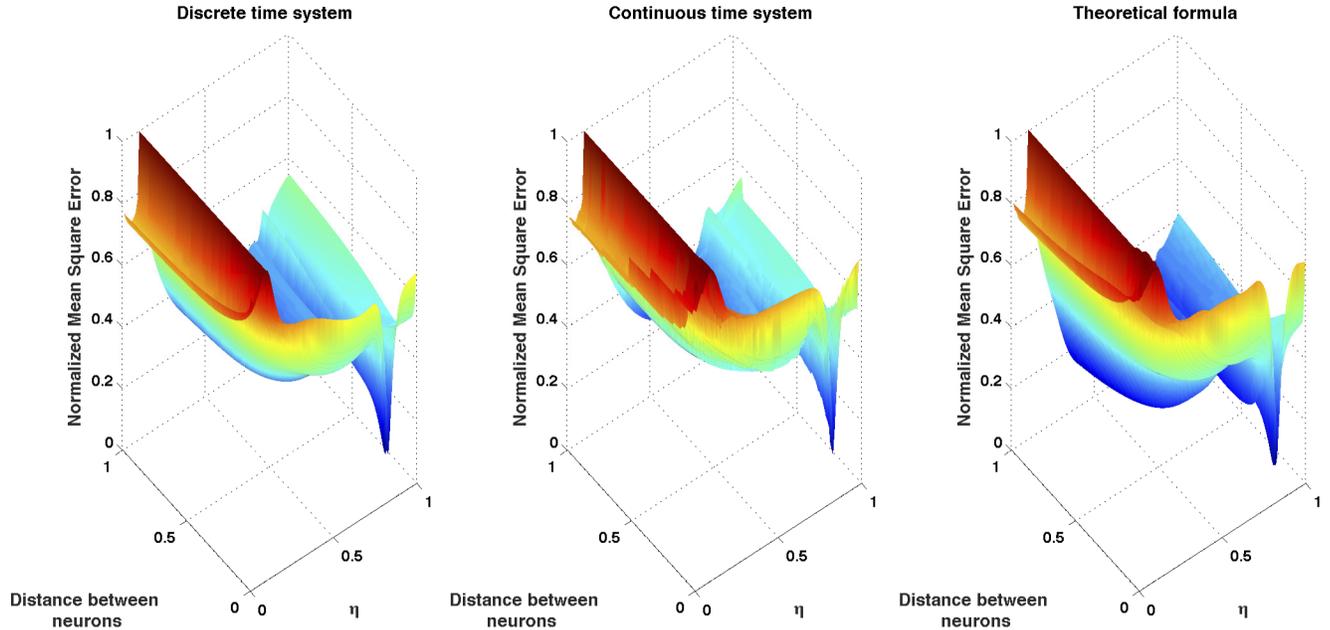


Figure 4: Normalized mean square error surfaces exhibited by an individually operating TDR constructed using a nonlinear Ikeda kernel (2.4) performing a quadratic 3-lags memory task on a 3-dimensional independent mean zero input signal with covariance matrix  $\Sigma_z$  given by  $\text{vech}(\Sigma_z) = (0.005, 0.0046, 0.0041, 0.0042, 0.0037, 0.004)$ . In these figures the input gain  $\gamma = 0.3724$  and the phase shift  $\phi = 0.7356$  are kept constant. The values of the input mask  $C \in \mathbb{M}_{20,3}$  are chosen randomly using a uniform distribution on the interval  $[-1, 1]$ . The left and middle panels show the error surfaces exhibited by the discrete and continuous time reservoirs computed via Monte Carlo simulations. The right panel shows the error produced by the reservoir model and computed evaluating the explicit capacity formula that can be written down in that case.

based on the reservoir model (3.30), we fix them we subsequently expose the corresponding reservoirs to different randomly specified memory tasks of a different specification, namely, 1000 different 9-lag quadratic tasks of the form  $Q = (\text{vec}(Q_{10}^*))^\top D_{10}$ , where  $Q_{10}^* \in \mathbb{M}_{10}$  is a randomly generated diagonal matrix with entries drawn from the uniform distribution on the interval  $[-10; 10]$ . Figure 6 contains the box plots corresponding to the performance distributions of the different configurations. In this case, parallel architectures offer an improvement on performance and robustness when compared to the single reservoir design. The most visible improvement is obtained in this case when using a parallel pool of two reservoirs.

## References

- [Appé 11] 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*, Vol. 2, p. 468, Jan. 2011.
- [Atiy 00] A. F. Atiya and A. G. Parlos. “New results on recurrent network training: unifying the algorithms and accelerating convergence”. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, Vol. 11, No. 3, pp. 697–709, Jan. 2000.

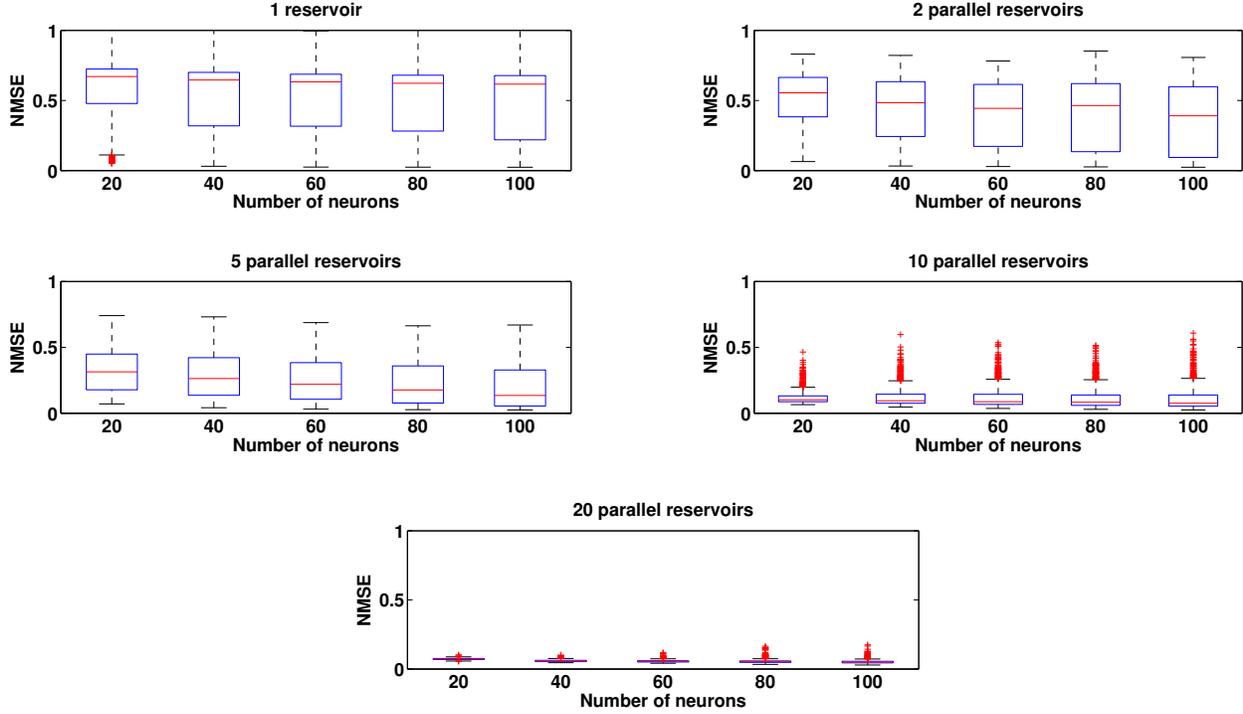


Figure 5: Performance distributions of parallel arrays of Mackey-Glass-based TDRs in a 9-lags quadratic memory task when the kernel parameters and the input mask are varied randomly.

- [Brun 13] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer. “Parallel photonic information processing at gigabyte per second data rates using transient states”. *Nature Communications*, Vol. 4, No. 1364, 2013.
- [Croo 07] N. Crook. “Nonlinear transient computation”. *Neurocomputing*, Vol. 70, pp. 1167–1176, 2007.
- [Crut 10] J. P. Crutchfield, W. L. Ditto, and S. Sinha. “Introduction to focus issue: intrinsic and designed computation: information processing in dynamical systems-beyond the digital hegemony”. *Chaos (Woodbury, N.Y.)*, Vol. 20, No. 3, p. 037101, Sep. 2010.
- [Grig 14] L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. “Stochastic time series forecasting using time-delay reservoir computers: performance and universality”. *Neural Networks*, Vol. 55, pp. 59–71, 2014.
- [Grig 15a] L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. “Optimal nonlinear information processing capacity in delay-based reservoir computers”. *Scientific Reports*, Vol. 5, No. 12858, pp. 1–11, 2015.
- [Grig 15b] L. Grigoryeva and J.-P. Ortega. “Ridge regularized multidimensional regressions: error evaluation with estimated regression parameters”. *Preprint*, 2015.

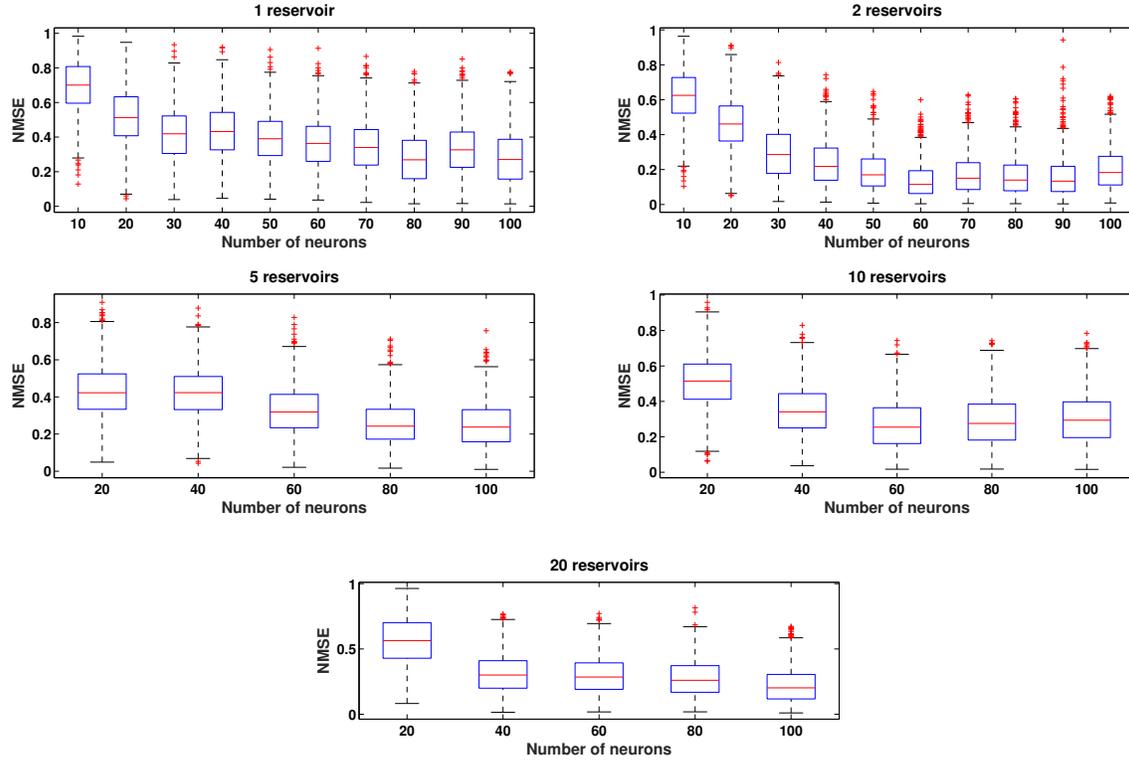


Figure 6: Robustness of the performance of parallel arrays of Mackey-Glass kernel based TDRs with respect to task misspecification. The box plots report the distribution of normalized mean square errors committed in the execution of 1000 randomly generated 9-lag diagonal quadratic memory tasks by different TDR configurations that had been initially optimized for a 3-lag quadratic memory task.

- [Gupt 00] A. Gupta and D. Nagar. *Matrix Variate Distributions*. Chapman and Hall/CRC, 2000.
- [Holm 88] B. Holmquist. “Moments and cumulants of the multivariate normal distribution”. *Stochastic Analysis and Applications*, Vol. 6, No. 3, pp. 273–278, Jan. 1988.
- [Ikeda 79] K. Ikeda. “Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system”. *Optics Communications*, Vol. 30, No. 2, pp. 257–261, Aug. 1979.
- [Jaeg 01] H. Jaeger. “The ‘echo state’ approach to analysing and training recurrent neural networks”. Tech. Rep., German National Research Center for Information Technology, 2001.
- [Jaeg 04] H. Jaeger and H. Haas. “Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication”. *Science*, Vol. 304, No. 5667, pp. 78–80, 2004.
- [Jaeg 07] H. Jaeger, M. Lukoševičius, D. Popovici, and U. Siewert. “Optimization and applications of echo state networks with leaky-integrator neurons”. *Neural Networks*, Vol. 20, No. 3, pp. 335–352, 2007.

- [Larg 12] 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*, Vol. 20, No. 3, p. 3241, Jan. 2012.
- [Luko 09] M. Lukoševičius and H. Jaeger. “Reservoir computing approaches to recurrent neural network training”. *Computer Science Review*, Vol. 3, No. 3, pp. 127–149, 2009.
- [Lutk 05] H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin, 2005.
- [Maas 02] W. Maass, T. Natschläger, and H. Markram. “Real-time computing without stable states: a new framework for neural computation based on perturbations”. *Neural Computation*, Vol. 14, pp. 2531–2560, 2002.
- [Maas 11] W. Maass. “Liquid state machines: motivation, theory, and applications”. In: S. S. Barry Cooper and A. Sorbi, Eds., *Computability In Context: Computation and Logic in the Real World*, Chap. 8, pp. 275–296, 2011.
- [Mack 77] M. C. Mackey and L. Glass. “Oscillation and chaos in physiological control systems”. *Science*, Vol. 197, pp. 287–289, 1977.
- [Meye 00] C. Meyer. *Matrix Analysis and Applied Linear Algebra Book and Solutions Manual*. Society for Industrial and Applied Mathematics, 2000.
- [Orti 12] S. Ortin, L. Pesquera, and J. M. Gutiérrez. “Memory and nonlinear mapping in reservoir computing with two uncoupled nonlinear delay nodes”. In: *Proceedings of the European Conference on Complex Systems*, pp. 895–899, 2012.
- [Paqu 12] Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar. “Optoelectronic reservoir computing”. *Scientific reports*, Vol. 2, p. 287, Jan. 2012.
- [Roda 11] A. Rodan and P. Tino. “Minimum complexity echo state network.”. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, Vol. 22, No. 1, pp. 131–44, Jan. 2011.
- [Tria 03] K. Triantafyllopoulos. “On the central moments of the multidimensional Gaussian distribution”. *The Mathematical Scientist*, Vol. 28, pp. 125–128, 2003.
- [Vers 07] D. Verstraeten, B. Schrauwen, M. DHaene, and D. Stroobandt. “An experimental unification of reservoir computing methods”. *Neural Networks*, Vol. 20, pp. 391–403, 2007.