Journal of Machine Learning Research 21 (2020) 1-61 Submitted 10/19; Published 11/20

Risk Bounds for Reservoir Computing

Lukas Gonon GONON@MATH.LMU.DE
Mathematics Institute
Ludwig-Maximilians-Universität München
Germany

Lyudmila Grigoryeva LYUDMILA.GRIGORYEVA@UNI-KONSTANZ.DE
Department of Mathematics and Statistics
Graduate School of Decision Sciences
Universität Konstanz
Germany

Juan-Pablo Ortega JUAN-PABLO.ORTEGA@UNISG.CH
Faculty of Mathematics and Statistics
Universität Sankt Gallen
Switzerland
Centre National de la Recherche Scientifique (CNRS)
France

Editor: Moritz Hardt

Abstract

We analyze the practices of reservoir computing in the framework of statistical learning theory. In particular, we derive finite sample upper bounds for the generalization error committed by specific families of reservoir computing systems when processing discrete-time inputs under various hypotheses on their dependence structure. Non-asymptotic bounds are explicitly written down in terms of the multivariate Rademacher complexities of the reservoir systems and the weak dependence structure of the signals that are being handled. This allows, in particular, to determine the minimal number of observations needed in order to guarantee a prescribed estimation accuracy with high probability for a given reservoir family. At the same time, the asymptotic behavior of the devised bounds guarantees the consistency of the empirical risk minimization procedure for various hypothesis classes of reservoir functionals.

Keywords: Reservoir computing, RC, echo state networks, ESN, state affine systems, SAS, random reservoirs, Rademacher complexity, weak dependence, empirical risk minimization, PAC bounds, risk bounds.

1. Introduction

Reservoir computing (RC) is a well established paradigm in the supervised learning of dynamic processes, which exploits the ability of specific families of semi-randomly generated state-space systems to solve computational and signal treatment tasks, both in deterministic and in stochastic setups. In recent years, both researchers and practitioners have been paying increasing attention to reservoir systems and their applications in learning. The main reasons behind this growing interest are threefold. Firstly, training strategies in reser-

voir computing are easy to implement as they simply consist in estimating the weights of memoryless static readouts, while the internal weights of the reservoir network are randomly created; this feature is closely linked to ideas originating in biology and the neurosciences in relation with the design of brain-inspired computing algorithms. Second, there is an important interplay between reservoir systems and the theory of dynamical systems, recurrent neural networks, and nonlinear discrete-time state-space systems, which makes the collection of tools available for their analysis very rich and explains in part why RC appears in the literature assimilated to other denominations, such as ***Liquid State Machines*** (see Maass and Sontag, 2000; Maass et al., 2002; Natschläger et al., 2002; Maass et al., 2004, 2007) or ***Echo State Networks*** (see Jaeger and Haas, 2004; Jaeger, 2010)). Finally, several families of reservoir systems have shown excellent performance in various classification and forecasting exercises including both standard machine learning benchmarks (see Lukoševičius and Jaeger, 2009, and references therein) and sophisticated applications that range from learning the attractors of chaotic nonlinear infinite dimensional dynamical systems (see Jaeger and Haas, 2004; Pathak et al., 2017, 2018; Lu et al., 2018) to the detection of Steady-State Visual Evoked Potentials (SSVEPs) in electroencephalographic signals as in Ibáñez-Soria et al. (2019). It is also important to point out that RC implementations with dedicated hardware have been proved to exhibit information processing speeds that significantly outperform standard Turing-type computers (see, for instance, Appeltant et al., 2011; Rodan and Tino, 2011; Vandoorne et al., 2011; Larger et al., 2012; Paquot et al., 2012; Brunner et al., 2013; Vandoorne et al., 2014; Vinckier et al., 2015; Laporte et al., 2018).

For a number of years, the reservoir computing community has worked hard on characterizing the key properties that explain the performance of reservoirs in classification, forecasting, and memory reconstruction tasks and also on formulating necessary conditions for a given state-space system to serve as a properly functioning reservoir system. Salient features of reservoir systems that have been shown to be important are the ***fading memory property (FMP)***, which appears in the context of systems theory (Volterra, 1959; Wiener, 1958; Boyd and Chua, 1985), computational neurosciences (Maass et al., 2004), physics (Coleman and Mizel, 1968), or mechanics (see Fabrizio et al., 2010, and references therein), the ***echo state property (ESP)*** (Jaeger, 2010; Yildiz et al., 2012; Manjunath and Jaeger, 2013), and the pairwise ***separation property (SP)*** (see, for instance, Legenstein and Maass, 2007; Lukoševičius and Jaeger, 2009; Maass, 2011, and references therein). Much effort has been made to provide rigorous definitions for these concepts and to characterize their relations under various hypotheses (see Jaeger, 2010; Grigoryeva and Ortega, 2019, and references therein). In particular, the crucial importance of these properties manifests itself in a series of universal approximation results which have been obtained for RC systems (see for example Maass et al., 2007; Grigoryeva and Ortega, 2018a,b; Gonon and Ortega, 2020b,a). This feature is a *dynamic* analog to well-established universal approximation properties for *static* machine learning paradigms, like neural networks (Cybenko, 1989; Funahashi, 1989; Hornik et al., 1989), for which the so-called ***approximation error*** (see Cucker and Smale, 2002; Smale and Zhou, 2003; Cucker and Zhou, 2007) can be made arbitrary small.

From the point of view of learning theory, the most important feature for any paradigm is its ability to generalize. Here this means that the performance of a given RC architecture on a training sample should be comparable to its behavior on previously unseen realizations

of the same data generation process. In the RC literature, this problem has been traditionally tackled using the notion of ***memory capacity***, that has been the subject of much research (Jaeger, 2002; White et al., 2004; Ganguli et al., 2008; Hermans and Schrauwen, 2010; Dambre et al., 2012; Grigoryeva et al., 2015; Couillet et al., 2016; Grigoryeva et al., 2016). Unfortunately, it has been recently shown that optimizing memory capacity does not necessarily lead to higher prediction performance (see Marzen, 2017). Moreover, the recently proved universal approximation properties of RC that we just brought up do not guarantee that a given universal reservoir system will exhibit small ***generalization errors***. In other words, they guarantee the availability of RC architectures that exhibit arbitrarily small training errors but give no control on their generalization power.

Following the standard learning theoretical approach to measure the generalization power would lead to consider the difference between the training error (empirical risk) and the testing error (statistical risk or generalization error) and aim at controlling it uniformly over a given class of reservoir systems by using a measure of the class complexity. A number of complexity measures for function classes have been proposed over the years. We can name the Vapnik-Chervonenkis (VC) dimension (Vapnik, 1998), Rademacher and Gaussian complexities (Bartlett and Mendelson, 2003), uniform stability (Mukherjee et al., 2006; Bousquet and Elisseeff, 2002; Poggio et al., 2004), and their modifications. In particular, a vast literature is available also on complexities and probably approximately correct (PAC) bounds for multilayer neural networks or recurrent neural networks (see, for instance, Haussler, 1992; Koiran and Sontag, 1998; Sontag, 1998; Anthony and Bartlett, 1999; Bartlett et al., 2017; Zhang et al., 2018, and references therein).

However, it is important to emphasize that using this traditional learning theoretical approach to formulate generalization error bounds in the case of reservoir systems is challenging and requires non-trivial extensions of this circle of ideas. Indeed, since a key motivation for our analysis are time series applications, the standard i.i.d. assumption on inputs and outputs cannot be invoked anymore, which makes a number of conventional tools unsuitable. Here the signals to be treated are stochastic processes with a particular dependence structure, which introduces mathematical difficulties that only a few works have analyzed in a learning theory context. In most of the available contributions on learning in a non-i.i.d. setting, stationarity and specific mixing properties of the input are key assumptions (see, for instance, McDonald et al., 2017; Kuznetsov and Mohri, 2017, 2018, and references therein). The time series applications that we are interested motivate us, however, to part with the latter. A common argument in this direction (Kuznetsov and Mohri, 2018) is that many standard time-series processes happen to be non-mixing; for example, one can easily construct AR(1) and ARFIMA processes which are not mixing (see Andrews, 1983; Baillie, 1996, respectively). On the other hand, it has been pointed out (see Adams and Nobel, 2010) that the convergence of empirical quantities to population-based ones can be arbitrarily slow for general stationary data and one cannot hope to obtain distribution-free probability bounds as they exist for the i.i.d. case. Motivated by these observations, in this article we restrict to a particular type of dependent processes, namely, we focus on dependence structures created by causal Bernoulli shifts (see, for instance, Dedecker et al., 2007; Alquier and Wintenberger, 2012) and hence the error bounds that we obtain are valid for any input process with such a dependence structure. Apart from trivially incorporating the i.i.d. case, the Bernoulli shifts category includes the VARMA time-series class of

models, financial econometric models such as various GARCH specifications (Engle, 1982; Bollerslev, 1986; Engle, 2009), and the ARFIMA (Beran, 1994) processes that allow the modeling of long memory behavior exhibited by many financial time series (for example realized variances). As we show later on, even though the bounds that we obtain depend on the weak dependence assumptions, they hold true without making precise the distributions of the input and the outputs, which are generally unknown. We hence place ourselves in a ***semi-agnostic setup***.

Regarding complexity measures, bounds for them are also customarily formulated in an i.i.d. setting. Recently, some authors addressed the question of constructing versions of Rademacher complexities for dependent inputs. For example, if one defines the risk in terms of conditional expectations, then the so-called sequential Rademacher complexities can be used to derive bounds (see, for instance, Rakhlin et al. (2010, Proposition 15) and Rakhlin, Sridharan, and Tewari (2014)). In this paper we pursue a more traditional approach in terms of the definition of the expected risk and hence the associated Rademacher complexity.

The main contribution of this paper is the formulation of *the first explicit generalization bounds for reservoir systems such as recurrent neural networks with input data exhibiting a sequential dependence structure for the classical notion of risk defined as an expected loss.* The uniform high-probability bounds which we state in this paper depend exclusively on the weak dependence behavior of the input and target processes and a quantitative measure of the capacity (Rademacher complexity) of the set of functionals generated by the considered reservoir systems. The finite sample guarantees provided by our generalization bounds explicitly answer practical questions concerning the bounds for the parameters within a particular reservoir family, the rates of uniform convergence, and hence the length of the training sample required to achieve a desired learning generalization quality within a given RC class. Finally, when one wishes to apply empirical risk minimization (ERM) in order to pick the reservoir functional within the hypothesis class, the asymptotic behavior of the devised bounds guarantees the consistency of ERM for reservoir systems.

The paper is organized as follows:

- Section 2 describes the notation used in the paper. We introduce reservoir systems, the associated filters and functionals, as well as a detailed description of various families of popular reservoir systems in the literature.

- Section 3 sets up the statistical learning problem for reservoir computing systems. It starts by introducing a general framework for the learning procedure, necessary risk definitions, and criteria of risk-consistency for the particular case of empirical risk minimization (ERM). The second subsection constitutes the main part of Section 3 and we present in it the setting in which reservoir systems are analyzed in the rest of the paper. First, three alternative core assumptions regarding the weak dependence structure of the input and target processes are analyzed and illustrated with examples. Second, the hypothesis classes of reservoir maps and functionals are constructed under a set of mild assumptions. Finally, the strategy for the derivation of risk bounds for a given choice of loss function is discussed.

- Section 4 contains the main results of the paper. Proofs and auxiliary results are postponed to the appendices. Section 4 is structured as follows. In the first subsec-

tion the expected value of the worst-case difference between the generalization and training errors over the class of reservoir functionals is shown to be bounded by its Rademacher complexity and terms related to the weak dependence structure of the input and target processes. The obtained rates differ depending on the various assumptions invoked. The second subsection provides explicit expressions for upper bounds of the Rademacher complexities associated to the families of reservoir systems presented in Section 2. The third subsection concludes with the formulation of high-probability finite-sample generalization bounds for reservoir systems. We emphasize that previously such bounds were not available in the literature. The asymptotic behavior of these bounds shows in passing the weak risk-consistency of the ERM procedure for reservoir systems. The last subsection contains a result that provides high-probability bounds for families of reservoir systems whose reservoir maps have been generated randomly. This result is a theoretical justification of the well-known good empirical properties of this standard *modus operandi* in reservoir computing.

2. Preliminaries

We start by specifying our notation and introducing reservoir computing systems for which in the following sections we will set up a statistical learning strategy. In the last subsection we provide a list of particular families of reservoir systems which are popular in the RC literature and in applications.

2.1 Notation

We use the symbol $\mathbb{N}$ (respectively, $\mathbb{N}^+$) to denote the set of natural numbers with the zero element included (respectively, excluded). $\mathbb{Z}$ denotes the set of all integers, and $\mathbb{Z}_-$ (respectively, $\mathbb{Z}_+$) stands for the set of the negative (respectively, positive) integers with the zero element included. Let $d, n, m \in \mathbb{N}^+$. Given an element $\mathbf{x} \in \mathbb{R}^n$, we denote by $\mathbb{R}[\mathbf{x}]$ the real-valued multivariate polynomials on $\mathbf{x}$ with real coefficients. Given a vector $\mathbf{v} \in \mathbb{R}^n$, the symbol $\|\mathbf{v}\|_2$ stands for its Euclidean norm. We denote by $\mathbb{M}_{m,n}$ the space of real $m \times n$ matrices. When $n = m$, we use the symbol $\mathbb{M}_n$ to refer to the space of square matrices of order n. For any $A \in \mathbb{M}_{m,n}$, $|||A|||_2$ denotes its matrix norm induced by the Euclidean norms in $\mathbb{R}^m$ and $\mathbb{R}^n$, which satisfies that $|||A|||_2 = \sigma_{max}(A)$ with $\sigma_{max}(A)$ the largest singular value of A. $|||A|||_2$ is sometimes referred to as the spectral norm of A (see Horn and Johnson, 2013).

When working in a deterministic setup, the inputs and outputs will be modeled using semi-infinite sequences $\mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-}$ and $\mathbf{y} \in (\mathbb{R}^m)^{\mathbb{Z}_-}$, respectively. We shall restrict very frequently to input sequences that exhibit additional convergence properties that are imposed with the help of weighting sequences. A weighting sequence w is a strictly decreasing sequence with zero limit $w : \mathbb{N} \longrightarrow (0, 1]$ such that $w_0 = 1$. We define the **weighted 1-norm** or the $(1, w)$**-norm** $\|\cdot\|_{1,w}$ in the space of semi-infinite sequences $(\mathbb{R}^d)^{\mathbb{Z}_-}$ as

$$\|\mathbf{z}\|_{1,w} := \sum_{t \in \mathbb{Z}_-} \|\mathbf{z}_t\|_2 \, w_{-t}, \quad \text{for any} \quad \mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-}.$$

We then set

$$\ell^{1,w}_-(\mathbb{R}^d) := \left\{ \mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-} \mid \|\mathbf{z}\|_{1,w} < \infty \right\}.$$

This weighted sequence space can be characterized as a Bochner space (Hytönen et al., 2016) by noticing that

$$\left(\ell_{-}^{1,w}(\mathbb{R}^d), \|\cdot\|_{1,w}\right) = \left(L^1(\mathbb{Z}_-, \mathcal{P}(\mathbb{Z}_-), \mu_w; \mathbb{R}^d), \|\cdot\|_{L^1(\mathbb{Z}_-;\mathbb{R}^d)}\right),$$

where $\mathcal{P}(\mathbb{Z}_-)$ stands for the power set of $\mathbb{Z}_-$ and μ_w is the measure defined on $(\mathbb{Z}_-, \mathcal{P}(\mathbb{Z}_-))$ generated by the assignments $\mu(\{t\}) := w_{-t}$, for any $t \in \mathbb{Z}_-$. This equality guarantees that the pair $\left(\ell_{-}^{1,w}(\mathbb{R}^d), \|\cdot\|_{1,w}\right)$ forms a separable Banach space.

Let now $\tau \in \mathbb{Z}_-$ and define the **time delay operator** $T_{-\tau} : (\mathbb{R}^d)^{\mathbb{Z}_-} \longrightarrow (\mathbb{R}^d)^{\mathbb{Z}_-}$ by $T_{-\tau}(\mathbf{z})_t := \mathbf{z}_{t+\tau}$, for any $t \in \mathbb{Z}_-$. We call $T_{-\tau}(\mathbf{z}) \in (\mathbb{R}^d)^{\mathbb{Z}_-}$ the **τ-shifted version** of the semi-infinite sequence $\mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-}$. It can be proved (Grigoryeva and Ortega, 2019) that $T_{-\tau}$ restricts to a continuous linear operator in $\left(\ell_{-}^{1,w}(\mathbb{R}^d), \|\cdot\|_{1,w}\right)$ and that the operator norm of the resulting maps $T_{-\tau} : \ell_{-}^{1,w}(\mathbb{R}^d) \longrightarrow \ell_{-}^{1,w}(\mathbb{R}^d)$ satisfies

$$|||T_1|||_{1,w} = L_w \text{ and } \quad |||T_{-\tau}|||_{1,w} \leq L_w^{-\tau}, \text{ for all } \tau \in \mathbb{Z}_-, \tag{1}$$

provided that the condition $L_w < \infty$ holds, where $1 \leq L_w \leq \infty$ is the **inverse decay ratio** of w defined as

$$L_w := \sup_{t\in\mathbb{N}} \left\{ \frac{w_t}{w_{t+1}} \right\}.$$

We define for future reference the ***decay ratio*** D_w of w as

$$D_w := \sup_{t\in\mathbb{N}} \left\{ \frac{w_{t+1}}{w_t} \right\} \leq 1.$$

The other weighted norm of much use in the context of reservoir computing is the **(∞, w)-norm**, defined by

$$\|\mathbf{z}\|_{\infty,w} := \sup_{t\in\mathbb{Z}_-} \{\|\mathbf{z}_t\|_2 w_{-t}\}, \quad \text{for any} \quad \mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-}.$$

We then set $\ell_{-}^{\infty,w}(\mathbb{R}^d) := \left\{ \mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-} \mid \|\mathbf{z}\|_{\infty,w} < \infty \right\}$. It can also be showed that the pair $(\ell_{-}^{\infty,w}(\mathbb{R}^d), \|\cdot\|_{\infty,w})$ is a Banach space (Grigoryeva and Ortega, 2018b). Additionally, the time delay operators also restrict to $\ell_{-}^{\infty,w}(\mathbb{R}^d)$ and the corresponding operator norms satisfy (1).

2.2 Filters and reservoir computing systems

The objects at the core of this paper are input/output maps of the form $U : \left(\mathbb{R}^d\right)^{\mathbb{Z}} \longrightarrow (\mathbb{R}^m)^{\mathbb{Z}}$. We will restrict to the case in which the maps U are **causal** and **time-invariant** (see Grigoryeva and Ortega, 2019, for definitions and the proofs of the facts that we now state) and hence it suffices to work with the restrictions $U : \left(\mathbb{R}^d\right)^{\mathbb{Z}_-} \longrightarrow (\mathbb{R}^m)^{\mathbb{Z}_-}$. Moreover, causal and time-invariant filters U uniquely determine functionals of the type $H_U : \left(\mathbb{R}^d\right)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}^m$ by

$$H_U(\mathbf{z}) := U(\mathbf{z})_0, \quad \text{for any} \quad \mathbf{z} \in \left(\mathbb{R}^d\right)^{\mathbb{Z}_-}.$$

In this setup, we shall say that $U:\left(\mathbb{R}^d\right)^{\mathbb{Z}_-}\longrightarrow(\mathbb{R}^m)^{\mathbb{Z}_-}$ is a **filter** and that $H_U:\left(\mathbb{R}^d\right)^{\mathbb{Z}_-}\longrightarrow\mathbb{R}^m$ is its corresponding **functional**. Conversely, given a functional $H:\left(\mathbb{R}^d\right)^{\mathbb{Z}_-}\longrightarrow\mathbb{R}^m$, there is a unique causal and time-invariant filter $U_H:\left(\mathbb{R}^d\right)^{\mathbb{Z}_-}\longrightarrow(\mathbb{R}^m)^{\mathbb{Z}_-}$ determined by it as

$$U_H(\mathbf{z})_t:=H(T_{-t}(\mathbf{z})),\quad \text{for any}\quad \mathbf{z}\in\left(\mathbb{R}^d\right)^{\mathbb{Z}_-},\ t\in\mathbb{Z}_-. \tag{2}$$

Suppose that given a weighting sequence w, the filter U restricts to a map between weighted (∞,w)-spaces, that is, $U:\ell_-^{\infty,w}(\mathbb{R}^d)\longrightarrow\ell_-^{\infty,w}(\mathbb{R}^m)$ and that, additionally, U is continuous with respect to the norm topology in those spaces. In that case we say that U has the **fading memory property (FMP)** with respect to w.

We shall provide an answer to the supervised learning of filters by estimating approximants built as **reservoir filters**. Reservoir filters are obtained out of a **reservoir system**, that is, a state-space system made out of two recurrent equations of the form:

$$\left\{\begin{array}{rcl}\mathbf{x}_t&=&F(\mathbf{x}_{t-1},\mathbf{z}_t),\\ \mathbf{y}_t&=&h(\mathbf{x}_t),\end{array}\right. \tag{3}$$

for all $t\in\mathbb{Z}_-$ and where $F\colon D_N\times D_d\longrightarrow D_N$ and $h:D_N\longrightarrow\mathbb{R}^m$ are maps, $D_d\subset\mathbb{R}^d$, $D_N\subset\mathbb{R}^N$. The sequences $\mathbf{z}\in(D_d)^{\mathbb{Z}_-}$ and $\mathbf{y}\in(\mathbb{R}^m)^{\mathbb{Z}_-}$ stand for the ***input*** and the ***output (target)*** of the system, respectively, and $\mathbf{x}\in(D_N)^{\mathbb{Z}_-}$ are the associated ***reservoir states***.

A reservoir system determines a filter when the first equation in (3) satisfies the so-called **echo state property (ESP**), that is, when for any $\mathbf{z}\in(D_d)^{\mathbb{Z}_-}$ there exists a unique $\mathbf{x}\in(D_N)^{\mathbb{Z}_-}$ such that (3) holds. In that case, we talk about the reservoir filter $U_h^F:(D_d)^{\mathbb{Z}_-}\longrightarrow(\mathbb{R}^m)^{\mathbb{Z}_-}$ associated to the reservoir system (3) that is defined by:

$$U_h^F:=h\circ U^F,\quad \text{where}\quad U^F(\mathbf{z}):=\mathbf{x},$$

with $\mathbf{z}\in(D_d)^{\mathbb{Z}_-}$ and $\mathbf{x}\in(D_N)^{\mathbb{Z}_-}$ linked by the first equation in (3) via the ESP. It is easy to show that reservoir filters are automatically causal and time-invariant (Grigoryeva and Ortega, 2018b, Proposition 2.1) and hence determine a reservoir functional $H_h^F:(D_d)^{\mathbb{Z}_-}\longrightarrow\mathbb{R}^m$.

As the following Proposition shows, a sufficient condition guaranteeing that the echo state property holds is that D_N is a closed ball and that the map F is continuous and a contraction in the first argument.

Proposition 1 *Let $S>0$, $\overline{B_S}=\{\mathbf{x}\in\mathbb{R}^N\colon\|\mathbf{x}\|_2\leq S\}$ and suppose that $F\colon\overline{B_S}\times D_d\to\overline{B_S}$ is continuous. Assume that F is a contraction in the first argument, that is, there exists $0<r<1$ such that for all $\mathbf{x}_1,\mathbf{x}_2\in\overline{B_S}$, $\mathbf{z}\in D_d$ it holds that*

$$\|F(\mathbf{x}_1,\mathbf{z})-F(\mathbf{x}_2,\mathbf{z})\|_2\leq r\|\mathbf{x}_1-\mathbf{x}_2\|_2. \tag{4}$$

Then the system (3) has the echo state property and hence its first equation determines a unique causal and time-invariant filter $U^F:(D_d)^{\mathbb{Z}_-}\longrightarrow(\overline{B_S})^{\mathbb{Z}_-}$ as well as a functional $H^F\colon(D_d)^{\mathbb{Z}_-}\to\overline{B_S}$ that are continuous (where both $(D_d)^{\mathbb{Z}_-}$ and $(\overline{B_S})^{\mathbb{Z}_-}$ are equipped with the product topologies).

Remark 2 If we have a continuous function $F\colon \mathbb{R}^N \times D_d \to \mathbb{R}^N$ which satisfies (4) for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^N$ and $\|F(\mathbf{0}, \mathbf{z})\|_2 \leq c$ for all $\mathbf{z} \in D_d$ and a certain $c > 0$, then choosing any $S \geq c/(1-r)$ it is easy to see that for all $\mathbf{u} \in \overline{B_S}$

$$\|F(\mathbf{u}, \mathbf{z})\|_2 \leq \|F(\mathbf{u}, \mathbf{z}) - F(\mathbf{0}, \mathbf{z})\|_2 + \|F(\mathbf{0}, \mathbf{z})\|_2 \leq r\|\mathbf{u}\|_2 + c \leq rS + c \leq S.$$

Thus, $F(\overline{B_S} \times D_d) \subset \overline{B_S}$ and so the restriction of F to $\overline{B_S} \times D_d$ satisfies the assumptions of Proposition 1.

2.3 Families of reservoir systems

The following paragraphs introduce various families of reservoir systems that appear in applications. We shall later on explicitly construct for these specific families the risk bounds contained in the main results of the paper.

Reservoir systems with linear reservoir maps (LRC)

In this case one associates to each input signal $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$ an output $\mathbf{y} \in (\mathbb{R}^m)^{\mathbb{Z}_-}$ via the two recurrent equations

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + C\mathbf{z}_t + \boldsymbol{\zeta}, \tag{5}$$

$$\mathbf{y}_t = h(\mathbf{x}_t), \tag{6}$$

with $t \in \mathbb{Z}_-$ and $A \in \mathbb{M}_N, C \in \mathbb{M}_{N,d}, \boldsymbol{\zeta} \in \mathbb{R}^N$. Systems with linear reservoir maps of the type (5) have been vastly studied in the literature in numerous contexts and under different denominations. In the RC setting, systems of the type (5)-(6) with polynomial readout maps $h : D_N \longrightarrow \mathbb{R}^m$ have been proved in Grigoryeva and Ortega (2018a) to be universal approximators in the category of fading memory filters either when presented with uniformly bounded inputs in the deterministic setup (see Corollary 3.4) or with almost surely uniformly bounded stochastic inputs (see Corollary 4.8). These boundedness hypotheses have been dropped in Gonon and Ortega (2020b) by considering density with respect to L^p norms, $1 \leq p < \infty$, defined using the law of the input data generating process. Sufficient conditions which ensure the echo state property and the fading memory property for these systems have been established (see Section 3.1 in Grigoryeva and Ortega, 2019). More specifically, consider the reservoir map $F^{A,C,\boldsymbol{\zeta}} : D_N \times D_d \longrightarrow D_N$ of the system (5)-(6) given by

$$F^{A,C,\boldsymbol{\zeta}}(\mathbf{x}, \mathbf{z}) = A\mathbf{x} + C\mathbf{z} + \boldsymbol{\zeta}.$$

It is easy to see that $F^{A,C,\boldsymbol{\zeta}}$ is a contraction in the first entry whenever the matrix A satisfies $|||A|||_2 < 1$. For these systems we consider only the case of uniformly bounded input signals:

Case with uniformly bounded inputs. Suppose now $|||A|||_2 < 1$. If the inputs are uniformly bounded, that is, if $D_d = \overline{B_M}$ for some $M > 0$ and so $\mathbf{z} \in K_M$ with $K_M := \left\{\mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-} \,|\, \|\mathbf{z}_t\|_2 \leq M \text{ for all } t \in \mathbb{Z}_-\right\}$, then the reservoir system (5)-(6) has the echo state property and defines a unique causal and time-invariant reservoir filter $U^{A,C,\boldsymbol{\zeta}} : K_M \longrightarrow (D_N)^{\mathbb{Z}_-}$ given by $U^{A,C,\boldsymbol{\zeta}}(\mathbf{z})_t := \sum_{j=0}^{\infty} A^j(C\mathbf{z}_{t-j} + \boldsymbol{\zeta})$, $t \in \mathbb{Z}_-$. Here $D_N = \overline{B_{M_F}}$ with $M_F = (|||C|||_2 M + \|\boldsymbol{\zeta}\|_2)/(1 - |||A|||_2)$ (see Remark 2 and part **(ii)** in the first example in Section 4.1 of Grigoryeva and Ortega, 2019). In particular, the corresponding functional

$H^{A,C,\boldsymbol{\zeta}}\colon K_M \to D_N$ satisfies that $\|H^{A,C,\boldsymbol{\zeta}}(\mathbf{z})\|_2 \leq M_F$ for all $\mathbf{z} \in K_M$. Additionally, it can be shown that the reservoir system (5)-(6) has the fading memory property with respect to any weighting sequence.

In what follows we consider a particular subfamily of systems (5)-(6), namely *reservoir systems with linear reservoir and linear readout maps*, in which case $h : D_N \longrightarrow \mathbb{R}^m$ is given by applying $W \in \mathbb{M}_{m,N}$.

Echo State Networks (ESN)

Echo State Networks (Matthews, 1992; Jaeger and Haas, 2004) are a family of reservoir systems that exhibit excellent performance in many practical applications and have been recently proved to have universal approximation properties. More specifically, Grigoryeva and Ortega (2018b) proved ESNs to be universal in the category of fading memory filters with semi-infinite uniformly bounded inputs in a deterministic setup, and Gonon and Ortega (2020b) obtained universality results for ESNs in the stochastic situation with respect to L^p-type criteria for stochastic discrete-time semi-infinite inputs.

An echo state network of dimension $N \in \mathbb{N}^+$ with reservoir matrix $A \in \mathbb{M}_N$, input mask $C \in \mathbb{M}_{N,d}$, input shift $\boldsymbol{\zeta} \in \mathbb{R}^N$, and readout matrix $W \in \mathbb{M}_{m,N}$ is the system

$$\mathbf{x}_t = \boldsymbol{\sigma}(A\mathbf{x}_{t-1} + C\mathbf{z}_t + \boldsymbol{\zeta}), \tag{7}$$

$$\mathbf{y}_t = W\mathbf{x}_t, \tag{8}$$

which for each $t \in \mathbb{Z}_-$ transforms the input $\mathbf{z}_t \in D_d \subset \mathbb{R}^d$ into the reservoir state $\mathbf{x}_t \in D_N \subset \mathbb{R}^N$ and, consequently, into the corresponding output $\mathbf{y}_t \in \mathbb{R}^m$. The reservoir map $F^{\sigma,A,C,\boldsymbol{\zeta}} : D_N \times D_d \longrightarrow D_N$ of the system (7)-(8) is given by

$$F^{\sigma,A,C,\boldsymbol{\zeta}}(\mathbf{x}, \mathbf{z}) = \boldsymbol{\sigma}(A\mathbf{x} + C\mathbf{z} + \boldsymbol{\zeta}),$$

where $\boldsymbol{\sigma}\colon \mathbb{R}^N \to \mathbb{R}^N$ is defined by the componentwise application of a given ***activation function*** $\sigma\colon \mathbb{R} \to \mathbb{R}$. Throughout, we assume that σ is Lipschitz-continuous with Lipschitz-constant L_σ. It is straightforward to verify that $F^{\sigma,A,C,\boldsymbol{\zeta}}$ is a contraction in the first entry whenever $L_\sigma \|\!|A|\!\|_2 < 1$. The sufficient conditions which ensure the echo state and the fading memory properties of (7)-(8) have been also carefully studied in the literature (see Buehner and Young, 2006; Yildiz et al., 2012; Grigoryeva and Ortega, 2018b, for details) and depend both on the type of the activation function $\sigma\colon \mathbb{R} \to \mathbb{R}$ and on the type of the input presented to the network. We consider the following two cases:

Case with arbitrary input signals and bounded activation function. In this situation, D_d is arbitrary and so generic input signals $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$ are considered, but we assume that the range of the activation function σ is bounded and contained in $[\sigma_{min}, \sigma_{max}]$ with $\sigma_{min} < \sigma_{max} \in \mathbb{R}$. Then, by Proposition 1, the condition $L_\sigma \|\!|A|\!\|_2 < 1$ suffices to ensure that the system (7)-(8) has the echo state property and hence defines a unique causal and time-invariant filter $U^{\sigma,A,C,\boldsymbol{\zeta}} : (D_d)^{\mathbb{Z}_-} \longrightarrow (D_N)^{\mathbb{Z}_-}$ as well as a functional $H^{\sigma,A,C,\boldsymbol{\zeta}}\colon (D_d)^{\mathbb{Z}_-} \to D_N$ that are additionally continuous with respect to the product topologies on the spaces $(D_d)^{\mathbb{Z}_-}$ and $(D_N)^{\mathbb{Z}_-}$. Here $D_N = \overline{B_{M_F}}$ with $M_F = \sqrt{N}\max(|\sigma_{min}|, |\sigma_{max}|)$ and in particular, we obviously have that $\|H^{\sigma,A,C,\boldsymbol{\zeta}}(\mathbf{z})\|_2 \leq M_F$.

Case with uniformly bounded inputs. Suppose that $L_\sigma |||A|||_2 < 1$ and $D_d = \overline{B_M}$ for some $M > 0$ and so the inputs are uniformly bounded, that is, $\mathbf{z} \in K_M$ with $K_M := \left\{\mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-} | \|\mathbf{z}_t\|_2 \leq M \text{ for all } t \in \mathbb{Z}_-\right\}$. In this case the reservoir system (7)-(8) has the echo state property and defines a unique causal and time-invariant reservoir filter $U^{\sigma,A,C,\boldsymbol{\zeta}} : K_M \longrightarrow (D_N)^{\mathbb{Z}_-}$ as well as a functional $H^{\sigma,A,C,\boldsymbol{\zeta}} \colon K_M \to D_N$, where $D_N = \overline{B_{M_F^1}}$ with $M_F^1 := [L_\sigma(|||C|||_2 M + \|\boldsymbol{\zeta}\|_2) + \sqrt{N}\sigma(0)]/(1 - L_\sigma |||A|||_2)$ (see Proposition 1 and Remark 2 or part **(ii)** in the first example in Section 4.1 of Grigoryeva and Ortega, 2019). Additionally, the fading memory property holds with respect to any weighting sequence. Note that these results hold true even though the range of the activation function is not assumed to be bounded and $\|H^{\sigma,A,C,\boldsymbol{\zeta}}(\mathbf{z})\|_2 \leq M_F$ with $M_F = M_F^1$ when the activation function σ has an unbounded range and with $M_F = \min(M_F^1, \sqrt{N}\max(|\sigma_{min}|, |\sigma_{max}|))$, otherwise.

State-Affine Systems (SAS)

The so-called homogeneous state-affine systems have been first introduced in the systems theory literature and were shown to exhibit universality properties in the discrete-time setting for compact times (see Fliess and Normand-Cyrot, 1980; Sontag, 1979a,b). A non-homogeneous version of these systems was introduced in Grigoryeva and Ortega (2018a), where they were proved to be universal approximants in the category of fading memory filters for the non-compact discrete-time deterministic setup. Trigonometric state-affine systems were later on studied in a stochastic setup in Gonon and Ortega (2020b), where their universality for stochastic discrete-time semi-infinite inputs with respect to L^p-criteria was established. State-affine systems serve as an excellent example of reservoir systems with easy-to-train linear readouts and even though little is known about their empirical performance in learning tasks, we find it important to provide explicit risk bounds for this family. In the rest of the paper we reserve the name State-Affine Systems (SAS) for the non-homogeneous version if not stated otherwise and leave the trigonometric family for future work.

The following notation for multivariate polynomials will be used: for any multi-index $\boldsymbol{\alpha} \in \mathbb{N}^d$ and any $\mathbf{z} \in \mathbb{R}^d$, we write $\mathbf{z}^{\boldsymbol{\alpha}} := z_1^{\alpha_1} \cdots z_d^{\alpha_d}$. Furthermore, the space $\mathbb{M}_{N,M}[\mathbf{z}]$, $N, M \in \mathbb{N}^+$, of polynomials in the variable $\mathbf{z} \in \mathbb{R}^d$ with matrix coefficients in $\mathbb{M}_{N,M}$ is the set of elements p of the form

$$p(\mathbf{z}) = \sum_{\boldsymbol{\alpha} \in V_p} \mathbf{z}^{\boldsymbol{\alpha}} A_{\boldsymbol{\alpha}}, \quad \mathbf{z} \in \mathbb{R}^d,$$

where $V_p \subset \mathbb{N}^d$ is a finite subset and the elements $A_{\boldsymbol{\alpha}} \in \mathbb{M}_{N,M}$ are matrix coefficients. The degree $\deg(p)$ of the polynomial p is defined as

$$\deg(p) = \max_{\boldsymbol{\alpha} \in V_p} \{\|\boldsymbol{\alpha}\|_1\}, \text{ where } \|\boldsymbol{\alpha}\|_1 := \alpha_1 + \cdots + \alpha_d.$$

We also define the following norm on $\mathbb{M}_{N,M}[\mathbf{z}]$:

$$|||p||| = \max_{\boldsymbol{\alpha} \in V_p} |||A_{\boldsymbol{\alpha}}|||_2. \tag{9}$$

The non-homogeneous state-affine system (SAS) of dimension $N \in \mathbb{N}^+$ associated to two given polynomials $p \in \mathbb{M}_{N,N}[\mathbf{z}]$ and $q \in \mathbb{M}_{N,1}[\mathbf{z}]$ with matrix and vector coefficients, respectively, is the reservoir system determined by the following state-space transformation

of each input signal $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$ into the output signal $\mathbf{y} \in (\mathbb{R}^m)^{\mathbb{Z}_-}$,

$$\mathbf{x}_t = p(\mathbf{z}_t)\mathbf{x}_{t-1} + q(\mathbf{z}_t), \tag{10}$$

$$\mathbf{y}_t = W\mathbf{x}_t, \tag{11}$$

for $t \in \mathbb{Z}_-$, with $W \in \mathbb{M}_{m,N}$ the readout map. The reservoir map $F^{p,q} : D_N \times D_d \longrightarrow D_N$ of the system (10)-(11) is given by

$$F^{p,q}(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})\mathbf{x} + q(\mathbf{z}). \tag{12}$$

Additionally, we define

$$M_p := \sup_{\mathbf{z} \in D_d} |||p(\mathbf{z})|||_2,$$

$$M_q := \sup_{\mathbf{z} \in D_d} |||q(\mathbf{z})|||_2.$$

First, we notice that for regular SAS defined by nontrivial polynomials, the set D_d needs to be bounded in order for M_p and M_q to be finite. It is easy to see that F in (12) is a contraction in the first entry with constant M_p whenever $M_p < 1$, which is a condition that we will assume holds true together with $M_q < \infty$ in the next paragraph.

Case with uniformly bounded input signals. Let $D_d = \overline{B_M}$ for some $M > 0$ so that we consider inputs $\mathbf{z} \in K_M$ with $K_M := \left\{\mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-} | \|\mathbf{z}_t\|_2 \leq M \text{ for all } t \in \mathbb{Z}_-\right\}$. In that case the system (10)-(11) has the echo state property and determines (see Proposition 1 and Remark 2 or part **(ii)** in the third example in Section 4.1 of Grigoryeva and Ortega, 2019) a unique reservoir filter $U^{p,q} : K_M \longrightarrow (D_N)^{\mathbb{Z}_-}$ as well as a functional $H^{p,q} \colon K_M \to D_N$, where $D_N = \overline{B_{M_F}}$ with $M_F = M_q/(1-M_p)$. In addition, the fading memory property holds with respect to any weighting sequence. Moreover, in this case the filter can be explicitly written as $(U^{p,q}(\mathbf{z}))_t = \sum_{j=0}^{\infty}(\prod_{k=0}^{j-1} p(\mathbf{z}_{t-k}))q(\mathbf{z}_{t-j})$, $t \in \mathbb{Z}_-$, and $\|H^{p,q}(\mathbf{z})\|_2 \leq M_F$, for all $\mathbf{z} \in K_M$.

3. The learning problem for reservoir computing systems

In this paper we work in the setting of supervised learning in a probabilistic framework and our goal is to provide performance estimates for reservoir systems from the statistical learning theory perspective. With that in mind, we start this section by stating the general learning problem for systems with stochastic input and target signals. We then introduce three alternative assumptions on the weak dependence of input and output processes which will be assumed later on in the paper and provide examples of important time series models that satisfy the conditions under consideration. We define the statistical risk and its empirical analogs for reservoir functionals and motivate the need for generalization error bounds. More specifically, on the one hand the in-class generalization error (risk) can be used to bound the estimation error of a class. On the other hand, whenever the learner follows the empirical risk minimization (ERM) strategy to select the reservoir computing system within the RC hypothesis class based on minimization of the empirical (training) error, generalization error bounds can be used to prove the weak universal risk-consistency of ERM for reservoir systems. If the inputs are i.i.d. (which is a particular case of our setup), this definition is essentially equivalent to saying that the hypothesis class of reservoir functionals is a (weak) uniform Glivenko-Cantelli class (see for example Mukherjee et al., 2006).

3.1 General setup of the learning procedure

Input and target stochastic processes. We fix a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ on which all random variables are defined. The triple consists of the sample space Ω, which is the set of possible outcomes, the σ-algebra $\mathcal{A}$ (a set of subsets of Ω (events)), and a probability measure $\mathbb{P} : \mathcal{A} \longrightarrow [0,1]$. The input and target signals are modeled by discrete-time stochastic processes $\mathbf{Z} = (\mathbf{Z}_t)_{t\in\mathbb{Z}_-}$ and $\mathbf{Y} = (\mathbf{Y}_t)_{t\in\mathbb{Z}_-}$ taking values in $D_d \subset \mathbb{R}^d$ and $\mathbb{R}^m$, respectively. Moreover, we write $\mathbf{Z}(\omega) = (\mathbf{Z}_t(\omega))_{t\in\mathbb{Z}_-}$ and $\mathbf{Y}(\omega) = (\mathbf{Y}_t(\omega))_{t\in\mathbb{Z}_-}$ for each outcome $\omega \in \Omega$ to denote the realizations or sample paths of $\mathbf{Z}$ and $\mathbf{Y}$, respectively. Since $\mathbf{Z}$ can be seen as a random sequence in $D_d \subset \mathbb{R}^d$, we write interchangeably $\mathbf{Z} : \mathbb{Z}_- \times \Omega \longrightarrow D_d$ and $\mathbf{Z} : \Omega \longrightarrow (D_d)^{\mathbb{Z}_-}$. The latter is necessarily measurable with respect to the Borel σ-algebra induced by the product topology in $(D_d)^{\mathbb{Z}_-}$. The same applies to the analogous assignments involving $\mathbf{Y}$.

Hypothesis class $\mathcal{H}$, loss functions, statistical, and empirical risk. Let $\mathcal{F}$ be the class of all measurable functionals $H\colon (D_d)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}^m$, $D_d \subset \mathbb{R}^d$, that is $\mathcal{F} := \{H\colon (D_d)^{\mathbb{Z}_-} \longrightarrow \mathbb{R}^m \mid H \text{ is measurable}\}$. Consider a smaller ***hypothesis class*** $\mathcal{H}$ of admissible functionals $\mathcal{H} \subset \mathcal{F}$. For a fixed ***loss***, that is, a measurable function[1] $L\colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ and for any functional $H \in \mathcal{F}$ we define the ***statistical risk*** (sometimes just referred to as ***risk***) or ***generalization error*** associated with H as

$$R(H) := \mathbb{E}[L(H(\mathbf{Z}), \mathbf{Y}_0)], \tag{13}$$

where by definition the expectation is taken with respect to the joint law of $(\mathbf{Z}, \mathbf{Y})$. The ultimate goal of the learning procedure consists in determining the ***Bayes functional*** $H^*_{\mathcal{F}} \in \mathcal{F}$ that exhibits the minimal statistical risk (***Bayes risk***) in the class of all measurable functionals, which we denote as

$$R^*_{\mathcal{F}} := R(H^*_{\mathcal{F}}) = \inf_{H\in\mathcal{F}} R(H). \tag{14}$$

Even though this task is generally infeasible, one may hope to solve it for the ***best-in-class*** functional $H^*_{\mathcal{H}} \in \mathcal{H}$ with the minimal associated in-class statistical risk (***Bayes in-class risk***), which is assumed achievable, and which we denote as

$$R^*_{\mathcal{H}} := R(H^*_{\mathcal{H}}) = \inf_{H\in\mathcal{H}} R(H).$$

The standard learning program is then based on the following error decomposition. For any $H \in \mathcal{H}$ we can write that

$$R(H) - R^*_{\mathcal{F}} = (R(H) - R^*_{\mathcal{H}}) + (R^*_{\mathcal{H}} - R^*_{\mathcal{F}}),$$

where the first term is called the ***estimation error*** and the second one is the ***approximation error***. In this paper we focus on upper bounds of the estimation component, while the same problem for the approximation error will be treated in the forthcoming work, namely

1. It is customary in the literature to consider nonnegative loss functions. This automatically guarantees that the expectation in (13) is well-defined, although it is not necessarily finite. In this paper, for the sake of mathematical convenience, we allow for general real-valued loss functions but carefully address technical questions where relevant.

Gonon, Grigoryeva, and Ortega (2020). We emphasize that since the universal approximation properties of reservoir systems have been established in numerous situations (see the introduction and Section 2.3) the approximation error can be made arbitrarily small by choosing an appropriate hypothesis class $\mathcal{H}$.

The distribution of $(\mathbf{Z}, \mathbf{Y})$ is generally unknown, and hence computing the risks (13) or (14) is in practice infeasible. This implies, in particular, that the estimation error cannot be explicitly evaluated. Therefore, the usual procedure is in this case to use an empirical counterpart for (13) which can be computed using a training dataset.

Suppose that a training sample for both the input and the target discrete-time stochastic processes is available up to some $n \in \mathbb{N}^+$ steps into the past, namely $(\mathbf{Z}_{-i}, \mathbf{Y}_{-i})_{i\in\{0,\ldots,n-1\}}$. For each time step $i \in \{0, \ldots, n-1\}$ we define the ***truncated training sample*** for the input stochastic process $\mathbf{Z}$ as

$$\mathbf{Z}_{-i}^{-n+1} := (\ldots, \mathbf{0}, \mathbf{0}, \mathbf{Z}_{-n+1}, \ldots, \mathbf{Z}_{-i-1}, \mathbf{Z}_{-i}). \tag{15}$$

In this time series context the ***training error*** or the ***empirical risk*** analog $\widehat{R}_n(H)$ of (13) is given by

$$\widehat{R}_n(H) = \frac{1}{n}\sum_{i=0}^{n-1} L(H(\mathbf{Z}_{-i}^{-n+1}), \mathbf{Y}_{-i}) = \frac{1}{n}\sum_{i=0}^{n-1} L(U_H(\mathbf{Z}_0^{-n+1})_{-i}, \mathbf{Y}_{-i}), \tag{16}$$

where U_H denotes the filter associated to the functional H as introduced in (2). In what follows we will also make use of what we call its ***idealized empirical risk*** version defined as

$$\widehat{R}_n^{\infty}(H) = \frac{1}{n}\sum_{i=0}^{n-1} L(H(\mathbf{Z}_{-i}^{-\infty}), \mathbf{Y}_{-i}) = \frac{1}{n}\sum_{i=0}^{n-1} L(U_H(\mathbf{Z}_0^{-\infty})_{-i}, \mathbf{Y}_{-i}), \tag{17}$$

which makes use of a larger training sample containing all the past values of the input process $\mathbf{Z}$.

Remark 3 The results of this paper are also valid if one replaces the zero elements in the truncated training sample (15) by an arbitrary sequence (deterministic, random, or dependent on the training sample). More specifically, consider an arbitrary function $\mathcal{I}\colon (D_d)^{\mathbb{Z}_-} \to (D_d)^{\mathbb{Z}_-}$ that we use to extend the input training sample, for each $i \in \{0, \ldots, n-1\}$, as

$$\widetilde{\mathbf{Z}}_{-i}^{-n+1} = (\ldots, (\mathcal{I}(\mathbf{Z}_0^{-n+1}))_{-1}, (\mathcal{I}(\mathbf{Z}_0^{-n+1}))_0, \mathbf{Z}_{-n+1}, \ldots, \mathbf{Z}_{-i-1}, \mathbf{Z}_{-i}), \tag{18}$$

and use this sample to define the empirical risk as in (16). Later on in Proposition 5, we show that the difference between the empirical risk (16) and its idealized counterpart (17) can be made arbitrarily small under various assumptions that we shall consistently invoke. The proof of that result remains valid for the more general definition of empirical risk using (18). Moreover, that result is used to justify why, in the rest of the paper, it will be sufficient to work almost exclusively with the idealized empirical risk (17).

Risk bounds and risk-consistency. As we have already discussed, the learner is interested in obtaining upper bounds of the estimation error $(R(H) - R_{\mathcal{H}}^*)$. In many cases, these bounds can be constructed by bounding the statistical risk (generalization error) or

$$\Delta_n := \sup_{H\in\mathcal{H}} \{R(H) - \widehat{R}_n(H)\}.$$

An upper bound of Δ_n (or its version with the absolute value of the difference) allows to quantify the worst in-class error between the statistical risk and its empirical analog. We emphasize that bounding this worst in-class error gives guarantees of performance for any learning algorithm which builds upon the idea of using the empirical risk to pick a concrete $\widehat{H}_n$ out of the hypothesis class $\mathcal{H}$ based on available training data.

A standard example of such learning rules is the so-called ***empirical risk minimization*** (ERM) principle for which generalization error bounds or bounds for Δ_n can be used in a straightforward manner to bound the estimation error and, moreover, to establish some important consistency properties.

More specifically, in the ERM procedure the learner chooses the desired functional $\widehat{H}_n$ out of the hypothesis class $\mathcal{H}$ of the admissible ones using (16) (the empirical version of (13)), that is,

$$\widehat{H}_n = \arg\min_{H\in\mathcal{H}} \widehat{R}_n(H), \tag{19}$$

which is well defined provided that such a minimizer exists and is unique; otherwise one may define $\widehat{H}_n$ to be an ϵ-minimizer of the empirical risk (see Alon et al., 1997, for details). We say that the ERM is ***strongly consistent*** within the hypothesis class $\mathcal{H}$ if the generalization error $R(\widehat{H}_n)$ (or statistical risk) and the training error $\widehat{R}_n(\widehat{H}_n)$ (or empirical risk) as defined in (13) and in (16), respectively, for a sequence of functionals $(\widehat{H}_n)_{n\in\mathbb{N}}$ picked by ERM from $\mathcal{H}$ using random samples of increasing length, both converge almost surely to the Bayes in-class risk $R^*_{\mathcal{H}}$ in (14), that is

$$\lim_{n\to\infty} R(\widehat{H}_n) = R^*_{\mathcal{H}} \text{ a.s.} \tag{20}$$

and

$$\lim_{n\to\infty} \widehat{R}_n(\widehat{H}_n) = R^*_{\mathcal{H}} \text{ a.s.} \tag{21}$$

When no assumptions on the distribution of $(\mathbf{Z},\mathbf{Y})$ are used to prove (20) and (21), then this means that (20) and (21) hold for all distributions and one talks about ***universal strong risk-consistency*** of the ERM principle over the class $\mathcal{H}$. This is essentially the case in our setting, since we are working in a semi-agnostic setup and only invoke assumptions on the temporal dependence (but not on the marginal distributions of the input and target stochastic processes $(\mathbf{Z},\mathbf{Y})$).

A standard approach to proving the strong risk-consistency of the ERM procedure for the hypothesis class of functionals $\mathcal{H}$ consists in finding a sequence $(\eta_n)_{n\in\mathbb{N}}$ converging to zero for which the inequality

$$\overline{\Delta}_n := \sup_{H\in\mathcal{H}} |\widehat{R}_n(H) - R(H)| \leq \eta_n,$$

holds $\mathbb{P}$-a.s. To see that this implies (20) and (21) one notes the following inequalities:

$$\begin{aligned} R(\widehat{H}_n) - R^*_{\mathcal{H}} &= \left(R(\widehat{H}_n) - \widehat{R}_n(\widehat{H}_n)\right) + \left(\widehat{R}_n(\widehat{H}_n) - \widehat{R}_n(H^*_{\mathcal{H}})\right) + \left(\widehat{R}_n(H^*_{\mathcal{H}}) - R^*_{\mathcal{H}}\right) \\ &\leq 2\eta_n + \left(\widehat{R}_n(\widehat{H}_n) - \widehat{R}_n(H^*_{\mathcal{H}})\right) \leq 2\eta_n, \end{aligned} \tag{22}$$

where the last inequality follows from the fact that, by definition (19), $\widehat{H}_n$ is a minimizer of the empirical risk $\widehat{R}_n$.

In the context of reservoir systems, we shall be working with a ***weak version of consistency*** which imposes all the convergence conditions to hold only in probability. In what follows we devise bounds for $\overline{\Delta}_n$ that allow us to establish the risk-consistency of the ERM procedure for reservoir systems. Additionally, we formulate high-probability bounds for $\overline{\Delta}_n$ which provide us with convergence rates for the ERM-based estimation of RC systems that, to our knowledge, are not yet available in the literature. It is well known that in some cases (for small classes with zero Bayes risk, see for example Bartlett et al., 2006) the argument that we just discussed results in unreasonably slow rates. We defer the discussion of possible refinements of the rates obtained in this paper to future projects.

3.2 Learning procedure for reservoir systems

The following paragraphs describe the implementation of the empirical risk minimization procedure in the setting of reservoir computing. We spell out the assumptions needed to derive the results in the next sections, construct the hypothesis classes, and set up the ERM learning strategy for the different families of reservoir systems discussed in Section 2.3.

Input and target stochastic processes. For both the input $\mathbf{Z}$ and the target $\mathbf{Y}$ processes we assume a ***causal Bernoulli shift*** structure (see for instance Dedecker et al., 2007; Alquier and Wintenberger, 2012). More precisely, for $I = y, z$ and $q_I \in \mathbb{N}^+$ suppose that the so-called causal functional $G^I \colon (\mathbb{R}^{q_I})^{\mathbb{Z}_-} \to D_{o_I}$ (with $o_z = d$ and $D_{o_y} = \mathbb{R}^m$) is measurable and that $\boldsymbol{\xi} = ((\boldsymbol{\xi}_t^y, \boldsymbol{\xi}_t^z))_{t\in\mathbb{Z}_-}$ are independent and identically distributed $\mathbb{R}^{q_y} \times \mathbb{R}^{q_z}$-valued random variables. We assume then that the input $\mathbf{Z}$ and target processes $\mathbf{Y}$ are Bernoulli shifts, that is, they are the (strictly) stationary processes determined by

$$\begin{aligned} \mathbf{Z}_t &= G^z(\ldots, \boldsymbol{\xi}_{t-1}^z, \boldsymbol{\xi}_t^z), \quad t \in \mathbb{Z}_-, \\ \mathbf{Y}_t &= G^y(\ldots, \boldsymbol{\xi}_{t-1}^y, \boldsymbol{\xi}_t^y), \quad t \in \mathbb{Z}_-, \end{aligned} \tag{23}$$

with $\mathbb{E}[\|\mathbf{Z}_0\|_2] < \infty$, $\mathbb{E}[\|\mathbf{Y}_0\|_2] < \infty$.

Many processes derived from stationary innovation sequences have causal Bernoulli shift structure including some that are of non-mixing type (see for instance the introduction in Dedecker et al., 2007). In order to obtain risk bounds for reservoir functionals as learning models, we need to additionally impose assumptions on the weak dependency of the processes (23). More specifically, each of the three main results provided in the next section is formulated under a different weak dependence assumption which we now spell out in detail.

We start with the strongest assumption of the three but which will allow us to obtain the strongest conclusions in terms of risk bounds for reservoir systems.

Assumption 1 *For $I = y, z$ the functional G^I is L_I-Lipschitz continuous when restricted to $(\ell_-^{1,w^I}(\mathbb{R}^{q_I}), \|\cdot\|_{1,w^I})$ for some strictly decreasing weighting sequence $w^I : \mathbb{N} \longrightarrow (0, 1]$ with finite mean, that is, $\sum_{j\in\mathbb{N}} j w_j^I < \infty$. More specifically, there exists $L_I > 0$ such that for all $\boldsymbol{u}^I = (\boldsymbol{u}_t^I)_{t\in\mathbb{Z}_-} \in \ell_-^{1,w^I}(\mathbb{R}^{q_I})$ and $\boldsymbol{v}^I = (\boldsymbol{v}_t^I)_{t\in\mathbb{Z}_-} \in \ell_-^{1,w^I}(\mathbb{R}^{q_I})$ it holds that*

$$\|G^I(\boldsymbol{u}^I) - G^I(\boldsymbol{v}^I)\|_2 \leq L_I \left\|\boldsymbol{u}^I - \boldsymbol{v}^I\right\|_{1,w^I}. \tag{24}$$

Additionally, let the innovations in (23) satisfy $\mathbb{E}[\|\boldsymbol{\xi}_0^I\|_2] < \infty$ for $I = y, z$.

The following example shows that one can easily construct causal Bernoulli shifts using reservoir functionals.

Example 1 (Causal Bernoulli shifts out of reservoir functionals) Let $I \in \{y, z\}$ and consider a reservoir system of the type (3) (see also the examples in Section 2.3) determined by the Lipschitz-continuous reservoir map $F\colon D_N \times D_d \longrightarrow D_N$ with $D_d \subset \mathbb{R}^d$, $D_N \subset \mathbb{R}^N$. Assume, additionally, that F is a r-contraction on the first entry and denote by $L > 0$ the Lipschitz constant of F with respect to the second entry. Let $w : \mathbb{N} \longrightarrow (0, 1]$ be a strictly decreasing weighting sequence with finite mean, that is $\sum_{j\in\mathbb{N}} jw_j < \infty$, and a finite associated inverse decay ratio L_w (see Section 2.1). Let now $V_d \subset (D_d)^{\mathbb{Z}_-} \cap \ell^{w,1}_-(\mathbb{R}^d)$ be a time-invariant set and consider inputs $\mathbf{z} \in V_d$. Suppose that the reservoir system (3) has a solution $(\mathbf{x}^0, \mathbf{z}^0) \in (D_N)^{\mathbb{Z}_-} \times V_d$, that is, $\mathbf{x}^0_t = F(\mathbf{x}^0_{t-1}, \mathbf{z}^0_t)$, for all $t \in \mathbb{Z}_-$. Then, by Theorem 4.1 and Remark 4.4 in Grigoryeva and Ortega (2019), if

$$rL_w < 1,$$

then the reservoir system associated to F with inputs in V_d has the echo state property and hence determines a unique continuous, causal, and time-invariant reservoir filter $U^F : (V_d, \|\cdot\|_{1,w}) \longrightarrow ((D_N)^{\mathbb{Z}_-}, \|\cdot\|_{1,w})$ which is Lipschitz-continuous with constant

$$L_{U^F} := \frac{L}{1 - rL_w}.$$

It hence also has the fading memory property with respect to w. The Lipschitz continuity of the filter U^F implies that the associated functional H_{U^F} is also Lipschitz-continuous with the same Lipschitz constant ((see Proposition 3.7 in Grigoryeva and Ortega, 2019). Taking $G^I := H_{U^F}$, it is easy to see that (24) indeed holds with $L_I = L_{U^F}$.

The next assumption is weaker and it is satisfied by many discrete-time stochastic processes. The results that we obtain in the following sections invoking this type of weak dependence will be also less strong than under Assumption 1.

Assumption 2 *For $I = y, z$ denote by $(\widetilde{\boldsymbol{\xi}}^I_t)_{t\in\mathbb{Z}_-}$ an independent copy of $(\boldsymbol{\xi}^I_t)_{t\in\mathbb{Z}_-}$ and define*

$$\theta^I(\tau) := \mathbb{E}[\|G^I(\ldots, \boldsymbol{\xi}^I_{-1}, \boldsymbol{\xi}^I_0) - G^I(\ldots, \widetilde{\boldsymbol{\xi}}^I_{-\tau-1}, \widetilde{\boldsymbol{\xi}}^I_{-\tau}, \boldsymbol{\xi}^I_{-\tau+1}, \ldots, \boldsymbol{\xi}^I_0)\|_2], \quad \tau \in \mathbb{N}^+. \tag{25}$$

Assume that for $I = y, z$ there exist $\lambda_I \in (0, 1)$ and $C_I > 0$ such that it holds that

$$\theta^I(\tau) \leq C_I \lambda^\tau_I, \quad \textit{for all } \tau \in \mathbb{N}^+. \tag{26}$$

Remark 4 Note that whenever the weighting sequence w^I in Assumption 1 can be chosen to be a geometric one, that is, $w^I_j = \lambda^j_I$, $j \in \mathbb{N}$ with $\lambda_I \in (0, 1)$, then Assumption 2 is also automatically satisfied. The argument proving this appears for instance in the proof of part **(i)** of Corollary 8.

The following example illustrates that for many widely used time series models this assumption does hold. In particular, we show that vector autoregressive VARMA processes with time-varying coefficients under mild conditions and, in particular, GARCH processes satisfy Assumption 2.

Example 2 (VARMA process with time-varying coefficients) Suppose $\mathbf{Z} = (\mathbf{Z}_t)_{t\in\mathbb{Z}_-}$ is a vector autoregressive process of first order with time-varying coefficients, which we write as

$$\mathbf{Z}_t = A_t\mathbf{Z}_{t-1} + \boldsymbol{\eta}_t, \quad t \in \mathbb{Z}_-, \tag{27}$$

where $(\boldsymbol{\eta}_t)_{t\in\mathbb{Z}_-} \sim$ IID with $\boldsymbol{\eta}_t \in \mathbb{R}^d$ and $\mathbb{E}[\|\boldsymbol{\eta}_0\|_2] < \infty$, and where $(A_t)_{t\in\mathbb{Z}_-} \sim$ IID with $A_t \in \mathbb{M}_d$ and $\mathbb{E}[|||A_0|||_2] < 1$. Under these hypotheses (see for instance Brandt, 1986; Bougerol and Picard, 1992, Theorem 1.1) there exists a unique stationary process satisfying (23) and (27) and $\mathbb{E}[\|\mathbf{Z}_0\|_2] < \infty$. Iterating (27) yields

$$\mathbf{Z}_0 = \boldsymbol{\eta}_0 + A_0\boldsymbol{\eta}_{-1} + \cdots + A_0\cdots A_{-\tau+1}\mathbf{Z}_{-\tau},$$

and so by definition, using the independence of $(A_t)_{t\in\mathbb{Z}_-}$ and stationarity, one gets

$$\theta^z(\tau) \le 2\mathbb{E}[|||A_0\cdots A_{-\tau+1}|||_2]\mathbb{E}[\|\mathbf{Z}_{-\tau}\|_2] \le 2\mathbb{E}[|||A_0|||_2]^\tau\mathbb{E}[\|\mathbf{Z}_0\|_2].$$

We now define $C_z := 2\mathbb{E}[\|\mathbf{Z}_0\|_2]$, $\lambda_z := \mathbb{E}[|||A_0|||_2]$ and immediately obtain that (26) indeed holds, that is, for all $\tau \in \mathbb{N}^+$

$$\theta^z(\tau) \le C_z\lambda_z^\tau,$$

as required.

We now consider a concrete example of an autoregressive process of the type (27) which is extensively used to describe and eventually to forecast the volatility of financial time series, namely the generalized autoregressive conditional heterostedastic (GARCH) family (Engle, 1982; Bollerslev, 1986; Francq and Zakoian, 2010).

Example 3 (GARCH process) Consider a GARCH(1,1) model given by the following equations:

$$r_t = \sigma_t\varepsilon_t, \quad \varepsilon_t \sim \mathrm{IID}(0,1), \quad t \in \mathbb{Z}_- \tag{28}$$

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta\sigma_{t-1}^2, \quad t \in \mathbb{Z}_- \tag{29}$$

with parameters that satisfy $\alpha, \beta, \omega \ge 0$, $\alpha + \beta < 1$, which guarantees the second order stationarity of the process $(r_t)_{t\in\mathbb{Z}_-}$ and the positivity of the conditional variances $(\sigma_t^2)_{t\in\mathbb{Z}_-}$. We now check if the GARCH(1,1) process in (28)-(29) falls in the framework (27) introduced in the previous example. Let $d = 2$ and define

$$\mathbf{Z}_t := \begin{pmatrix} r_t^2 \\ \sigma_t^2 \end{pmatrix}, \quad \boldsymbol{\eta}_t := \begin{pmatrix} \omega\varepsilon_t^2 \\ \omega \end{pmatrix}, \quad A_t := \begin{pmatrix} \alpha\varepsilon_t^2 & \beta\varepsilon_t^2 \\ \alpha & \beta \end{pmatrix}, \quad t \in \mathbb{Z}_-.$$

It is easy to verify that with this choice of matrix A_t, $t \in \mathbb{Z}_-$, one has $\mathbb{E}[|||A_0|||_2] = \mathbb{E}[\alpha\varepsilon_0^2 + \beta] = \alpha + \beta < 1$, by the stationarity condition. Additionally, $\mathbb{E}[\|\boldsymbol{\eta}_0\|_2] = \omega\mathbb{E}[\sqrt{\varepsilon_0^4 + 1}] \le \omega\mathbb{E}[\varepsilon_0^2 + 1] = 2\omega < \infty$. Hence, the GARCH(1,1) model in (28)-(29) can be represented as (27) and automatically satisfies Assumption 2.

Assumption 3 *Assume that for $I = y, z$ there exist $\alpha_I \in (0,\infty)$ and $C_I > 0$ such that,*

$$\theta^I(\tau) \le C_I\tau^{-\alpha_I}, \quad \text{for all } \tau \in \mathbb{N}^+, \tag{30}$$

with θ^I as in (25).

Example 4 (ARFIMA process) Let $\overline{d} \in (-\frac{1}{2}, \frac{1}{2})$ and suppose that $\mathbf{Z} = (Z_t)_{t\in\mathbb{Z}_-}$ is an autoregressive fractionally integrated moving average ARFIMA $(0, \overline{d}, 0)$ process (see, for instance, Hosking, 1981; Beran, 1994, for details). The process $\mathbf{Z}$ admits an infinite moving average (MA(∞)) representation

$$Z_t = \sum_{k=0}^{\infty} \phi_k \varepsilon_{t-k}, \quad t \in \mathbb{Z}_-,$$

with innovations $(\varepsilon_t)_{t\in\mathbb{Z}_-} \sim \mathrm{IID}(0,1)$ and where the coefficients are given by $\phi_k = \frac{\Gamma(k+\overline{d})}{\Gamma(k+1)\Gamma(\overline{d})}$ so that $\Gamma(\overline{d})k^{1-\overline{d}}\phi_k \to 1$, as $k \to \infty$. Using this asymptotic behavior and the independence of the innovations one obtains

$$\theta^z(\tau) \leq 2\mathbb{E}\left[\left|\sum_{k=\tau}^{\infty} \phi_k \varepsilon_{-k}\right|\right] \leq 2\left(\sum_{k=\tau}^{\infty} \phi_k^2\right)^{1/2} \leq 2 \sup_{l\in\mathbb{N}^+}\{l^{1-\overline{d}}\phi_l\}\left(\sum_{k=\tau}^{\infty} k^{2\overline{d}-2}\right)^{1/2}.$$

Comparing the sum to the integral $\int_\tau^\infty x^{2\overline{d}-2}\mathrm{d}x = \frac{1}{1-2\overline{d}}\tau^{2\overline{d}-1}$, it is easy to see that (30) is satisfied with $\alpha_z = \frac{1}{2} - \overline{d} > 0$.

Hypothesis classes of reservoir maps $\mathcal{F}^{RC}$ and reservoir functionals $\mathcal{H}^{RC}$. The next step in order to set up the learning program in the context of reservoir systems is to construct the associated hypothesis classes. These classes need to be chosen beforehand and consist of candidate functionals associated to causal time-invariant reservoir filters of the type discussed in Sections 2.2 and 2.3.

For fixed $N, d \in \mathbb{N}^+$, consider a class $\mathcal{F}^{RC}$ of reservoir maps $F\colon D_N \times D_d \longrightarrow D_N$, $\mathbf{0} \in D_N \subset \mathbb{R}^N$, $D_d \subset \mathbb{R}^d$ that we assume is (a subset of and) separable in the space of bounded continuous functions when equipped with the supremum norm and, additionally, satisfies the following assumptions:

Assumption 4 *There exist $r \in (0,1)$ and $L_R > 0$ such that for each $F \in \mathcal{F}^{RC}$:*

(i) *for any $\mathbf{z} \in D_d$, $F(\cdot, \mathbf{z})$ is an r-contraction,*

(ii) *for any $\mathbf{x} \in D_N$, $F(\mathbf{x}, \cdot)$ is L_R-Lipschitz.*

Assumption 5 *For any $F \in \mathcal{F}^{RC}$ the (first equation in the) system (3) has the echo state property. If H^F is the functional associated to it, we assume that H^F is measurable with respect to the Borel σ-algebra associated to the product topology on its domain.*

Notice that if the state space D_N is a closed ball, then Assumption 4 implies Assumption 5 by Proposition 1. This implication holds for any reservoir system with bounded reservoir maps, an example of which are the elements of the echo state networks family with bounded activation functions σ in Section 2.3.

Assumption 6 *There exists $M_{\mathcal{F}} > 0$ such that*

$$\|H^F(\mathbf{z})\|_2 \leq M_{\mathcal{F}}, \quad \textit{for all } \mathbf{z} \in (D_d)^{\mathbb{Z}_-} \textit{ and for each } F \in \mathcal{F}^{RC}.$$

This assumption automatically holds for many families of reservoir systems. We carefully addressed this question in Section 2.3, where we discussed various families and input types for which the reservoir functionals are indeed bounded. For example, Assumption 6 is satisfied by construction in the case of bounded inputs for all the families in Section 2.3. In the presence of generic unbounded inputs, Assumption 6 obviously holds for echo state networks (ESN) with bounded activation function. In addition, the condition in Assumption 6 appears in many applications. For instance, in the recent paper by Verzelli et al. (2019) it is shown that using a so-called self-normalizing activation function allows one to achieve high performances in standard benchmark tasks. It is not difficult to see that self-normalizing functions yield $\|H^F(\mathbf{z})\|_2 \leq 1$.

Our assumptions also guarantee that various suprema over the classes $\mathcal{H}^{RC}$ and $\mathcal{F}^{RC}$ that will appear in the sequel are measurable random variables. There are very general conditions that guarantee such a fact holds (see Dudley, 2014, Corollary 5.25) but here we simply assume that H_F for all $F \in \mathcal{F}^{RC}$ is bounded (see Assumption 6) and that $\mathcal{F}^{RC}$ is separable in the space of bounded continuous functions when equipped with the supremum norm. This condition together with the continuity assumptions imposed below on the loss function allows us to conclude the measurability of the suprema over $\mathcal{H}^{RC}$ and $\mathcal{F}^{RC}$ (see Lemma 17 in Appendix 5.1 for the details).

Once we have spelled out Assumptions 4-6 that define the class $\mathcal{F}^{RC}$, we proceed to construct the corresponding hypothesis class of reservoir functionals $\mathcal{H}^{RC}$. Since in most of the cases considered in the literature the readouts h in (3) are either polynomial (as in the case of reservoir systems with linear reservoir maps and polynomial readouts) or linear (as in the case of reservoir systems with linear reservoir maps and linear readouts, ESNs, and SAS in Section 2.3), we shall treat the case of generic Lipschitz readouts and the linear case separately:

(i) Reservoir functionals hypothesis class $\mathcal{H}^{RC}$ with Lipschitz readouts. We consider a set $\mathcal{F}^O$ of readout maps $h\colon D_N \to \mathbb{R}^m$ that are Lipschitz-continuous with Lipschitz constant $L_h > 0$. We assume that for all the members of the class it holds that $L_h \leq \overline{L_h}$ and $\|h(\mathbf{0})\|_2 \leq L_{h,0}$, for some fixed $\overline{L_h}, L_{h,0} > 0$, and that the class contains the zero function and is separable in the space of bounded continuous functions when equipped with the supremum norm. In this situation, we define the hypothesis class $\mathcal{H}^{RC}$ of reservoir functionals as

$$\mathcal{H}^{RC} := \{H\colon (D_d)^{\mathbb{Z}_-} \to \mathbb{R}^m \mid H(\mathbf{z}) = h(H^F(\mathbf{z})), h \in \mathcal{F}^O, F \in \mathcal{F}^{RC}\}. \tag{31}$$

(ii) Reservoir functionals hypothesis class $\mathcal{H}^{RC}$ with linear readouts. Most of the examples of reservoir systems which we discussed in Section 2.3 are constructed using linear readout maps, which are known to be easier to train and popular in many practical applications. We hence treat this case separately. Let now the readouts h be given by maps of the type $h(\mathbf{x}) = W\mathbf{x} + \boldsymbol{a}$, $\mathbf{x} \in D_N$, with $W \in \mathbb{M}_{m,N}$ and $\boldsymbol{a} \in \mathbb{R}^m$. We assume that for all the members of the class it holds that $|||W|||_2 \leq \overline{L_h}$ and $\|h(\mathbf{0})\|_2 = \|\boldsymbol{a}\|_2 \leq L_{h,0}$, for some fixed $\overline{L_h}, L_{h,0} > 0$. In this case, such a class of readouts is automatically separable in the space of bounded continuous functions when equipped with the supremum norm. In this situation we hence define the hypothesis

class $\mathcal{H}^{RC}$ of reservoir functionals as

$$\mathcal{H}^{RC} := \{H\colon (D_d)^{\mathbb{Z}_-} \to \mathbb{R}^m \mid H(\mathbf{z}) = WH^F(\mathbf{z}) + \boldsymbol{a}, W \in \mathbb{M}_{m,N}, \boldsymbol{a} \in \mathbb{R}^m, \\ |||W|||_2 \leq \overline{L_h}, \|\boldsymbol{a}\|_2 \leq L_{h,0}, F \in \mathcal{F}^{RC}\}. \tag{32}$$

Loss function. The choice of loss function is often key to the success in quantifying risk bounds for learning models. In this paper we work with distance-based loss functions of the form

$$L(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} f_i(x_i - y_i), \tag{33}$$

for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, where for each $i \in \{1, \ldots, m\}$, the so-called representing functions $f_i\colon \mathbb{R} \longrightarrow \mathbb{R}$ are all Lipschitz-continuous with the same Lipschitz constant $L_L/\sqrt{m}$ and satisfy $f_i(0) = 0$. The assumption of Lipschitz-continuity on the loss L in the case in which its codomain is restricted to $\mathbb{R}^+$ guarantees that it is also a Nemitski loss of order $p = 1$ (see Christmann and Steinwart, 2008, for detailed discussion of Nemitski losses and their associated risks). Notice that our assumptions imply in particular that

$$\mathbb{E}[|L(\mathbf{0}, \mathbf{Y}_0)|] < \infty \tag{34}$$

and

$$|L(\mathbf{x}, \mathbf{y}) - L(\overline{\mathbf{x}}, \overline{\mathbf{y}})| \leq L_L(\|\mathbf{x} - \overline{\mathbf{x}}\|_2 + \|\mathbf{y} - \overline{\mathbf{y}}\|_2), \;\; \mathbf{x}, \overline{\mathbf{x}}, \mathbf{y}, \overline{\mathbf{y}} \in \mathbb{R}^m. \tag{35}$$

Additionally, we notice that since we restrict to reservoir systems satisfying the echo state property and the hypothesis class $\mathcal{H}^{RC}$ contains their associated reservoir functionals, for $H = h \circ H^F$ the idealized empirical risk (17) can be written as

$$\widehat{R}_n^{\infty}(H) = \frac{1}{n}\sum_{i=0}^{n-1} L(h(H^F(\mathbf{Z}_{-i}^{-\infty})), \mathbf{Y}_{-i}) = \frac{1}{n}\sum_{i=0}^{n-1} L(U_h^F(\mathbf{Z}_0^{-\infty})_{-i}, \mathbf{Y}_{-i}) = \frac{1}{n}\sum_{i=0}^{n-1} L(h(\mathbf{X}_{-i}), \mathbf{Y}_{-i}),$$

where $\mathbf{X}$ is the solution of the reservoir system

$$\mathbf{X}_t = F(\mathbf{X}_{t-1}, \mathbf{Z}_t), \quad t \in \mathbb{Z}_-.$$

Risk consistency and risk bounds of reservoir systems. As discussed in Section 3.1 we are interested in generalization error bounds or, in particular, in deriving uniform bounds for $\Delta_n = \sup_{H \in \mathcal{H}^{RC}}\{R(H) - \widehat{R}_n(H)\}$. In order to proceed, we first decompose Δ_n and write

$$\begin{aligned}
\Delta_n = \sup_{H \in \mathcal{H}^{RC}} \{R(H) - \widehat{R}_n(H)\} &\leq \sup_{H \in \mathcal{H}^{RC}} \{R(H) - \widehat{R}_n(H) - \widehat{R}_n^{\infty}(H) + \widehat{R}_n^{\infty}(H)\} \\
&\leq \sup_{H \in \mathcal{H}^{RC}} \left\{\widehat{R}_n^{\infty}(H) - \widehat{R}_n(H)\right\} + \sup_{H \in \mathcal{H}^{RC}} \left\{R(H) - \widehat{R}_n^{\infty}(H)\right\} \\
&\leq \sup_{H \in \mathcal{H}^{RC}} \left|\widehat{R}_n(H) - \widehat{R}_n^{\infty}(H)\right| + \sup_{H \in \mathcal{H}^{RC}} \left|R(H) - \widehat{R}_n^{\infty}(H)\right|.
\end{aligned} \tag{36}$$

This means that one can find upper bounds for both Δ_n and $\overline{\Delta}_n = \sup_{H \in \mathcal{H}^{RC}} |\widehat{R}_n(H) - R(H)|$ by controlling the two summands in the right hand side of the last inequality.

Coming back to the example of the ERM procedure that we discussed in Section 3.1, the previous expression can also be used to deduce the weak (essentially universal) risk-consistency of the ERM for the class $\mathcal{H}^{RC}$. More specifically, in line with classical results due to Vapnik (1991), from the inequalities (22) it follows that in order to establish the weak (essentially universal) risk-consistency of ERM for reservoir functionals one simply needs to show that for any $\epsilon, \delta > 0$ there exists $n_0 \in \mathbb{N}^+$ such that for all $n \geq n_0$ it holds that

$$\mathbb{P}\left(\overline{\Delta}_n > \epsilon\right) = \mathbb{P}\left(\sup_{H \in \mathcal{H}^{RC}} \left|R(H) - \widehat{R}_n(H)\right| > \epsilon\right) \leq \delta.$$

Whenever the inputs are i.i.d. (which is a particular case of our setup), this definition is essentially equivalent to saying that $\mathcal{H}^{RC}$ is a (weak) uniform Glivenko-Cantelli class (see for example Mukherjee et al., 2006). From expression (36) it also follows then that in order to establish the (weak) risk-consistency, it suffices to show the two-sided uniform convergence over the class $\mathcal{H}^{RC}$ of reservoir functionals, first, of the truncated versions of empirical risk to their idealized counterparts and, second, of these idealized versions of the empirical risk to the generalization error (or statistical risk). More explicitly, we shall separately show that for any $\epsilon_1, \delta_1 > 0$ and $\epsilon_2, \delta_2 > 0$ there exist $n_1 \in \mathbb{N}^+$ and $n_2 \in \mathbb{N}^+$ such that for all $n \geq n_1$ and $n \geq n_2$, respectively, it holds that

$$\mathbb{P}\left(\sup_{H \in \mathcal{H}^{RC}} \left|\widehat{R}_n(H) - \widehat{R}_n^{\infty}(H)\right| > \epsilon_1\right) \leq \delta_1 \tag{37}$$

and

$$\mathbb{P}\left(\sup_{H \in \mathcal{H}^{RC}} \left|R(H) - \widehat{R}_n^{\infty}(H)\right| > \epsilon_2\right) \leq \delta_2. \tag{38}$$

One needs to start by showing that the suprema of both these differences over the class $\mathcal{H}^{RC}$ are indeed random variables. This fact has been proved in Lemma 17 in the Appendix 5.1. Next, we need to show that the difference between the idealized and the truncated empirical risks can be made as small as one wants by choosing an appropriate length $n \in \mathbb{N}^+$ of the training sample. This fact is contained in the following result.

Proposition 5 *Consider the hypothesis class $\mathcal{H}^{RC}$ of reservoir functionals defined in* (31)*. Define*

$$C_0 := \frac{2rL_L\overline{L_h}M_{\mathcal{F}}}{1-r}. \tag{39}$$

Then, for any $n \in \mathbb{N}^+$

$$\sup_{H \in \mathcal{H}^{RC}} \left|\widehat{R}_n(H) - \widehat{R}_n^{\infty}(H)\right| \leq \frac{C_0(1-r^n)}{n}$$

holds $\mathbb{P}$-a.s.

This proposition implies that (37) indeed holds. In order to complete the uniform convergence argument, we also need to show that (38) holds. Even though in order to prove the (essentially universal) risk-consistency of the ERM procedure for reservoir systems it

is sufficient to show that the upper bounds of $\overline{\Delta}_n$ (and Δ_n) can be made as small as one wants with $n \to \infty$, for hyper-parameter selection in practical applications the availability of non-asymptotic bounds is also of much importance. We see in the following section that depending on the particular weak dependence assumption imposed (Assumptions 1-3) we will be able to use more or less strong concentration inequalities that yield finite-sample size bounds with different rates of convergence.

4. Main Results

In this section we provide high-probability risk bounds for reservoir computing systems. The main ingredients of the probability bounds are the expected values of $\Gamma_n := \sup_{H\in\mathcal{H}^{RC}}\{R(H)-\widehat{R}_n^{\infty}(H)\}$ and of $\overline{\Gamma}_n := \sup_{H\in\mathcal{H}^{RC}}|R(H)-\widehat{R}_n^{\infty}(H)|$, which are the maximum difference of the idealized training and the generalization errors over the class $\mathcal{H}^{RC}$ and the maximum of the absolute value of this difference, respectively. Since the random variables in the training sample of the input and the output discrete-time processes are not independent and identically distributed, bounding the expected values of Γ_n and $\overline{\Gamma}_n$ is a challenging task. In the first subsection we show that this problem may be circumvented using the following idea: one may compute the empirical risk by partitioning the training sample into blocks of appropriate length and then exploiting the weak dependence of the input and output stochastic processes spelled out in Assumptions 1-3. We first make use of this "block-partitioning" idea in order to derive the bounds for the expected values of the random variables Γ_n and $\overline{\Gamma}_n$ in the setting of each of those three assumptions. These bounds are expressed in terms of the so-called Rademacher complexities of the reservoir hypothesis classes and the weak dependence coefficients of the input and the target stochastic processes. We provide details concerning the complexity bounds for particular families in the second subsection. In the third subsection we then use the fact that the random fluctuations of Γ_n and $\overline{\Gamma}_n$ around their expected values can be controlled using concentration inequalities, which, as we show further, can be done either with the help of the Markov inequality under the weaker Assumptions 2-3, or using stronger exponential concentration inequalities (Propositions 19, 20) under the stronger Assumption 1. This approach yields explicit expressions for non-asymptotic high-probability bounds for $\overline{\Delta}_n$ and hence for Δ_n, which we spell out in the third subsection. Finally, showing that these upper bounds can be made as small as one wants as $n \to \infty$ proves the desired (weak and essentially universal) risk-consistency of the ERM-selected reservoir systems used as learning models. All the proofs of the main results given in this section are provided in the appendices.

4.1 Bounding the expected value

The main ingredient that needs to be introduced in order to bound the expected value of both Γ_n and $\overline{\Gamma}_n$ is a complexity measure for the hypothesis classes of reservoir functionals $\mathcal{H}^{RC}$. Many complexity measures have been discussed in the literature in recent years (see for example Vapnik and Chervonenkis, 1968; Ledoux and Talagrand, 1991; Bartlett and Mendelson, 2003; Ben-David and Shalev-Shwartz, 2014; Rakhlin et al., 2010). In this paper we use the so-called ***(multivariate) Rademacher-type complexity*** associated to a given $\mathcal{H}^{RC}$, which we denote as $\mathcal{R}_k(\mathcal{H}^{RC})$. More explicitly, let $k \in \mathbb{N}^+$ and consider $\varepsilon_0, \ldots, \varepsilon_{k-1}$

independent and identically distributed Rademacher random variables and let $\widetilde{\mathbf{Z}}^{(j)}$, $j = 0, \ldots, k-1$, denote independent copies of $\mathbf{Z}$ (***ghost processes***), which are also independent of $\varepsilon_0, \ldots, \varepsilon_{k-1}$. The Rademacher-type complexity $\mathcal{R}_k(\mathcal{H}^{RC})$ over k ghost processes is defined as

$$\mathcal{R}_k(\mathcal{H}^{RC}) = \frac{1}{k}\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left\|\sum_{j=0}^{k-1}\varepsilon_j H(\widetilde{\mathbf{Z}}^{(j)})\right\|_2\right]. \tag{40}$$

Note that $\mathcal{R}_k(\mathcal{H}^{RC})$ is not an empirical Rademacher complexity and that the expectation is taken with respect to all randomness. In this paper we do not use the standard approach consisting in bounding the theoretical Rademacher complexity using its empirical analogue (conditional on $(\widetilde{\mathbf{Z}}^{(j)})_{j\in\{0,\ldots,k-1\}}$), since in the context of reservoir systems the ghost processes $\widetilde{\mathbf{Z}}^{(j)}$ have no empirical interpretation due to the fact that it is usually only a single trajectory and not i.i.d. samples of input data which are available to the learner.

The following results provide upper bounds for the expected and the expected absolute value of the largest deviation of the statistical risk from its idealized empirical analogue within the hypothesis class of reservoir functionals $\mathcal{H}^{RC}$. The two upper bounds (41) and (42) in the next proposition share the same first three terms, up to a factor 2 due to the absolute value. The first term is related to the weak dependence coefficients of the input and target signals, θ^z and θ^y, respectively. The second term involves the Rademacher-type complexity (40) of the hypothesis class of reservoir functionals $\mathcal{H}^{RC}$. Finally, the third term is always of order $\frac{\tau}{n}$, where τ is the block length, which needs to be carefully chosen depending on the rates of decay of θ^z and θ^y, as we show later in Corollary 8. The upper bound for the expected absolute value (42) contains an additional term of order $\frac{\sqrt{\tau}}{\sqrt{n}}$.

Proposition 6 *Let $\mathcal{H}^{RC}$ be the hypothesis class of reservoir functionals associated to the reservoir maps in the class $\mathcal{F}^{RC}$ as given in* (31) *or* (32)*. Let both the input process $\mathbf{Z}$ and the target process $\mathbf{Y}$ have a causal Bernoulli shift structure as in* (23) *and take values in $(D_d)^{\mathbb{Z}_-}$ and $(\mathbb{R}^m)^{\mathbb{Z}_-}$, respectively. Then, there exist $B > 0$, $M > 0$, and $\{a_\tau\}_{\tau\in\mathbb{N}^+}$ with $a_\tau \in (0,\infty)$, such that for any $\tau, n \in \mathbb{N}^+$ with $\tau < n$ it holds that*

$$\mathbb{E}\left[\Gamma_n\right] = \mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left\{R(H) - \widehat{R}_n^\infty(H)\right\}\right] \leq \frac{k\tau}{n}a_\tau + \frac{Bk\tau}{n}\mathcal{R}_k(\mathcal{H}^{RC}) + \frac{2M(n-k\tau)}{n}, \tag{41}$$

where $k = \lfloor n/\tau \rfloor$ and

$$\begin{aligned}\mathbb{E}\left[\overline{\Gamma}_n\right] = \mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left|R(H) - \widehat{R}_n^\infty(H)\right|\right] &\leq \frac{k\tau}{n}a_\tau + \frac{2Bk\tau}{n}\mathcal{R}_k(\mathcal{H}^{RC}) + \frac{2M(n-k\tau)}{n} \\ &\quad + \frac{4\tau\sqrt{k}}{n}L_L\mathbb{E}\left[\|\mathbf{Y}_0\|_2^2\right]^{1/2}. \end{aligned}\tag{42}$$

In these expressions, the Rademacher complexity $\mathcal{R}_k(\mathcal{H}^{RC})$ of the hypothesis class $\mathcal{H}^{RC}$ of reservoir functionals is defined as in (40)*, the constants and the sequence $\{a_\tau\}_{\tau\in\mathbb{N}^+}$ can be explicitly expressed as*

$$B = 2\sqrt{m}L_L, \tag{43}$$

$$M = L_L\overline{L_h}M_{\mathcal{F}} + \mathbb{E}[|L(\mathbf{0}, \mathbf{Y}_0)|] + L_{h,0}L_L, \tag{44}$$

$$a_\tau = L_L(2r^\tau M_{\mathcal{F}}\overline{L_h} + \theta^y(\tau) + L_R\overline{L_h}\sum_{l=0}^{\tau-1} r^l\theta^z(\tau-l)), \tag{45}$$

where for $I = y, z$ the weak dependence coefficients θ^I for $\tau \in \mathbb{N}^+$ are defined as in (25)*, namely,*

$$\theta^I(\tau) = \mathbb{E}[\|G^I(\ldots, \boldsymbol{\xi}^I_{-1}, \boldsymbol{\xi}^I_0) - G^I(\ldots, \overline{\boldsymbol{\xi}}^I_{-\tau-1}, \overline{\boldsymbol{\xi}}^I_{-\tau}, \boldsymbol{\xi}^I_{-\tau+1}, \ldots, \boldsymbol{\xi}^I_0)\|_2], \tag{46}$$

with $(\overline{\boldsymbol{\xi}}^I_t)_{t\in\mathbb{Z}_-}$ an independent copy of $(\boldsymbol{\xi}^I_t)_{t\in\mathbb{Z}_-}$.

We now explore the conditions required for the upper bounds in (41) and (42) to be finite and exhibit a certain decay as a function of τ and n. Notice that (up to a factor 2) the right-hand sides of the two inequalities are equal up to the last summand in (42) and hence one needs to impose that $\mathbb{E}\left[\|\mathbf{Y}_0\|_2^2\right] < \infty$. Additionally, in order to better understand the behavior of both bounds as a function of τ and n one needs to study two more ingredients, namely the sequence $\{a_\tau\}_{\tau\in\mathbb{N}^+}$ and the Rademacher complexity $\mathcal{R}_k(\mathcal{H}^{RC})$. The behavior of the sequence $\{a_\tau\}_{\tau\in\mathbb{N}^+}$ is exclusively determined by the properties of the input and target processes, while the Rademacher complexity of $\mathcal{H}^{RC}$ is fully characterized by the type of reservoir and readout maps of the given family of reservoir systems. In the following remark we argue that under either the stronger Assumption 1 or the weaker Assumptions 2-3 the sequence $\{a_\tau\}_{\tau\in\mathbb{N}^+}$ converges to zero.

Remark 7 A condition guaranteeing that the sequence $\{a_\tau\}_{\tau\in\mathbb{N}^+}$ in (45) converges to zero is, for instance, that

$$\sum_{\tau=1}^{\infty}\theta^I(\tau) < \infty, \quad I = y, z. \tag{47}$$

To verify this, observe that this condition also implies that

$$\sum_{\tau=1}^{\infty}\sum_{l=0}^{\tau-1} r^l\theta^z(\tau-l) = \sum_{l=0}^{\infty} r^l \sum_{\tau=l+1}^{\infty}\theta^z(\tau-l) = \frac{1}{1-r}\sum_{\tau=1}^{\infty}\theta^z(\tau) < \infty,$$

where we used that $r \in (0,1)$. This proves that $\sum_{\tau=1}^{\infty} a_\tau < \infty$, which necessarily implies that $\lim_{\tau\to\infty} a_\tau = 0$ as required.

It is easy to verify that under Assumption 1 one has that

$$\begin{aligned}\sum_{\tau=1}^{\infty}\theta^I(\tau) &\leq \sum_{\tau=1}^{\infty} 2L_I\mathbb{E}[\|\boldsymbol{\xi}^I_0\|_2]\sum_{j=\tau}^{\infty} w^I_j = 2L_I\mathbb{E}[\|\boldsymbol{\xi}^I_0\|_2]\sum_{j=1}^{\infty}\sum_{\tau=1}^{j} w^I_j \\ &= 2L_I\mathbb{E}[\|\boldsymbol{\xi}^I_0\|_2]\sum_{j=1}^{\infty} jw^I_j < \infty,\end{aligned}$$

which immediately implies that condition (47) is satisfied. Additionally, notice that under Assumption 2, condition (47) is also automatically satisfied. However, under Assumption 3 condition (47) may not be satisfied, but a straightforward argument (see the proof of part **(iii)** of Corollary 8) shows that $\lim_{\tau\to\infty} a_\tau = 0$.

We just argued that under any of the three assumptions 1-3, the convergence to zero of the sequence $\{a_\tau\}_{\tau\in\mathbb{N}^+}$ is guaranteed, which implies that if we establish the finiteness and a certain decay of the Rademacher complexity term we shall have proved that the upper bounds given in (41) and (42) are finite and tend to 0 as $n\to\infty$. The rate of convergence of these bounds is however affected by the particular dependence assumption adopted. We address this important issue in the following corollary where we assume that the Rademacher complexity is finite and exhibits a certain decay and we prove decay rates for the bounds in (41) and (42) that are valid under the different assumptions 1-3. The boundedness of the Rademacher complexities is studied in detail later on in Section 4.2 for the different hypothesis classes of reservoir systems that we introduced in Section 2.3

Corollary 8 *Assume that there exists $C_{RC}>0$ such that for all $k\in\mathbb{N}^+$ the Rademacher-type complexity $\mathcal{R}_k(\mathcal{H}^{RC})$ of the class $\mathcal{H}^{RC}$ of reservoir functionals satisfies*

$$\mathcal{R}_k(\mathcal{H}^{RC})\leq\frac{C_{RC}}{\sqrt{k}}. \tag{48}$$

Consider the following three cases that correspond to Assumptions 1, 2, and 3, respectively:

(i) *Suppose that Assumption 1 holds and that, additionally, for $I=y,z$ the weighting sequences $w^I:\mathbb{N}\longrightarrow(0,1]$ are such that the associated decay ratios $D_{w^I}:=\sup_{i\in\mathbb{N}}\left\{\frac{w_{i+1}^I}{w_i^I}\right\}<1$. Let $\lambda_{max}:=\max(r,D_{w^y},D_{w^z})$. Then, there exist C_1, C_2, C_3, $C_{3,abs}>0$ such that for all $n\in\mathbb{N}^+$ satisfying $\log(n)<n\log(\lambda_{max}^{-1})$ it holds that*

$$\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left\{R(H)-\widehat{R}_n^\infty(H)\right\}\right]\leq\frac{C_1}{n}+\frac{C_2\log(n)}{n}+\frac{C_3\sqrt{\log(n)}}{\sqrt{n}} \tag{49}$$

and

$$\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left|R(H)-\widehat{R}_n^\infty(H)\right|\right]\leq\frac{C_1}{n}+\frac{C_2\log(n)}{n}+\frac{C_{3,abs}\sqrt{\log(n)}}{\sqrt{n}}. \tag{50}$$

The constants can be explicitly chosen as

$$C_1=\frac{2M_{\mathcal{F}}L_L\overline{L_h}+L_LC_y}{\lambda_{max}},\qquad C_2=\frac{2M}{\log(\lambda_{max}^{-1})}+\frac{L_LL_R\overline{L_h}C_z}{\lambda_{max}\log(\lambda_{max}^{-1})}, \tag{51}$$

$$C_3=\frac{2\sqrt{m}L_LC_{RC}}{\sqrt{\log(\lambda_{max}^{-1})}},\qquad C_{3,abs}=2C_3+\frac{4L_L\mathbb{E}\left[\|\mathbf{Y}_0\|_2^2\right]^{1/2}}{\sqrt{\log(\lambda_{max}^{-1})}}, \tag{52}$$

where M is as in (44) and $C_I=\dfrac{2L_I\mathbb{E}[\|\boldsymbol{\xi}_0^I\|_2]}{1-D_{w^I}}$ for $I=y,z$.

(ii) *Suppose that Assumption 2 holds and let $\lambda_{max}:=\max(r,\lambda_y,\lambda_z)$ with λ_y,λ_z as in (26). Then there exist C_1, C_2, C_3, $C_{3,abs}>0$ such that for all $n\in\mathbb{N}^+$ satisfying $\log(n)<n\log(\lambda_{max}^{-1})$ the bounds in (49) and (50) hold. The constants can be explicitly chosen as in (51)-(52) with C_I as in (26).*

(iii) *Suppose that Assumption 3 holds and denote* $\alpha := \min(\alpha_y, \alpha_z)$. *Then there exist* $C_1, C_2, C_{1,abs} > 0$ *such that for all* $n \in \mathbb{N}^+$ *it holds that*

$$\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left\{R(H) - \widehat{R}_n^{\infty}(H)\right\}\right] \leq C_1 n^{-\frac{1}{2+\alpha^{-1}}} + C_2 n^{-\frac{2}{2+\alpha^{-1}}}$$

and

$$\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left|R(H) - \widehat{R}_n^{\infty}(H)\right|\right] \leq C_{1,abs} n^{-\frac{1}{2+\alpha^{-1}}} + C_2 n^{-\frac{2}{2+\alpha^{-1}}}.$$

The constants can be explicitly chosen as

$$C_1 = L_L(2M_{\mathcal{F}}\overline{L_h}r^{-\gamma_\alpha} + L_R\overline{L_h}C_zC_\alpha + C_y) + BC_{RC}, \quad C_2 = 2M, \tag{53}$$

$$C_{1,abs} = C_1 + 4L_L\mathbb{E}\left[\|\mathbf{Y}_0\|_2^2\right]^{1/2} + BC_{RC}, \tag{54}$$

with M, B *as in* (43)-(44) *and*

$$\gamma_\alpha = \max_{\tau\in\mathbb{N}^+}\left\{\frac{\log(\tau)\alpha_z}{\log(r^{-1})} - \frac{\tau}{4}\right\}, \tag{55}$$

$$C_\alpha = \max(2^{\alpha_z}, r^{-\gamma_\alpha})(1-\sqrt{r})^{-1}, \tag{56}$$

and C_I, α_I, *for* $I = y, z$ *as in* (30).

4.2 Rademacher complexity of reservoir systems

In this section we show that for the most important hypothesis classes of reservoir systems, the Rademacher complexity tends to 0 as $k \to \infty$ at the rates required in (48) of Corollary 8. More specifically, in the next propositions we will provide upper bounds for the Rademacher complexities of the most popular reservoir families that we spelled out in Section 2.3. For these propositions we assume that the corresponding parameter set Θ is separable in the respective (Euclidean) space. This is required in order to ensure that the supremum of random variables appearing in the definition of the Rademacher complexity (40) is again a random variable. For instance, if Θ is an open set, then it is separable.

Reservoir systems with linear reservoir map (LRC) and linear readout

We now provide a bound for the Rademacher complexity of classes of reservoir functionals associated to linear reservoir maps and readouts. We recall that in this case we always work with uniformly bounded inputs $\mathbf{Z}$ (see Section 2.3) by some constant $M > 0$, that is, $D_d = \overline{B_M}$ and so the random variable $\mathbf{Z}$ takes values in the set

$$K_M := \left\{\mathbf{z} \in (\mathbb{R}^d)^{\mathbb{Z}_-} \,|\, \|\mathbf{z}_t\|_2 \leq M \text{ for all } t \in \mathbb{Z}_-\right\}. \tag{57}$$

Proposition 9 *Let* $N, d \in \mathbb{N}^+$ *and let* $\Theta \subset \mathbb{M}_N \times \mathbb{M}_{N,d} \times \mathbb{R}^N$. *Define the classes of linear reservoir maps as*

$$\mathcal{F}^{RC} := \{F^{A,C,\boldsymbol{\zeta}} \,|\, (A, C, \boldsymbol{\zeta}) \in \Theta\}$$

and let $\mathcal{H}^{RC}$ *be a class of reservoir functionals of the type defined in* (32)*, associated to reservoir systems with linear reservoir maps and readouts. Additionally, define*

$$\lambda^{A}_{max} := \sup_{(A,C,\boldsymbol{\zeta})\in\Theta} |||A|||_2, \tag{58}$$

$$\lambda^{C}_{max} := \sup_{(A,C,\boldsymbol{\zeta})\in\Theta} |||C|||_2, \tag{59}$$

$$\lambda^{\boldsymbol{\zeta}}_{max} := \sup_{(A,C,\boldsymbol{\zeta})\in\Theta} \|\boldsymbol{\zeta}\|_2. \tag{60}$$

If for the class $\mathcal{F}^{RC}$ *it holds that*

$$0 < \lambda^{A}_{max} < 1, \quad \lambda^{C}_{max} < \infty, \quad \lambda^{\boldsymbol{\zeta}}_{max} < \infty, \tag{61}$$

then Assumptions 4-6 are satisfied and the Rademacher complexity of the associated class of reservoir functionals satisfies

$$\mathcal{R}_k(\mathcal{H}^{RC}) \leq \frac{C_{LRC}}{\sqrt{k}}, \tag{62}$$

for any $k \in \mathbb{N}^+$*, where*

$$C_{LRC} = \frac{\overline{L_h}}{1-\lambda^{A}_{max}} \left(\lambda^{C}_{max} \mathbb{E}\left[\|\mathbf{Z}_0\|_2^2 \right]^{1/2} + \lambda^{\boldsymbol{\zeta}}_{max} \right) + L_{h,0}, \tag{63}$$

and with $\mathbf{Z}$ *the input process.*

Remark 10 Due to the uniform boundedness of the inputs, the constant $\mathbb{E}\left[\|\mathbf{Z}_0\|_2^2 \right]^{1/2}$ in (63) is bounded by the value M that defines the set K_M in which the inputs take values. Nevertheless, $\mathbb{E}\left[\|\mathbf{Z}_0\|_2^2 \right]^{1/2}$ can obviously be much smaller than M.

Echo State Networks (ESN)

The following proposition provides an estimate for the Rademacher complexity of hypothesis classes constructed using echo state networks. We recall that in this case we either work with arbitrary inputs and a bounded activation function σ or with a possibly unbounded activation function σ and uniformly bounded inputs $\mathbf{Z}$ (see Section 2.3) by some constant $M > 0$, that is, $D_d = \overline{B_M}$.

Proposition 11 *Let* $N, d \in \mathbb{N}^+$*, let* $\Theta \subset \mathbb{M}_N \times \mathbb{M}_{N,d} \times \mathbb{R}^N$ *be a subset, and let* $\mathcal{F}^{RC}$ *be a family of echo state reservoir systems defined as*

$$\mathcal{F}^{RC} := \{ F^{\sigma,A,C,\boldsymbol{\zeta}} \,|\, (A,C,\boldsymbol{\zeta}) \in \Theta \}$$

and let $\mathcal{H}^{RC}$ *be a class of reservoir functionals of the type defined in* (32)*, associated to reservoir systems with linear reservoir maps and readouts. Suppose that the class* $\mathcal{F}^{RC}$ *is such that for any* $F^{\sigma,A,C,\boldsymbol{\zeta}} \in \mathcal{F}^{RC}$ *one necessarily has that* $-F^{\sigma,A,C,\boldsymbol{\zeta}}(-\cdot,\cdot) \in \mathcal{F}^{RC}$*.*[2] *Define*

$$\lambda^{A}_{max} := L_\sigma \sum_{l=1}^{N} \sup_{(A,C,\boldsymbol{\zeta})\in\Theta} \|A_{l,\cdot}\|_\infty,$$

2. This is satisfied for example if σ is odd and $(A,C,\boldsymbol{\zeta}) \in \Theta \Leftrightarrow (A,-C,-\boldsymbol{\zeta}) \in \Theta$.

$$\lambda^C_{max} := L_\sigma \sum_{l=1}^{N} \sup_{(A,C,\boldsymbol{\zeta})\in\Theta} \|C_{l,\cdot}\|_2,$$
$$\lambda^{\boldsymbol{\zeta}}_{max} := L_\sigma \sum_{l=1}^{N} \sup_{(A,C,\boldsymbol{\zeta})\in\Theta} |\zeta_l|.$$

If for class $\mathcal{F}^{RC}$ it holds that

$$0 < \lambda^A_{max} < 1, \quad \lambda^C_{max} < \infty, \quad \lambda^{\boldsymbol{\zeta}}_{max} < \infty,$$

then Assumptions 4-6 are satisfied and the Rademacher complexity of the associated class of reservoir functionals satisfies

$$\mathcal{R}_k(\mathcal{H}^{RC}) \leq \frac{C_{ESN}}{\sqrt{k}}, \quad \text{for any } k \in \mathbb{N}^+, \tag{64}$$

where

$$C_{ESN} = \frac{\overline{L_h}}{1-\lambda^A_{max}} \left(\lambda^C_{max} \mathbb{E}\left[\|\mathbf{Z}_0\|_2^2 \right]^{1/2} + \lambda^{\boldsymbol{\zeta}}_{max} \right) + L_{h,0}, \tag{65}$$

and with $\mathbf{Z}$ the input process.

Remark 12 Notice that by (62)-(63) and by (64)-(65) the Rademacher complexities of the hypothesis classes formed by reservoir systems with linear reservoir maps and linear readouts or by echo state networks are finite whenever the second moment of the input process is finite, which is not directly implied by any of the assumptions 1-3 and hence needs to be separately assumed.

State Affine Systems (SAS)

In the following proposition we provide an estimate for the Rademacher complexity of hypothesis classes constructed using state affine systems. In this case we also work with uniformly bounded inputs (see Section 2.3) in a set of the type K_M as in (57) with $M = 1$.

Proposition 13 *Let $\Theta \subset \mathbb{M}_{N,N}[\mathbf{z}] \times \mathbb{M}_{N,1}[\mathbf{z}]$, and define the class of SAS reservoir maps as*

$$\mathcal{F}^{RC} := \{F^{p,q} \,|\, (p,q) \in \Theta\}.$$

Assume that there is a finite set $I_{max} \subset \mathbb{N}^d$ such that for any $P(\mathbf{z}) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^d} A_{\boldsymbol{\alpha}} \mathbf{z}^{\boldsymbol{\alpha}}$ with $P = p$ or $P = q$, $(p,q) \in \Theta$ one has $A_{\boldsymbol{\alpha}} = 0$ for $\boldsymbol{\alpha} \notin I_{max}$ and define $|I_{max}| := \mathrm{card}(I_{max})$,

$$\begin{aligned} \lambda^{SAS} &:= \sup_{(p,q)\in\Theta} |||p|||, \\ c^{SAS} &:= \sup_{(p,q)\in\Theta} |||q|||, \end{aligned} \tag{66}$$

where the norm $|||\cdot|||$ was introduced in (9). Let $\mathcal{H}^{RC}$ be the hypothesis class of reservoir systems with linear readouts associated to $\mathcal{F}^{RC}$ as in (32). Then, if for the class $\mathcal{F}^{RC}$ it

holds that $\lambda^{SAS} < 1/|I_{max}|$ *and* $c^{SAS} < \infty$*, then Assumptions 4-6 are satisfied and for any* $k \in \mathbb{N}^+$ *it holds that*

$$\mathcal{R}_k(\mathcal{H}^{RC}) \leq \frac{C_{SAS}}{\sqrt{k}}$$

with

$$C_{SAS} = \overline{L_h} \frac{c^{SAS}|I_{max}|}{1 - |I_{max}|\lambda^{SAS}} + L_{h,0}. \tag{67}$$

4.3 High-probability risk bounds for reservoir systems

We now use the previous results and the three assumptions 1-3 in conjunction with different concentration inequalities to produce three families of high-probability bounds for $\overline{\Delta}_n := \sup_{H\in\mathcal{H}} |\widehat{R}_n(H) - R(H)|$ of different strength for reservoir systems, which prove in passing the (weak) universal risk-consistency of ERM for reservoir functionals. High-probability finite-sample generalization RC bounds of this type were not available in the literature previously.

Theorem 14 *Let* $\mathcal{H}^{RC}$ *be a hypothesis class of reservoir functionals of the type specified in* (32) *associated to a class* $\mathcal{F}^{RC}$ *of reservoir maps that satisfies Assumptions 4-6 and assume that the Rademacher complexity of* $\mathcal{H}^{RC}$ *satisfies* (48)*. Suppose that both the input* $\mathbf{Z}$ *and the target* $\mathbf{Y}$ *processes have a causal Bernoulli shift structure as in* (23) *and that they take values in* $(\mathbb{R}^d)^{\mathbb{Z}_-}$ *and* $(\mathbb{R}^m)^{\mathbb{Z}_-}$*,* $d, m \in \mathbb{N}^+$*, respectively.*

(i) *Suppose that Assumption 1 is satisfied and that, additionally, for* $I = y, z$ *the strictly decreasing weighting sequences* $w^I : \mathbb{N} \longrightarrow (0, 1]$ *are such that the associated decay ratios* $D_{w^I} := \sup_{i\in\mathbb{N}} \frac{w^I_{i+1}}{w^I_i} < 1$*. Let* $\lambda_{max} := \max(r, D_{w^y}, D_{w^z})$*.*

(a) *Assume that the innovations are bounded, that is, there exists* $\overline{M} > 0$ *such that* $\|\boldsymbol{\xi}_t\|_2 \leq \overline{M}$ *for all* $t \in \mathbb{Z}_-$*. Then there exist constants* $C_0, C_1, C_2, C_3, C_{bd} > 0$ *such that for all* $n \in \mathbb{N}^+$ *satisfying* $\log(n) < n\log(\lambda_{max}^{-1})$ *and for all* $\delta \in (0, 1)$*, the following bound holds*

$$\mathbb{P}\left(\sup_{H\in\mathcal{H}^{RC}} |\widehat{R}_n(H) - R(H)| \leq \frac{(1-r^n)C_0 + C_1}{n} + \frac{C_2\log(n)}{n} + \frac{C_3\sqrt{\log(n)}}{\sqrt{n}} + \frac{C_{bd}\sqrt{\log(\frac{4}{\delta})}}{\sqrt{2n}} \right) \geq 1 - \delta, \tag{68}$$

where the constant C_0 *is explicitly given in* (39)*,* C_1, C_2 *are given in* (51)*,* C_3 *in* (52)*, and* C_{bd} *in* (93)*.*

(b) *Assume that for* $\Phi(x) = x^p$*,* $p > 1$ *or* $\Phi(x) = \exp(x) - 1$ *the innovations possess* Φ^2*-moments, that is, for any* $u > 0$*,* $\mathbb{E}[\Phi(u\|\boldsymbol{\xi}_0\|_2)^2] < \infty$*. Then there exist constants* $C_0, C_1, C_2, C_3 > 0$ *such that for all* $n \in \mathbb{N}^+$ *satisfying* $\log(n) < n\log(\lambda_{max}^{-1})$ *and for all* $\delta \in (0, 1)$ *it holds that*

$$\mathbb{P}\left(\sup_{H\in\mathcal{H}^{RC}} |\widehat{R}_n(H) - R(H)| \leq \frac{(1-r^n)C_0 + C_1}{n} + \frac{C_2\log(n)}{n} + \frac{C_3\sqrt{\log(n)}}{\sqrt{n}} + B_\Phi(n, \delta) \right) \geq 1 - \delta, \tag{69}$$

where $B_\Phi(n,\delta)$ is given in (112). *The constants are explicitly given: C_0 in* (39), *C_1, C_2 are given in* (51), *and C_3 in* (52).

(ii) *Suppose that Assumption 2 is satisfied and let $\lambda_{max} := \max(r, \lambda_y, \lambda_z)$ with λ_y, λ_z as in* (26). *Then there exist constants C_0, C_1, C_2, $C_{3,abs} > 0$ such that for all $n \in \mathbb{N}^+$ satisfying $\log(n) < n\log(\lambda_{max}^{-1})$ and for all $\delta \in (0,1)$ it holds that*

$$\mathbb{P}\left(\sup_{H\in\mathcal{H}^{RC}} |\widehat{R}_n(H) - R(H)| \leq \frac{(1-r^n)C_0}{n} + \frac{2}{\delta}\left(\frac{C_1}{n} + \frac{C_2\log(n)}{n} + \frac{C_{3,abs}\sqrt{\log(n)}}{\sqrt{n}}\right)\right) \geq 1-\delta. \tag{70}$$

The constants are explicitly given: C_0 in (39), *C_1, C_2 are given in* (51), *and $C_{3,abs}$ in* (52) *with C_I as in* (26).

(iii) *Suppose that Assumption 3 is satisfied. Denote $\alpha = \min(\alpha_y, \alpha_z)$ with α_y, α_z as in* (30). *Then there exist constants C_0, $C_{1,abs}$, $C_2 > 0$ such that for all $n \in \mathbb{N}^+$, $\delta \in (0,1)$,*

$$\mathbb{P}\left(\sup_{H\in\mathcal{H}^{RC}} |\widehat{R}_n(H) - R(H)| \leq \frac{(1-r^n)C_0}{n} + \frac{2}{\delta}\left(C_{1,abs}n^{-\frac{1}{2+\alpha^{-1}}} + C_2 n^{-\frac{2}{2+\alpha^{-1}}}\right)\right) \geq 1-\delta. \tag{71}$$

The constants are explicitly given: C_0 in (39), *$C_{1,abs}$ is given in* (54) *and C_2 in* (53) *together with* (55)-(56).

In order to obtain explicit high-probability risk bounds for particular families of reservoir systems, one can use the bounds that we obtained for the Rademacher complexities of various families in Section 4.2. For example, let $\mathcal{F}^{RC}$ be a family of state affine systems that satisfies the assumptions of Proposition 13; in that case one takes the value C_{RC} appearing in various constants above (for example in C_3 given in (52)) as $C_{RC} = C_{SAS}$ with C_{SAS} given in (67). The same applies to other families: for echo state networks one takes $C_{RC} = C_{ESN}$ with C_{ESN} given in (65). For the family of reservoir systems with linear reservoir map one takes $C_{RC} = C_{LRC}$, with C_{LRC} given in (63).

Remark 15 The result in part **(ii)** requires Assumption 2 which, as we saw in Remark 4, is implied by Assumption 1 with geometric weighting sequences, that is, $w_s^z = \lambda_z^s$ and $w_s^y = \lambda_y^s$, for some $\lambda_z, \lambda_y \in (0,1)$. Therefore, both (68) and (70) provide bounds in this case. We also emphasize that the result in part **(iii)** allows the treatment of long-memory processes as inputs (see, for instance Example 4).

4.4 High-probability risk bounds for randomly generated reservoir systems

We now show that the results in Theorem 14 can be reformulated for echo state networks whose parameters A, C, and $\boldsymbol{\zeta}$ have been randomly generated. This statement is a theoretical justification of the good empirical properties of this standard *modus operandi* in reservoir computing. Even though results of this type could be formulated for all the reservoir families introduced in Section 2.3 and all the different settings considered in Theorem 14, we restrict our study in the next proposition to echo state networks and part **(ii)**.

Proposition 16 (Random reservoirs) *Let $\mathbf{A}, \mathbf{C}, \boldsymbol{\zeta}$ be independent random variables with values in $\mathbb{M}_N$, $\mathbb{M}_{N,d}$, and in $\mathbb{R}^N$, respectively. Consider now echo state networks that have those random values as parameters and whose activation function σ is odd, that is, consider the random class of reservoir maps defined as*

$$\boldsymbol{\mathcal{F}}^{RC} := \{F^{\sigma,\rho_A\mathbf{A},\rho_C\mathbf{C},\rho_{\boldsymbol{\zeta}}\boldsymbol{\zeta}} \,|\, (\rho_A,\rho_C,\rho_{\boldsymbol{\zeta}}) \in (-\frac{a}{\lambda^{\boldsymbol{A}}},\frac{a}{\lambda^{\boldsymbol{A}}}) \times [-c,c] \times [-s,s]\}$$

for some $a \in (0,1)$, $c, s > 0$ and with

$$\lambda^{\mathbf{A}} = L_\sigma \left(\sum_{l=1}^{N} \|\mathbf{A}_{l,\cdot}\|_\infty \right).$$

Suppose also that the input process $\mathbf{Z}$ and the target process $\mathbf{Y}$ are independent of the parameter random variables $\mathbf{A}, \mathbf{C}, \boldsymbol{\zeta}$ and that Assumption 2 is satisfied. Let $\lambda_{max} := \max(r, \lambda_y, \lambda_z)$ with λ_y, λ_z as in (26). Then there exist constants C_0, C_1, C_2, $C_{3,abs} > 0$ such that for all $n \in \mathbb{N}^+$ satisfying $\log(n) < n\log(\lambda_{max}^{-1})$ and for all $\delta \in (0,1)$ it holds that

$$\mathbb{P}\left(\sup_{H \in \boldsymbol{\mathcal{H}}^{RC}} |\widehat{R}_n(H) - R(H)| \leq \frac{(1-r^n)C_0}{n} + \frac{2}{\delta}\left(\frac{C_1}{n} + \frac{C_2 \log(n)}{n} + \frac{C_{3,abs}\sqrt{\log(n)}}{\sqrt{n}} \right) \right) \geq 1 - \delta, \tag{72}$$

where $\boldsymbol{\mathcal{H}}^{RC}$ is the hypothesis class of reservoir functionals associated to $\boldsymbol{\mathcal{F}}^{RC}$ and with linear readouts as in (32). The constants are explicitly given. More specifically, C_0 is given in (39), C_1, C_2 are given in (51), $C_{3,abs}$ in (52) with C_I as in (26). Additionally, the constant C_{RC} appearing in (52) is given by

$$C_{RC} = \frac{\overline{L_h}}{1-a}(\mathbb{E}[\lambda^{\mathbf{C}}]\mathbb{E}[\|\mathbf{Z}_0\|_2^2]^{1/2} + \mathbb{E}[\lambda^{\boldsymbol{\zeta}}]) + L_{h,0}, \tag{73}$$

where

$$\lambda^{\mathbf{C}} = cL_\sigma \left(\sum_{l=1}^{N} \|\mathbf{C}_{l,\cdot}\|_2 \right), \; \lambda^{\boldsymbol{\zeta}} = sL_\sigma \left(\sum_{l=1}^{N} \|\zeta_l\|_2 \right).$$

5. Appendices

These appendices contain preliminary results and the proofs of all the main results of the paper.

5.1 Preliminary results

The following Lemma shows that the supremum appearing for instance in (68) is indeed a random variable. More precisely, $\sup_{H \in \mathcal{H}^{RC}} |\widehat{R}_n(H) - R(H)|$ is a measurable mapping from $(\Omega, \mathcal{A})$ to $\mathbb{R}$ equipped with its Borel sigma-algebra. An analogous argument can be used in the case of the other suprema, for instance the supremum in (40), considered in the paper.

Lemma 17 *Let $\mathcal{H}^{RC}$ be the reservoir hypothesis class introduced in (31) or in (32) and let R and $\widehat{R}_n$ be the statistical and empirical risk introduced in (13) and (16), respectively. Then,*

$$\sup_{H \in \mathcal{H}^{RC}} |\widehat{R}_n(H) - R(H)|$$

is a random variable.

Proof For any $H \in \mathcal{H}^{RC}$ set $\Delta(H) := |\widehat{R}_n(H) - R(H)|$. Then, for any $H, \bar{H} \in \mathcal{H}^{RC}$

$$\begin{aligned}\Delta(H) - \Delta(\bar{H}) &\leq |\Delta(H) - \Delta(\bar{H})| \leq |\widehat{R}_n(H) - R(H) - \widehat{R}_n(\bar{H}) + R(\bar{H})| \\ &\leq |\widehat{R}_n(H) - \widehat{R}_n(\bar{H})| + |R(H) - R(\bar{H})| \\ &\leq \frac{L_L}{n}\sum_{i=0}^{n-1} \|H(\mathbf{Z}_{-i}^{-n+1}) - \bar{H}(\mathbf{Z}_{-i}^{-n+1})\|_2 + L_L \mathbb{E}[\|H(\mathbf{Z}) - \bar{H}(\mathbf{Z})\|_2] \\ &\leq 2L_L \sup_{\mathbf{z}\in(D_d)^{\mathbb{Z}_-}} \|H(\mathbf{z}) - \bar{H}(\mathbf{z})\|_2, \end{aligned} \tag{74}$$

where we used the (reverse) triangle inequality and the Lipschitz property (35) of the loss function. Further, by using the definition (31) of $\mathcal{H}^{RC}$, Assumption 6 on the boundedness of the functionals associated to reservoir maps, and again the triangle inequality, one obtains

$$\sup_{H\in\mathcal{H}^{RC}} \Delta(H) = \sup_{H\in\mathcal{H}^{RC}} |\widehat{R}_n(H) - R(H)| \leq \Delta(0) + 4L_L[\overline{L_h} M_{\mathcal{F}} + L_{h,0}]$$

and so (34) yields that $\sup_{H\in\mathcal{H}^{RC}} |\widehat{R}_n(H) - R(H)|$ is finite, $\mathbb{P}$-a.s.

It remains to prove the measurability. The separability assumption imposed on $\mathcal{F}^{RC}$ guarantees the existence of a countable subset $\{F_j\}_{j\in\mathbb{N}^+} \subset \mathcal{F}^{RC}$ which is dense with respect to the supremum norm. Let also $\{h_k\}_{k\in\mathbb{N}^+} \subset \mathcal{F}^O$ be a countable dense subset of readouts that by hypothesis exists. This can be used to construct a countable dense subset of $\mathcal{H}^{RC}$. Indeed, for any $H \in \mathcal{H}^{RC}$, $H = h(H^F)$, one may choose indices $(j_l, k_l)_{l\in\mathbb{N}^+}$ such that $\|F_{j_l} - F\|_\infty \to 0$ and $\|h_{k_l} - h\|_\infty \to 0$ as $l \to \infty$. Consequently, using the triangle inequality and an argument as in part **(iii)** of Theorem 3.1 in Grigoryeva and Ortega (2018b), one obtains for any $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$ that

$$\begin{aligned}\|H(\mathbf{z}) - h_{k_l}(H^{F_{j_l}}(\mathbf{z}))\|_2 &\leq \|h - h_{k_l}\|_\infty + \overline{L_h}\|H^F(\mathbf{z}) - H^{F_{j_l}}(\mathbf{z})\|_2 \\ &\leq \|h - h_{k_l}\|_\infty + \frac{1}{1-r}\|F_{j_l} - F\|_\infty.\end{aligned}$$

Combining this with (74) and setting $H_l = h_{k_l}(H^{F_{j_l}})$ one obtains that

$$\lim_{l\to\infty} |\Delta(H) - \Delta(H_l)| \leq \lim_{l\to\infty} 2L_L \left(\|h - h_{k_l}\|_\infty + \frac{1}{1-r}\|F_{j_l} - F\|_\infty \right) = 0.$$

In particular this shows that for any $H \in \mathcal{H}^{RC}$, $\Delta(H) \leq \sup_{j,k\in\mathbb{N}^+} \Delta(h_k(H^{F_j}))$. Taking the supremum over $H \in \mathcal{H}^{RC}$ thus shows that

$$\sup_{H\in\mathcal{H}^{RC}} |\widehat{R}_n(H) - R(H)| = \sup_{j,k\in\mathbb{N}^+} \Delta(h_k(H^{F_j}))$$

is measurable, as it is the supremum of a countable collection of random variables. ■

5.2 Proof of Proposition 1

Consider the map

$$\begin{array}{rccc} \mathcal{F}: & (\overline{B_S})^{\mathbb{Z}_-} \times (D_d)^{\mathbb{Z}_-} & \longrightarrow & (\overline{B_S})^{\mathbb{Z}_-} \\ & (\mathbf{x}, \mathbf{z}) & \longmapsto & (\mathcal{F}(\mathbf{x}, \mathbf{z}))_t := F(\mathbf{x}_{t-1}, \mathbf{z}_t),\ t \in \mathbb{Z}_-, \end{array}$$

and endow $(D_d)^{\mathbb{Z}_-}$ and $(\overline{B_S})^{\mathbb{Z}_-}$ with the relative topologies induced by the product topologies in $(\mathbb{R}^d)^{\mathbb{Z}_-}$ and $(\mathbb{R}^N)^{\mathbb{Z}_-}$, respectively. Notice that $\mathcal{F}$ can be written as

$$\mathcal{F} = \prod_{t \in \mathbb{Z}_-} F_t \quad \text{with} \quad F_t := F \circ p_t \circ \left(T_1 \times \mathrm{id}_{(D_d)^{\mathbb{Z}_-}}\right) : (\overline{B_S})^{\mathbb{Z}_-} \times (D_d)^{\mathbb{Z}_-} \longrightarrow \overline{B_S},$$

where p_t yields the canonical projection of any sequence onto its t-th component. It is easy to see that the maps p_t and T_1 are continuous with respect to the product topologies and hence $\mathcal{F}$ is a Cartesian product of continuous functions, which is always continuous in the product topology.

We now recall that, by the compactness of $\overline{B_S}$, we have that $(\overline{B_S})^{\mathbb{Z}_-} \subset \ell_-^{\infty,w}(\mathbb{R}^N)$ and that by Corollary 2.7 in Grigoryeva and Ortega (2018b), the product topology on $(\overline{B_S})^{\mathbb{Z}_-}$ coincides with the norm topology induced by $\ell_-^{\infty,w}(\mathbb{R}^N)$, for any weighting sequence w, that we choose in the sequel satisfying the inequality $rL_w < 1$.

We now show that $\mathcal{F}$ is a contraction in the first entry with constant $rL_w < 1$. Indeed, for any $\mathbf{x}^1, \mathbf{x}^2 \in (\overline{B_S})^{\mathbb{Z}_-}$ and any $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$, we have

$$\begin{aligned} \left\|\mathcal{F}(\mathbf{x}^1, \mathbf{z}) - \mathcal{F}(\mathbf{x}^2, \mathbf{z})\right\|_{\infty,w} &= \sup_{t \in \mathbb{Z}_-} \left\{\left\|F(\mathbf{x}^1_{t-1}, \mathbf{z}_t) - F(\mathbf{x}^2_{t-1}, \mathbf{z}_t)\right\|_2 w_{-t}\right\} \\ &\leq \sup_{t \in \mathbb{Z}_-} \left\{\left\|\mathbf{x}^1_{t-1} - \mathbf{x}^2_{t-1}\right\|_2 r w_{-t}\right\}, \end{aligned}$$

where we used that F is a contraction in the first entry. Now,

$$\sup_{t \in \mathbb{Z}_-} \left\{\left\|\mathbf{x}^1_{t-1} - \mathbf{x}^2_{t-1}\right\|_2 r w_{-t}\right\} = r \sup_{t \in \mathbb{Z}_-} \left\{\left\|\mathbf{x}^1_{t-1} - \mathbf{x}^2_{t-1}\right\|_2 w_{-(t-1)} \frac{w_{-t}}{w_{-(t-1)}}\right\} \leq r L_w \left\|\mathbf{x}^1 - \mathbf{x}^2\right\|_{\infty,w},$$

which shows that $\mathcal{F}$ is a family of contractions with constant $rL_w < 1$ that is continuously parametrized by the elements in $(D_d)^{\mathbb{Z}_-}$. In view of these facts and given that the product topology in $(D_d)^{\mathbb{Z}_-} \subset (\mathbb{R}^d)^{\mathbb{Z}_-}$ is metrizable (see Munkres, 2014, Theorem 20.5) and that $(\overline{B_S})^{\mathbb{Z}_-} \subset (\mathbb{R}^N)^{\mathbb{Z}_-}$ is compact by Tychonoff's Theorem (see Munkres, 2014, Theorem 37.3) in the product topology and hence complete, Theorem 6.4.1 in Sternberg (2010) implies the existence of a unique fixed point of $\mathcal{F}(\cdot, \mathbf{z})$ for each $\mathbf{z} \in (D_d)^{\mathbb{Z}_-}$, which establishes the ESP. Moreover, that result also shows the continuity of the associated filter $U^F : (D_d)^{\mathbb{Z}_-} \longrightarrow ((\overline{B_S})^{\mathbb{Z}_-}, \|\cdot\|_{\infty,w})$. ■

5.3 Proof of Proposition 5

In order to proceed with the proof of this proposition, we first need the following lemma.

Lemma 18 *For any $F \in \mathcal{F}^{RC}$ and any $\mathbf{z}, \overline{\mathbf{z}} \in (D_d)^{\mathbb{Z}_-}$ the following holds for all $i \in \mathbb{N}^+$:*

$$\|H^F(\mathbf{z}) - H^F(\overline{\mathbf{z}})\|_2 \leq 2r^i M_{\mathcal{F}} + L_R \sum_{j=0}^{i-1} r^j \|\mathbf{z}_{-j} - \overline{\mathbf{z}}_{-j}\|_2. \tag{75}$$

Proof of Lemma 18. Let $\mathbf{z}, \overline{\mathbf{z}} \in (D_d)^{\mathbb{Z}_-}$ and denote by $\mathbf{x}$ the solution to (3) and by $\overline{\mathbf{x}}$ the solution to (3) with $\mathbf{z}$ replaced by $\overline{\mathbf{z}}$. Then the triangle inequality and Assumption 4 on $F \in \mathcal{F}^{RC}$ yield

$$\begin{aligned}\|H^F(\mathbf{z}) - H^F(\overline{\mathbf{z}})\|_2 &= \|\mathbf{x}_0 - \overline{\mathbf{x}}_0\|_2 \\ &\leq \|F(\mathbf{x}_{-1}, \mathbf{z}_0) - F(\mathbf{x}_{-1}, \overline{\mathbf{z}}_0)\|_2 + \|F(\mathbf{x}_{-1}, \overline{\mathbf{z}}_0) - F(\overline{\mathbf{x}}_{-1}, \overline{\mathbf{z}}_0)\|_2 \\ &\leq L_R\|\mathbf{z}_0 - \overline{\mathbf{z}}_0\|_2 + r\|\mathbf{x}_{-1} - \overline{\mathbf{x}}_{-1}\|_2.\end{aligned}$$

By iterating this estimate one obtains

$$\|H^F(\mathbf{z}) - H^F(\overline{\mathbf{z}})\|_2 \leq r^i\|\mathbf{x}_{-i} - \overline{\mathbf{x}}_{-i}\|_2 + L_R\sum_{j=0}^{i-1} r^j\|\mathbf{z}_{-j} - \overline{\mathbf{z}}_{-j}\|_2,$$

from which the claim follows by Assumption 6. ▼
We now proceed to prove Proposition 5. Let $\widetilde{\mathbf{Z}} := \mathbf{Z}_0^{-n+1}$ and write for any $H \in \mathcal{H}^{RC}$

$$\begin{aligned}|\widehat{R}_n(H) - \widehat{R}_n^{\infty}(H)| &= \left|\frac{1}{n}\sum_{i=0}^{n-1} L(H(\mathbf{Z}_{-i}^{-n+1}), \mathbf{Y}_{-i}) - L(H(\mathbf{Z}_{-i}^{-\infty}), \mathbf{Y}_{-i})\right| \\ &\leq \frac{1}{n}\sum_{i=0}^{n-1} L_L\|h(H^F(\widetilde{\mathbf{Z}}_{-i}^{-\infty})) - h(H^F(\mathbf{Z}_{-i}^{-\infty}))\|_2 \\ &\leq \frac{1}{n}\sum_{i=0}^{n-1} L_L\overline{L_h}(2r^{n-i}M_{\mathcal{F}} + L_R\sum_{j=0}^{n-i-1} r^j\|\widetilde{\mathbf{Z}}_{-j-i} - \mathbf{Z}_{-j-i}\|_2) \\ &= \frac{2L_L\overline{L_h}M_{\mathcal{F}}}{n}\sum_{i=0}^{n-1} r^{n-i} \\ &= \frac{1-r^n}{n}\frac{2rL_L\overline{L_h}M_{\mathcal{F}}}{1-r} \\ &= \frac{1-r^n}{n}C_0.\end{aligned}$$

In these derivations, the first inequality follows from the triangle inequality and the Lipschitz continuity of the loss function (35), the second one is a consequence of the Lipschitz continuity of the readout map and of (75) in Lemma 18, which finally yield the claim with the choice of constant C_0 in (39). ■

5.4 Proof of Proposition 6

In order to simplify the notation, for any $i \in \mathbb{N}$ we define an $(\mathbb{R}^d)^{\mathbb{Z}_-} \times \mathbb{R}^m$-valued random variable $\mathbf{V}_{-i}$ as $\mathbf{V}_{-i} := (\mathbf{Z}_{-i}^{-\infty}, \mathbf{Y}_{-i})$ and denote its associated loss by $L_H(\mathbf{V}_{-i}) := L(H(\mathbf{Z}_{-i}^{-\infty}), \mathbf{Y}_{-i})$. We start by using the assumptions on the Lipschitz-continuity of both the loss function (35) and of the reservoir readout map and hence for any $i \in \mathbb{N}$ and $H \in \mathcal{H}^{RC}$ write

$$\begin{aligned}|L_H(\mathbf{V}_{-i})| &\leq |L_H(\mathbf{V}_{-i}) - L(\mathbf{0}, \mathbf{Y}_{-i})| + |L(\mathbf{0}, \mathbf{Y}_{-i})| \\ &\leq L_L \|H(\mathbf{Z}_{-i}^{-\infty})\|_2 + |L(\mathbf{0}, \mathbf{Y}_{-i})| \\ &\leq L_L \|h(H^F(\mathbf{Z}_{-i}^{-\infty})) - h(\mathbf{0})\|_2 + L_L \|h(\mathbf{0})\|_2 + |L(\mathbf{0}, \mathbf{Y}_{-i})| \\ &\leq L_L \overline{L_h} M_{\mathcal{F}} + |L(\mathbf{0}, \mathbf{Y}_{-i})| + L_{h,0} L_L.\end{aligned}$$

We continue by decomposing $n = k\tau + (n - k\tau)$ with $k = \lfloor \frac{n}{\tau} \rfloor$. For the last $(n - k\tau)$ elements one estimates

$$\begin{aligned}\mathbb{E}\left[\sup_{H \in \mathcal{H}^{RC}} \sum_{i=k\tau}^{n-1} \left\{\mathbb{E}\left[L_H(\mathbf{V}_{-i})\right] - L_H(\mathbf{V}_{-i})\right\}\right] &\leq 2(n - k\tau)(L_L \overline{L_h} M_{\mathcal{F}} + \mathbb{E}\left[|L(\mathbf{0}, \mathbf{Y}_0)|\right] + L_{h,0} L_L) \\ &= 2M(n - k\tau)\end{aligned}$$

with M as in (44). Subsequently using the definitions of the generalization error (13) and the idealized empirical risk (17) one obtains

$$\begin{aligned}&\mathbb{E}\left[\sup_{H \in \mathcal{H}^{RC}} \left\{R(H) - \widehat{R}_n^{\infty}(H)\right\}\right] \\ &= \mathbb{E}\left[\sup_{H \in \mathcal{H}^{RC}} \mathbb{E}[L(H(\mathbf{Z}), \mathbf{Y}_0)] - \left\{\frac{1}{n} \sum_{i=0}^{n-1} L(H(\mathbf{Z}_{-i}^{-\infty}), \mathbf{Y}_{-i})\right\}\right] \\ &= \mathbb{E}\left[\sup_{H \in \mathcal{H}^{RC}} \frac{1}{n} \sum_{i=0}^{n-1} \left\{\mathbb{E}[L_H(\mathbf{V}_{-i})] - L_H(\mathbf{V}_{-i})\right\}\right] \\ &\leq \frac{1}{n} \mathbb{E}\left[\sup_{H \in \mathcal{H}^{RC}} \sum_{i=0}^{k\tau-1} \left\{\mathbb{E}[L_H(\mathbf{V}_{-i})] - L_H(\mathbf{V}_{-i})\right\}\right] + \frac{2M(n - k\tau)}{n} \\ &= \frac{1}{n} \mathbb{E}\left[\sup_{H \in \mathcal{H}^{RC}} \sum_{j=0}^{k-1} \sum_{i=0}^{\tau-1} \left\{\mathbb{E}\left[L_H(\mathbf{V}_{-(\tau j+i)})\right] - L_H(\mathbf{V}_{-(\tau j+i)})\right\}\right] + \frac{2M(n - k\tau)}{n} \\ &\leq \frac{\tau}{n} \mathbb{E}\left[\sup_{H \in \mathcal{H}^{RC}} \sum_{j=0}^{k-1} \left\{\mathbb{E}[L_H(\mathbf{V}_{-\tau j})] - L_H(\mathbf{V}_{-\tau j})\right\}\right] + \frac{2M(n - k\tau)}{n}. \qquad (76)\end{aligned}$$

In order to obtain a bound for the first summand in the last expression, we introduce ghost samples and use tools that hinge on the independence between them. Let $\overline{\boldsymbol{\xi}}^{(j)} = (\overline{\boldsymbol{\xi}}_t^{y,(j)}, \overline{\boldsymbol{\xi}}_t^{z,(j)})_{t \in \mathbb{Z}_-}$, $j = 0, \ldots, k-1$ denote independent copies of $\boldsymbol{\xi}$. Next, for $I = y, z$ define $\boldsymbol{\xi}^{I,(j)}$ by setting $\boldsymbol{\xi}_i^{I,(j)} = \boldsymbol{\xi}_i^I$ for $i = -\tau(j+1) + 1, \ldots, 0$ and $\boldsymbol{\xi}_i^{I,(j)} = \overline{\boldsymbol{\xi}}_i^{I,(j)}$ for $i \leq -\tau(j+1)$. Additionally, let $\overline{\mathbf{Z}}_t^{(j)} := G^z(\ldots, \boldsymbol{\xi}_{t-1}^{z,(j)}, \boldsymbol{\xi}_t^{z,(j)})$ and $\overline{\mathbf{Y}}_t^{(j)} := G^y(\ldots, \boldsymbol{\xi}_{t-1}^{y,(j)}, \boldsymbol{\xi}_t^{y,(j)})$ for $t \in \mathbb{Z}_-$ and define the $(\mathbb{R}^d)^{\mathbb{Z}_-} \times \mathbb{R}^m$-valued random variables $\mathbf{U}^{(j)} := (\overline{\mathbf{Z}}_{-\tau j}^{-\infty,(j)}, \overline{\mathbf{Y}}_{-\tau j}^{(j)})$, $j = 0, \ldots, k-1$.

Then one has

$$\begin{aligned}\overline{\mathbf{Z}}^{(j)}_{t-\tau j} &= G^z(\ldots, \boldsymbol{\xi}^{z,(j)}_{t-\tau j-1}, \boldsymbol{\xi}^{z,(j)}_{t-\tau j}) \\ &= \begin{cases} G^z(\ldots, \overline{\boldsymbol{\xi}}^{z,(j)}_{-\tau j-\tau}, \boldsymbol{\xi}^{z}_{-\tau j-\tau+1}, \ldots, \boldsymbol{\xi}^{z}_{t-\tau j}), & t = -\tau+1, \ldots, 0, \\ G^z(\ldots, \overline{\boldsymbol{\xi}}^{z,(j)}_{t-\tau j-1}, \overline{\boldsymbol{\xi}}^{z,(j)}_{t-\tau j}), & t \leq -\tau \end{cases}\end{aligned}$$

and so, for any $j = 0, \ldots, k-1$, the random variable $\mathbf{U}^{(j)}$ is measurable with respect to the σ-algebra generated by $(\boldsymbol{\xi}_{-\tau j+t})_{t=-\tau+1,\ldots,0}$ and $\overline{\boldsymbol{\xi}}^{(j)}$. The assumption of independence between the ghost samples implies that $\mathbf{U}^{(0)}, \ldots, \mathbf{U}^{(k-1)}$ are also independent and identically distributed with the same distribution as $\mathbf{V}_0 = (\mathbf{Z}, \mathbf{Y}_0)$ introduced above. Hence one can rewrite the first summand of the right hand side of the last inequality in (76) as

$$\begin{aligned}&\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}} \sum_{j=0}^{k-1} \{\mathbb{E}[L_H(\mathbf{V}_{-\tau j})] - L_H(\mathbf{V}_{-\tau j})\}\right] \\ &\leq \mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}} \sum_{j=0}^{k-1} \left\{\mathbb{E}[L_H(\mathbf{V}_{-\tau j})] - L_H(\mathbf{U}^{(j)})\right\}\right] + \mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}} \sum_{j=0}^{k-1} \left\{L_H(\mathbf{U}^{(j)}) - L_H(\mathbf{V}_{-\tau j})\right\}\right] \\ &= \mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}} \sum_{j=0}^{k-1} \left\{\mathbb{E}[L_H(\mathbf{U}^{(j)})] - L_H(\mathbf{U}^{(j)})\right\}\right] + \mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}} \sum_{j=0}^{k-1} \left\{L_H(\mathbf{U}^{(j)}) - L_H(\mathbf{V}_{-\tau j})\right\}\right].\end{aligned} \tag{77}$$

We now analyze these two terms separately. For the second term, we first note that for any $H \in \mathcal{H}^{RC}$ it holds that

$$\begin{aligned}\left|L_H(\mathbf{V}_{-\tau j}) - L_H(\mathbf{U}^{(j)})\right| &= \left|L(H(\mathbf{Z}^{-\infty}_{-\tau j}), \mathbf{Y}_{-\tau j}) - L(H(\overline{\mathbf{Z}}^{-\infty,(j)}_{-\tau j}), \overline{\mathbf{Y}}^{(j)}_{-\tau j})\right| \\ &\leq L_L\left(\|H(\mathbf{Z}^{-\infty}_{-\tau j}) - H(\overline{\mathbf{Z}}^{-\infty,(j)}_{-\tau j})\|_2 + \|\mathbf{Y}_{-\tau j} - \overline{\mathbf{Y}}^{(j)}_{-\tau j}\|_2\right).\end{aligned} \tag{78}$$

Next, we use the Lipschitz-continuity of the readout maps (see (31)) together with the estimate (75) in Lemma 18 and compute

$$\sup_{H\in\mathcal{H}^{RC}} \|H(\mathbf{Z}^{-\infty}_{-\tau j}) - H(\overline{\mathbf{Z}}^{-\infty,(j)}_{-\tau j})\|_2 \leq 2r^\tau M_{\mathcal{F}} \overline{L_h} + L_R \overline{L_h} \sum_{l=0}^{\tau-1} r^l \|\mathbf{Z}_{-l-\tau j} - \overline{\mathbf{Z}}^{(j)}_{-l-\tau j}\|_2. \tag{79}$$

Combining (79) with (78) we now estimate the second term in (77) by

$$\begin{aligned}&\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}} \sum_{j=0}^{k-1} \left\{L_H(\mathbf{U}^{(j)}) - L_H(\mathbf{V}_{-\tau j})\right\}\right] \\ &\leq L_L \sum_{j=0}^{k-1} \left(2r^\tau M_{\mathcal{F}} \overline{L_h} + \mathbb{E}[\|\mathbf{Y}_{-\tau j} - \overline{\mathbf{Y}}^{(j)}_{-\tau j}\|_2] + L_R \overline{L_h} \sum_{l=0}^{\tau-1} r^l \mathbb{E}[\|\mathbf{Z}_{-l-\tau j} - \overline{\mathbf{Z}}^{(j)}_{-l-\tau j}\|_2]\right) \\ &= k a_\tau\end{aligned} \tag{80}$$

with a_τ as in (45).

In order to estimate the first term in (77), one relies on techniques which are common in the case of independent and identically distributed random variables. Here we start by introducing real Rademacher random variables $\varepsilon_0, \ldots, \varepsilon_{k-1}$ (see for example Hytönen et al., 2016, Definition 3.2.9), which are independent of all the other random variables considered so far. In what follows we need to use the structure for the loss function introduced in (33) as well as the other hypotheses on it that we spelled out in Section 3.2. The fact that the loss functions are a sum of representing functions $f_i : \mathbb{R} \longrightarrow \mathbb{R}$, implies that their evaluation on the hypothesis class $\mathcal{H}^{RC}$ can be expressed through the evaluation of each representing function on the sets $\mathcal{H}_i^{RC}$, $i \in \{1, \ldots, m\}$, defined by

$$\mathcal{H}_i^{RC} := \{\widetilde{H} \colon (D_d)^{\mathbb{Z}_-} \times \mathbb{R}^m \to \mathbb{R} \mid \widetilde{H}(\mathbf{x}, \mathbf{y}) := (H(\mathbf{x}))_i - y_i, H \in \mathcal{H}^{RC}\}.$$

Using independence and a symmetrization trick by Giné and Zinn (1984) (see for example Ledoux and Talagrand (1991, Lemma 6.3) or the proof of Bartlett and Mendelson (2003, Theorem 8)) one writes

$$\begin{aligned}\frac{1}{k}\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\sum_{j=0}^{k-1}\left\{\mathbb{E}[L_H(\mathbf{U}^{(j)})] - L_H(\mathbf{U}^{(j)})\right\}\right] &\leq \frac{2}{k}\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\sum_{j=0}^{k-1}\varepsilon_j L_H(\mathbf{U}^{(j)})\right] \\ &\leq \frac{2}{k}\sum_{i=1}^{m}\mathbb{E}\left[\sup_{\widetilde{H}\in\mathcal{H}_i^{RC}}\sum_{j=0}^{k-1}\varepsilon_j(f_i\circ\widetilde{H})(\mathbf{U}^{(j)})\right].\end{aligned} \tag{81}$$

Applying the contraction principle for Rademacher random variables (see Ben-David and Shalev-Shwartz (2014, Lemma 26.9) and also Ledoux and Talagrand (1991, Theorem 4.12)) to the last expression one obtains

$$\begin{aligned}&\frac{1}{k}\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\sum_{j=0}^{k-1}\left\{\mathbb{E}[L_H(\mathbf{U}^{(j)})] - L_H(\mathbf{U}^{(j)})\right\}\right] \\ &\leq \frac{2L_L}{\sqrt{m}k}\sum_{i=1}^{m}\mathbb{E}\left[\sup_{\widetilde{H}\in\mathcal{H}_i^{RC}}\sum_{j=0}^{k-1}\varepsilon_j\widetilde{H}(\mathbf{U}^{(j)})\right] \\ &= \frac{2L_L}{\sqrt{m}k}\sum_{i=1}^{m}\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\sum_{j=0}^{k-1}\varepsilon_j\left(H(\overline{\mathbf{Z}}_{-\tau j}^{-\infty,(j)}) - \overline{\mathbf{Y}}_{-\tau j}^{(j)}\right)_i\right] \\ &\leq \frac{2L_L}{\sqrt{m}k}\sum_{i=1}^{m}\left(\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left\|\sum_{j=0}^{k-1}\varepsilon_j H(\overline{\mathbf{Z}}_{-\tau j}^{-\infty,(j)})\right\|_2\right] + \mathbb{E}\left[-\sum_{j=0}^{k-1}\varepsilon_j(\overline{\mathbf{Y}}_{-\tau j}^{(j)})_i\right]\right) \\ &\leq \frac{2\sqrt{m}L_L}{k}\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left\|\sum_{j=0}^{k-1}\varepsilon_j H(\overline{\mathbf{Z}}_{-\tau j}^{-\infty,(j)})\right\|_2\right] = 2\sqrt{m}L_L\mathcal{R}_k(\mathcal{H}^{RC}),\end{aligned} \tag{82}$$

with the Rademacher complexity defined as in (40). Note that the last term in the fourth line is equal to zero due to the independence and the fact that the expectation of Rademacher random variables is zero.

We now come back to the estimate of the expected maximum difference between the in-class statistical risk and the idealized empirical risk and rewrite the expression (76) using (77), (80), and (82)

$$\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left\{R(H)-\widehat{R}_n^{\infty}(H)\right\}\right]\leq\frac{\tau}{n}\left\{2k\sqrt{m}L_L\mathcal{R}_k(\mathcal{H}^{RC})+ka_\tau\right\}+\frac{2M(n-k\tau)}{n},$$

which then yields (41) as required.

It remains to prove (42). To do so, notice that the triangle inequality and the same arguments used in (76), (77), (80) and (81) may be applied in the presence of absolute values to obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left|\widehat{R}_n^{\infty}(H)-R(H)\right|\right]\leq\frac{k\tau}{n}a_\tau+\frac{2\tau}{n}\sum_{i=1}^{m}\mathbb{E}\left[\sup_{\widetilde{H}\in\mathcal{H}_i^{RC}}\left|\sum_{j=0}^{k-1}\varepsilon_j(f_i\circ\widetilde{H})(\mathbf{U}^{(j)})\right|\right]\\+\frac{2M(n-k\tau)}{n}.\end{aligned}\tag{83}$$

Applying again the contraction principle for Rademacher random variables (Bartlett and Mendelson, 2003, Theorem 12.4), one estimates, for any $i=1,\dots,m$,

$$\begin{aligned}&\frac{1}{k}\sum_{i=1}^{m}\mathbb{E}\left[\sup_{\widetilde{H}\in\mathcal{H}_i^{RC}}\left|\sum_{j=0}^{k-1}\varepsilon_j(f_i\circ\widetilde{H})(\mathbf{U}^{(j)})\right|\right]\\&\leq\frac{2L_L}{\sqrt{mk}}\sum_{i=1}^{m}\mathbb{E}\left[\sup_{\widetilde{H}\in\mathcal{H}_i^{RC}}\left|\sum_{j=0}^{k-1}\varepsilon_j\widetilde{H}(\mathbf{U}^{(j)})\right|\right]\\&\leq\frac{2L_L}{\sqrt{mk}}\sum_{i=1}^{m}\left(\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left|\sum_{j=0}^{k-1}\varepsilon_j(H(\overline{\mathbf{Z}}_{-\tau j}^{-\infty,(j)}))_i\right|\right]+\mathbb{E}\left[\left|\sum_{j=0}^{k-1}\varepsilon_j(\overline{\mathbf{Y}}_{-\tau j}^{(j)})_i\right|\right]\right)\\&\leq\frac{2L_L}{\sqrt{mk}}\left(m\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left\|\sum_{j=0}^{k-1}\varepsilon_jH(\overline{\mathbf{Z}}_{-\tau j}^{-\infty,(j)})\right\|_2\right]+\sum_{i=1}^{m}\mathbb{E}\left[\left|\sum_{j=0}^{k-1}\varepsilon_j(\overline{\mathbf{Y}}_{-\tau j}^{(j)})_i\right|^2\right]^{1/2}\right)\\&\leq\frac{2L_L}{\sqrt{mk}}\left(mk\mathcal{R}_k(\mathcal{H}^{RC})+\sum_{i=1}^{m}\mathbb{E}\left[\left|\sum_{j=0}^{k-1}\varepsilon_j(\overline{\mathbf{Y}}_{-\tau j}^{(j)})_i\right|^2\right]^{1/2}\right)\end{aligned}\tag{84}$$

with the Rademacher complexity defined as in (40). Finally, using the independence of the Rademacher sequence and the ghost samples $(\overline{\mathbf{Y}}^{(j)})_{j=0,\dots,k-1}$ as well as the stationarity properties of the latter, one has

$$\sum_{i=1}^{m}\left(\mathbb{E}\left[\left|\sum_{j=0}^{k-1}\varepsilon_j(\overline{\mathbf{Y}}_{-\tau j}^{(j)})_i\right|^2\right]\right)^{1/2}=\sum_{i=1}^{m}\left(\sum_{j=0}^{k-1}\mathbb{E}\left[(\overline{\mathbf{Y}}_{-\tau j}^{(j)})_i^2\right]\right)^{1/2}\leq\sqrt{k}\sqrt{m}\mathbb{E}\left[\|\mathbf{Y}_0\|_2^2\right]^{1/2}.$$

The combination of this inequality with (83) and (84) yields (42), as required. ■

5.5 Proof of Corollary 8

Proof of part (i) We start by noticing that under Assumption 1, the weak dependence coefficients θ^I defined in (46) for $I = y, z$ and $\tau \in \mathbb{N}^+$ can be estimated as

$$\begin{aligned}\theta^I(\tau) &= \mathbb{E}[\|G^I(\ldots, \boldsymbol{\xi}^I_{-1}, \boldsymbol{\xi}^I_0) - G^I(\ldots, \widetilde{\boldsymbol{\xi}}^I_{-\tau-1}, \widetilde{\boldsymbol{\xi}}^I_{-\tau}, \boldsymbol{\xi}^I_{-\tau+1}, \ldots, \boldsymbol{\xi}^I_0)\|_2] \\ &\leq \mathbb{E}\left[L_I \sum_{j=\tau}^{\infty} w_j^I \|\boldsymbol{\xi}^I_{-j} - \widetilde{\boldsymbol{\xi}}^I_{-j}\|_2\right] \\ &\leq 2L_I \mathbb{E}[\|\boldsymbol{\xi}^I_0\|_2] \sum_{j=\tau}^{\infty} w_j^I \leq 2L_I \mathbb{E}[\|\boldsymbol{\xi}^I_0\|_2] \sum_{j=\tau}^{\infty} (D_{w^I})^j = 2L_I \mathbb{E}[\|\boldsymbol{\xi}^I_0\|_2] \frac{(D_{w^I})^\tau}{1 - D_{w^I}},\end{aligned}$$

where we used that by hypothesis $D_{w^I} < 1$. Consequently, if we set $C_I := \dfrac{2L_I \mathbb{E}[\|\boldsymbol{\xi}^I_0\|_2]}{1 - D_{w^I}}$, condition (26) does hold for all $\tau \in \mathbb{N}^+$. We now define $c_0 := 2L_L\overline{L_h}M_{\mathcal{F}}$, $c_1 := L_L L_R C_z \overline{L_h}$, and $c_2 := L_L C_y$ and with the notation $\lambda_{max} := \max(r, D_{w^y}, D_{w^z}) \in (0, 1)$ write (45) for any $\tau \in \mathbb{N}^+$ as

$$a_\tau \leq c_0 r^\tau + c_1 \sum_{l=0}^{\tau-1} r^l (D_{w^z})^{\tau-l} + c_2 (D_{w^y})^\tau \leq \lambda_{max}^\tau (c_0 + \tau c_1 + c_2). \tag{85}$$

Next, let $\tau \in \mathbb{N}^+$ with $\tau < n$ and set $k = \lfloor n/\tau \rfloor$. Inserting assumption (48) in (41) and then using that $n/\tau - 1 \leq k \leq n/\tau$, one obtains

$$\begin{aligned}\mathbb{E}\left[\sup_{H \in \mathcal{H}^{RC}} \left\{R(H) - \widehat{R}_n^\infty(H)\right\}\right] &\leq \frac{k\tau}{n} a_\tau + \frac{BC_{RC}\sqrt{k\tau}}{n} + \frac{2M(n - k\tau)}{n} \\ &\leq a_\tau + \frac{BC_{RC}\sqrt{\tau}}{\sqrt{n}} + \frac{2M\tau}{n}.\end{aligned} \tag{86}$$

Our goal now is to choose the length of the block τ depending on λ_{max}. Recall that by hypothesis $\log(n) < n \log(\lambda_{max}^{-1})$, which means that in order to be able to apply the blocking technique for a given value $\lambda_{max} \in (0, 1)$ the number of observations $n \in \mathbb{N}^+$ should be sufficiently large. In this situation one can choose $\tau = \lfloor \log(n)/\log(\lambda_{max}^{-1}) \rfloor$, which is then guaranteed to satisfy $\tau < n$. Notice that then $\lambda_{max}^{\tau+1} \leq 1/n$ and consequently (49) follows from (85) and (86) with the appropriate choice of constants as given in (51)-(52). Finally, the last term in (50) follows by noticing that

$$\frac{4\tau\sqrt{k} L_L \mathbb{E}\left[\|\mathbf{Y}_0\|_2^2\right]^{1/2}}{n} \leq \frac{4\sqrt{\tau} L_L \mathbb{E}\left[\|\mathbf{Y}_0\|_2^2\right]^{1/2}}{\sqrt{n}}, \tag{87}$$

and hence one gets (50) as required.

Proof of part (ii) Recall that by Assumption 2 for $I = y, z$ there exist $\lambda_I \in (0, 1)$ and $C_I > 0$ such that, for all $\tau \in \mathbb{N}^+$, it holds that

$$\theta^I(\tau) \leq C_I \lambda_I^\tau.$$

Mimicking the proof of part **(i)** with $\lambda_{max} := \max(r, \lambda_y, \lambda_z)$ yields the claim.

Proof of part (iii) By our choice of γ_α it holds that $\tau/4 + \gamma_\alpha \geq \log(\tau)\alpha_z / \log(r^{-1})$ for all $\tau \in \mathbb{N}^+$. One has $r^{l/2}(\tau - l)^{-\alpha_z} \leq 2^{\alpha_z}\tau^{-\alpha_z}$ for $l \leq \tau/2$ and $r^{l/2}(\tau-l)^{-\alpha_z} \leq r^{\tau/4} \leq r^{-\gamma_\alpha}\tau^{-\alpha_z}$, for $\tau/2 \leq l \leq \tau - 1$. Setting $C_\alpha = \max(2^{\alpha_z}, r^{-\gamma_\alpha})(1-\sqrt{r})^{-1}$ one has for all $\tau \in \mathbb{N}^+$ that

$$\sum_{l=0}^{\tau-1} r^l(\tau - l)^{-\alpha_z} \leq \sum_{l=0}^{\infty} r^{l/2} \max(2^{\alpha_z}, r^{-\gamma_\alpha})\tau^{-\alpha_z} = C_\alpha \tau^{-\alpha_z}.$$

Defining c_0, c_1, and c_2 as in the proof of part **(i)**, applying (30) and inserting the above estimate thus allows us to bound (45) for any $\tau \in \mathbb{N}^+$ by

$$a_\tau \leq c_0 r^\tau + c_1 \sum_{l=0}^{\tau-1} r^l(\tau - l)^{-\alpha_z} + c_2\tau^{-\alpha_y} \leq \tau^{-\alpha}\left(r^{-\gamma_\alpha}c_0 + C_\alpha c_1 + c_2\right).$$

Furthermore, (86) remains valid and so choosing $\tau = n^\beta$ yields

$$\mathbb{E}\left[\sup_{H \in \mathcal{H}^{RC}} \left\{R(H) - \widehat{R}_n^\infty(H)\right\}\right] \leq \frac{(r^{-\gamma_\alpha}c_0 + C_\alpha c_1 + c_2)}{n^{\alpha\beta}} + \frac{BC_{RC}}{n^{1/2-\beta/2}} + \frac{2M}{n^{1-\beta}}.$$

Taking $\beta = \frac{1}{2}(\alpha + \frac{1}{2})^{-1}$ yields $1/2 - \beta/2 = \frac{\alpha}{2}(\alpha + \frac{1}{2})^{-1}$ and hence the desired result. The last term in (42) can be bounded by proceeding analogously to part **(i)** and noticing that (87) remains valid, which concludes the proof. ■

5.6 Proof of Proposition 9 (Reservoir systems with linear reservoir and readout maps)

The condition (61) together with Proposition 1 ensure that the ESP property of the reservoir systems in the hypothesis class is guaranteed and that for any $\mathbf{z} \in K_M$ we have that $H^{A,C,\boldsymbol{\zeta}}(\mathbf{z}) = \sum_{i=0}^\infty A^i(C\mathbf{z}_{-i} + \boldsymbol{\zeta})$. Using the definition of Rademacher complexity, one estimates

$$\begin{aligned}
&\mathbb{E}\left[\sup_{H \in \mathcal{H}^{RC}} \left\|\sum_{j=0}^{k-1} \varepsilon_j H(\widetilde{\mathbf{Z}}^{(j)})\right\|_2\right] \\
&= \mathbb{E}\left[\sup_{\substack{(A,C,\boldsymbol{\zeta}) \in \Theta \\ W:\|\!|W|\!\|_2 \leq \overline{L_h} \\ \boldsymbol{a}:\|\boldsymbol{a}\|_2 \leq L_{h,0}}} \left\|\sum_{j=0}^{k-1} \varepsilon_j (W H^{A,C,\boldsymbol{\zeta}}(\widetilde{\mathbf{Z}}^{(j)}) + \boldsymbol{a})\right\|_2\right] \\
&\leq \mathbb{E}\left[\sup_{\substack{(A,C,\boldsymbol{\zeta}) \in \Theta \\ W:\|\!|W|\!\|_2 \leq \overline{L_h}}} \left\|\sum_{j=0}^{k-1} \varepsilon_j W H^{A,C,\boldsymbol{\zeta}}(\widetilde{\mathbf{Z}}^{(j)})\right\|_2\right] + \mathbb{E}\left[\sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2 \leq L_{h,0}} \left\|\sum_{j=0}^{k-1} \varepsilon_j \boldsymbol{a}\right\|_2\right] \\
&\leq \sup_{W:\|\!|W|\!\|_2 \leq \overline{L_h}} |\!|\!|W|\!|\!|_2 \mathbb{E}\left[\sup_{(A,C,\boldsymbol{\zeta}) \in \Theta} \left\|\sum_{j=0}^{k-1} \varepsilon_j H^{A,C,\boldsymbol{\zeta}}(\widetilde{\mathbf{Z}}^{(j)})\right\|_2\right] + \mathbb{E}\left[\sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2 \leq L_{h,0}} \left\|\sum_{j=0}^{k-1} \varepsilon_j \boldsymbol{a}\right\|_2\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \overline{L_h}\sum_{l=0}^{\infty}\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}\left|\!\left|\!\left|A^l\right|\!\right|\!\right|_2\mathbb{E}\left[\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}\left\|\sum_{j=0}^{k-1}\varepsilon_j(C\widetilde{\mathbf{Z}}_{-l}^{(j)}+\boldsymbol{\zeta})\right\|_2\right]+\mathbb{E}\left[\sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2\leq L_{h,0}}\left\|\sum_{j=0}^{k-1}\varepsilon_j\boldsymbol{a}\right\|_2\right]\\
&\leq \overline{L_h}\sum_{l=0}^{\infty}\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}\left|\!\left|\!\left|A^l\right|\!\right|\!\right|_2\left(\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}|\!|\!|C|\!|\!|_2\mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\varepsilon_j\widetilde{\mathbf{Z}}_{-l}^{(j)}\right\|_2\right]+\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}\|\boldsymbol{\zeta}\|_2\mathbb{E}\left[\left|\sum_{j=0}^{k-1}\varepsilon_j\right|\right]\right)\\
&\quad+\mathbb{E}\left[\sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2\leq L_{h,0}}\left\|\sum_{j=0}^{k-1}\varepsilon_j\boldsymbol{a}\right\|_2\right]\\
&\leq \overline{L_h}\sum_{l=0}^{\infty}(\lambda_{max}^A)^l\left(\lambda_{max}^C\mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\varepsilon_j\widetilde{\mathbf{Z}}_0^{(j)}\right\|_2\right]+\lambda_{max}^{\boldsymbol{\zeta}}\mathbb{E}\left[\left|\sum_{j=0}^{k-1}\varepsilon_j\right|\right]\right)+\mathbb{E}\left[\sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2\leq L_{h,0}}\left\|\sum_{j=0}^{k-1}\varepsilon_j\boldsymbol{a}\right\|_2\right],
\end{aligned}
\tag{88}
$$

where we used stationarity, the fact that $|\!|\!|W|\!|\!|_2\leq\overline{L_h}$ for all readout maps from the class $H\in\mathcal{H}^{RC}$, and constants as in (58)-(60). For the first summand in this expression we have

$$
\mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\varepsilon_j\widetilde{\mathbf{Z}}_0^{(j)}\right\|_2\right]^2\leq\mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\varepsilon_j\widetilde{\mathbf{Z}}_0^{(j)}\right\|_2^2\right]=\sum_{j=0}^{k-1}\mathbb{E}\left[\left\|\widetilde{\mathbf{Z}}_0^{(j)}\right\|_2^2\right]\mathbb{E}[\varepsilon_j^2]=k\mathbb{E}\left[\|\mathbf{Z}_0\|_2^2\right],
$$

where in the first step we used the Jensen inequality, the next equality is obtained using the independence of the ghost samples and the Rademacher sequence and also the fact that $\mathbb{E}[\varepsilon_{j'}\varepsilon_j]=0$ when $j\neq j'$ for Rademacher variables. The second summand in (88) is bounded using the inequality by Khintchine (1923)

$$
\mathbb{E}\left[\left|\sum_{j=0}^{k-1}\varepsilon_j\right|\right]\leq\sqrt{k}.
$$

We bound the third term as the first one and obtain

$$
\mathbb{E}\left[\sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2\leq L_{h,0}}\left\|\sum_{j=0}^{k-1}\varepsilon_j\boldsymbol{a}\right\|_2\right]^2\leq kL_{h,0}^2,
\tag{89}
$$

where we took into account that $\|\boldsymbol{a}\|_2\leq L_{h,0}$ for all readout maps from the class $H\in\mathcal{H}^{RC}$. Finally, (88) can be rewritten as

$$
\begin{aligned}
\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left\|\sum_{j=0}^{k-1}\varepsilon_jH(\widetilde{\mathbf{Z}}^{(j)})\right\|_2\right]&\leq\sqrt{k}\ \overline{L_h}\sum_{l=0}^{\infty}(\lambda_{max}^A)^l\left(\lambda_{max}^C\mathbb{E}\left[\|\mathbf{Z}_0\|_2^2\right]^{1/2}+\lambda_{max}^{\boldsymbol{\zeta}}\right)+\sqrt{k}L_{h,0}\\
&=\sqrt{k}\left(\frac{\overline{L_h}}{1-\lambda_{max}^A}\left(\lambda_{max}^C\mathbb{E}\left[\|\mathbf{Z}_0\|_2^2\right]^{1/2}+\lambda_{max}^{\boldsymbol{\zeta}}\right)+L_{h,0}\right),
\end{aligned}
$$

where we used that $\lambda_{max}^A\in(0,1)$. Finally, the choice of constants which satisfy conditions (61) yields (62), as required. ■

5.7 Proof of Proposition 11 (Echo State Networks)

Firstly, note that for any $\mathbf{x} \in D_N$, it holds that $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 = \sum_{i=1}^N |x_i|$ and hence one can write

$$\begin{aligned}
\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left\|\sum_{j=0}^{k-1}\varepsilon_j H(\widetilde{\mathbf{Z}}^{(j)})\right\|_2\right] &= \mathbb{E}\left[\sup_{\substack{F\in\mathcal{F}^{RC}\\ W:\|W\|_2\leq\overline{L_h}\\ \boldsymbol{a}:\|\boldsymbol{a}\|_2\leq L_{h,0}}}\left\|\sum_{j=0}^{k-1}\varepsilon_j(WH^F(\widetilde{\mathbf{Z}}^{(j)})+\boldsymbol{a})\right\|_2\right]\\
&\leq \mathbb{E}\left[\sup_{\substack{F\in\mathcal{F}^{RC}\\ W:\|W\|_2\leq\overline{L_h}}}\left\|\sum_{j=0}^{k-1}\varepsilon_j WH^F(\widetilde{\mathbf{Z}}^{(j)})\right\|_2\right]+\mathbb{E}\left[\sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2\leq L_{h,0}}\left\|\sum_{j=0}^{k-1}\varepsilon_j\boldsymbol{a}\right\|_2\right]\\
&\leq \overline{L_h}\sum_{l=1}^N\mathbb{E}\left[\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}\left|\sum_{j=0}^{k-1}\varepsilon_j H_l^{\sigma,A,C,\boldsymbol{\zeta}}(\widetilde{\mathbf{Z}}^{(j)})\right|\right]+\sqrt{k}L_{h,0},
\end{aligned}$$

where we used the same arguments as in (89). Using the assumed symmetry of the family $\mathcal{F}^{RC}$ in the first step and the contraction principle in the second step one may estimate

$$\begin{aligned}
&\sum_{l=1}^N\mathbb{E}\left[\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}\left|\sum_{j=0}^{k-1}\varepsilon_j H_l^{\sigma,A,C,\boldsymbol{\zeta}}(\widetilde{\mathbf{Z}}^{(j)})\right|\right]\\
&\quad=\sum_{l=1}^N\mathbb{E}\left[\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}\sum_{j=0}^{k-1}\varepsilon_j H_l^{\sigma,A,C,\boldsymbol{\zeta}}(\widetilde{\mathbf{Z}}^{(j)})\right]\\
&\quad\leq L_\sigma\sum_{l=1}^N\mathbb{E}\left[\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}\sum_{j=0}^{k-1}\varepsilon_j(A_{l,\cdot}H^{\sigma,A,C,\boldsymbol{\zeta}}(\widetilde{\mathbf{Z}}_{-1}^{-\infty,(j)})+C_{l,\cdot}\widetilde{\mathbf{Z}}_0^{(j)}+\zeta_l)\right]\\
&\quad\leq L_\sigma\sum_{l=1}^N\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}\|A_{l,\cdot}\|_\infty\mathbb{E}\left[\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}\left\|\sum_{j=0}^{k-1}\varepsilon_j H^{\sigma,A,C,\boldsymbol{\zeta}}(\widetilde{\mathbf{Z}}_{-1}^{-\infty,(j)})\right\|_1\right]\\
&\qquad+L_\sigma\sum_{l=1}^N\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}\|C_{l,\cdot}\|_2\mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\varepsilon_j\widetilde{\mathbf{Z}}_0^{(j)}\right\|_2\right]+L_\sigma\sum_{l=1}^N\mathbb{E}\left[\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}\sum_{j=0}^{k-1}\varepsilon_j\zeta_l\right].
\end{aligned}$$

Since $\lambda_{max}^A \in (0,1)$ by assumption and using that our hypotheses ensure that Assumption 6 is satisfied, one may iterate the above inequality to obtain

$$\sum_{l=1}^N\mathbb{E}\left[\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}\left|\sum_{j=0}^{k-1}\varepsilon_j H_l^{\sigma,A,C,\boldsymbol{\zeta}}(\widetilde{\mathbf{Z}}^{(j)})\right|\right]\leq\sum_{l=0}^\infty(\lambda_{max}^A)^l\left(\lambda_{max}^C\mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\varepsilon_j\widetilde{\mathbf{Z}}_0^j\right\|_2\right]+\lambda_{max}^{\boldsymbol{\zeta}}\mathbb{E}\left[\left|\sum_{j=0}^{k-1}\varepsilon_j\right|\right]\right). \tag{90}$$

For the first summand in this expression we obtain

$$\mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\varepsilon_j\widetilde{\mathbf{Z}}_0^{(j)}\right\|_2\right]^2 \leq \mathbb{E}\left[\left\|\sum_{j=0}^{k-1}\varepsilon_j\widetilde{\mathbf{Z}}_0^{(j)}\right\|_2^2\right] = \sum_{j=0}^{k-1}\mathbb{E}\left[\left\|\widetilde{\mathbf{Z}}_0^{(j)}\right\|_2^2\right]\mathbb{E}[\varepsilon_j^2] = k\mathbb{E}\left[\|\mathbf{Z}_0\|_2^2\right],$$

where in the first step we used Jensen's inequality, the next equality is obtained using the independence of the ghost samples and the Rademacher sequence and also the fact that $\mathbb{E}[\varepsilon_{j'}\varepsilon_j] = 0$ for $j \neq j'$ for Rademacher variables. The last step trivially follows again from the definition of ghost samples and the definition of Rademacher variables.

The second summand in (90) is bounded using Khintchine's inequality (Khintchine, 1923)

$$\mathbb{E}\left[\left|\sum_{j=0}^{k-1}\varepsilon_j\right|\right] \leq \sqrt{k}$$

and hence in (90) we obtain

$$\begin{aligned}\sum_{l=1}^{N}\mathbb{E}\left[\sup_{(A,C,\boldsymbol{\zeta})\in\Theta}\left|\sum_{j=0}^{k-1}\varepsilon_j H_l^{\sigma,A,C,\boldsymbol{\zeta}}(\widetilde{\mathbf{Z}}^{(j)})\right|\right] &\leq \sqrt{k}\sum_{l=0}^{\infty}(\lambda_{max}^A)^l\left(\lambda_{max}^C\mathbb{E}\left[\|\mathbf{Z}_0\|_2^2\right]^{1/2}+\lambda_{max}^{\boldsymbol{\zeta}}\right)\\ &= \frac{\sqrt{k}}{1-\lambda_{max}^A}\left(\lambda_{max}^C\mathbb{E}\left[\|\mathbf{Z}_0\|_2^2\right]^{1/2}+\lambda_{max}^{\boldsymbol{\zeta}}\right),\end{aligned}$$

which finally gives

$$\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left\|\sum_{j=0}^{k-1}\varepsilon_j H(\widetilde{\mathbf{Z}}^{(j)})\right\|_2\right] \leq \sqrt{k}\left(\frac{\overline{L_h}}{1-\lambda_{max}^A}\left(\lambda_{max}^C\mathbb{E}\left[\|\mathbf{Z}_0\|_2^2\right]^{1/2}+\lambda_{max}^{\boldsymbol{\zeta}}\right)+L_{h,0}\right)$$

which with the choice of constant in (65) gives (64) as required. ■

5.8 Proof of Proposition 13 (State-Affine Systems)

Let $(p,q) \in \Theta$. Then, the conditions on the quantities λ^{SAS} and c^{SAS} introduced in (66) imply that

$$M_p = \max_{\mathbf{z}\in\overline{B_{\|\cdot\|}(\mathbf{0},M)}}\{|||p(\mathbf{z})|||_2\} < 1 \tag{91}$$

and so $F^{p,q}$ is a contraction on the first entry. Thus, Proposition 1 implies that the system (3) has the echo state property. Moreover, by Grigoryeva and Ortega (2018a, Proposition 14), for any $\mathbf{z} \in K_M$, we have that

$$H^{p,q}(\mathbf{z}) = \sum_{i=0}^{\infty} p(\mathbf{z}_0)p(\mathbf{z}_{-1})\cdots p(\mathbf{z}_{-i+1})q(\mathbf{z}_{-i}).$$

Next, write $p(\mathbf{z}) = \sum_{\boldsymbol{\alpha}\in I_{max}}\mathbf{z}^{\boldsymbol{\alpha}}B_{\boldsymbol{\alpha}}$ and $q(\mathbf{z}) = \sum_{\boldsymbol{\alpha}\in I_{max}}\mathbf{z}^{\boldsymbol{\alpha}}C_{\boldsymbol{\alpha}}$. Again, by the conditions on the quantities λ^{SAS} and c^{SAS} introduced in (66) one has that the image of $H^{p,q}$ is bounded

and so, combining this with (91) one obtains

$$
\begin{aligned}
&\Big\| \sum_{j=0}^{k-1} \varepsilon_j H^{p,q}(\widetilde{\mathbf{Z}}^{(j)}) \Big\|_2 \\
&\quad \leq \sum_{i=0}^{\infty} \left\| \sum_{j=0}^{k-1} \varepsilon_j p(\widetilde{\mathbf{Z}}_0^{(j)}) \cdots p(\widetilde{\mathbf{Z}}_{-i+1}^{(j)}) q(\widetilde{\mathbf{Z}}_{-i}^{(j)}) \right\|_2 \\
&\quad \leq \sum_{i=0}^{\infty} \sum_{\boldsymbol{\alpha} \in I_{max}} \left\| B_{\boldsymbol{\alpha}} \sum_{j=0}^{k-1} \varepsilon_j (\widetilde{\mathbf{Z}}_0^{(j)})^{\boldsymbol{\alpha}} p(\widetilde{\mathbf{Z}}_{-1}^{(j)}) \cdots p(\widetilde{\mathbf{Z}}_{-i+1}^{(j)}) q(\widetilde{\mathbf{Z}}_{-i}^{(j)}) \right\|_2 \\
&\quad \leq \sum_{i=0}^{\infty} \sum_{\boldsymbol{\alpha}^1 \in I_{max}} \cdots \sum_{\boldsymbol{\alpha}^i \in I_{max}} |\!|\!| B_{\boldsymbol{\alpha}^1} |\!|\!|_2 \cdots |\!|\!| B_{\boldsymbol{\alpha}^i} |\!|\!|_2 \left\| \sum_{j=0}^{k-1} \varepsilon_j (\widetilde{\mathbf{Z}}_0^{(j)})^{\boldsymbol{\alpha}^1} \cdots (\widetilde{\mathbf{Z}}_{-i+1}^{(j)})^{\boldsymbol{\alpha}^i} q(\widetilde{\mathbf{Z}}_{-i}^{(j)}) \right\|_2 \\
&\quad \leq \sum_{i=0}^{\infty} \sum_{\boldsymbol{\alpha}^1 \in I_{max}} \cdots \sum_{\boldsymbol{\alpha}^i \in I_{max}} |\!|\!| B_{\boldsymbol{\alpha}^1} |\!|\!|_2 \cdots |\!|\!| B_{\boldsymbol{\alpha}^i} |\!|\!|_2 \sum_{\boldsymbol{\alpha} \in I_{max}} \|C_{\boldsymbol{\alpha}}\|_2 \left| \sum_{j=0}^{k-1} \varepsilon_j (\widetilde{\mathbf{Z}}_0^{(j)})^{\boldsymbol{\alpha}^1} \cdots (\widetilde{\mathbf{Z}}_{-i+1}^{(j)})^{\boldsymbol{\alpha}^i} (\widetilde{\mathbf{Z}}_{-i}^{(j)})^{\boldsymbol{\alpha}} \right|.
\end{aligned}
$$

Therefore, by this expression and using the same type of arguments for linear readout as in the example of echo state networks one obtains

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{H \in \mathcal{H}^{RC}} \left\| \sum_{j=0}^{k-1} \varepsilon_j H(\widetilde{\mathbf{Z}}^j) \right\|_2 \right] \\
&\quad = \mathbb{E}\left[\sup_{\substack{F \in \mathcal{F}^{RC} \\ W : |\!|\!| W |\!|\!|_2 \leq \overline{L_h} \\ \boldsymbol{a} : \|\boldsymbol{a}\|_2 \leq L_{h,0}}} \left\| \sum_{j=0}^{k-1} \varepsilon_j (W H^F(\widetilde{\mathbf{Z}}^{(j)}) + \boldsymbol{a}) \right\|_2 \right] \\
&\quad \leq \overline{L_h} \sum_{i=0}^{\infty} \sum_{\boldsymbol{\alpha}^1 \in I_{max}} \cdots \sum_{\boldsymbol{\alpha}^i \in I_{max}} (\lambda^{SAS})^i c^{SAS} \sum_{\boldsymbol{\alpha} \in I_{max}} \mathbb{E}\left[\left| \sum_{j=0}^{k-1} \varepsilon_j (\widetilde{\mathbf{Z}}_0^{(j)})^{\boldsymbol{\alpha}^1} \cdots (\widetilde{\mathbf{Z}}_{-i+1}^{(j)})^{\boldsymbol{\alpha}^i} (\widetilde{\mathbf{Z}}_{-i}^{(j)})^{\boldsymbol{\alpha}} \right| \right] + \sqrt{k} L_{h,0} \\
&\quad \leq \overline{L_h} \sum_{i=0}^{\infty} \sum_{\boldsymbol{\alpha}^1 \in I_{max}} \cdots \sum_{\boldsymbol{\alpha}^i \in I_{max}} (\lambda^{SAS})^i c^{SAS} \sum_{\boldsymbol{\alpha} \in I_{max}} \mathbb{E}\left[\left| \sum_{j=0}^{k-1} \varepsilon_j (\widetilde{\mathbf{Z}}_0^{(j)})^{\boldsymbol{\alpha}^1} \cdots (\widetilde{\mathbf{Z}}_{-i+1}^{(j)})^{\boldsymbol{\alpha}^i} (\widetilde{\mathbf{Z}}_{-i}^{(j)})^{\boldsymbol{\alpha}} \right|^2 \right]^{1/2} + \sqrt{k} L_{h,0} \\
&\quad = \sqrt{k} \overline{L_h} \sum_{i=0}^{\infty} \sum_{\boldsymbol{\alpha}^1 \in I_{max}} \cdots \sum_{\boldsymbol{\alpha}^i \in I_{max}} (\lambda^{SAS})^i c^{SAS} \sum_{\boldsymbol{\alpha} \in I_{max}} \mathbb{E}\left[\left| (\mathbf{Z}_0)^{\boldsymbol{\alpha}^1} \cdots (\mathbf{Z}_{-i+1})^{\boldsymbol{\alpha}^i} (\mathbf{Z}_{-i})^{\boldsymbol{\alpha}} \right|^2 \right]^{1/2} + \sqrt{k} L_{h,0} \\
&\quad \leq \sqrt{k} \left(\overline{L_h} \frac{c^{SAS} |I_{max}|}{1 - |I_{max}| \lambda^{SAS}} + L_{h,0} \right). \quad \blacksquare
\end{aligned}
$$

5.9 Proof of Theorem 14

We start by working out two concentration inequalities that are needed in the proof. These are contained in the two propositions 19 and 20 and are used in part **(i)** of the theorem in relation with the use of Assumption 1.

5.9.1 (Exponential) concentration inequalities

Proposition 19 *Define* $\Gamma_n := \sup_{H\in\mathcal{H}^{RC}}\{R(H)-\widehat{R}_n^{\infty}(H)\}$. *Suppose that Assumptions 4-6 hold, that Assumption 1 is satisfied, and that the Bernoulli shifts innovations are bounded, that is, there exists* $\overline{M} > 0$ *such that for* $I = y, z$ *and for all* $t \in \mathbb{Z}_-$, $\|\boldsymbol{\xi}_t^I\|_2 \leq \overline{M}$. *Then there exists* $C_{bd} > 0$ *such that for any* $\eta > 0$, $n \in \mathbb{N}^+$ *it holds that*

$$\mathbb{P}\left(|\Gamma_n - \mathbb{E}[\Gamma_n]| \geq \eta\right) \leq 2\exp\left(-\frac{2n\eta^2}{C_{bd}^2}\right). \tag{92}$$

The constant C_{bd} *is explicitly given by*

$$C_{bd} = 2L_L\left(\frac{\overline{L_h}}{1-r}\left(M_{\mathcal{F}}r + L_R\overline{M}L_z\|w^z\|_1\right) + \overline{M}L_y\|w^y\|_1\right). \tag{93}$$

Proof. The main idea of the proof is to exploit the Bernoulli shift structure and apply McDiarmid's inequality (Boucheron et al., 2013, see for example), which out of the bound of differences of functions constructed in a particular manner yields the bound in (92). In order to ease the notation, we first define $\mathcal{Y} := (\overline{B_{\overline{M}}} \times \overline{B_{\overline{M}}}) \subset (\mathbb{R}^{q_y} \times \mathbb{R}^{q_z})$. Consider now a function $\phi\colon \mathcal{Y}^{n-1} \times \mathcal{Y}^{\mathbb{Z}_-} \to \mathbb{R}$, which is defined for $\boldsymbol{u}_i = (\boldsymbol{u}_{-i}^y, \boldsymbol{u}_{-i}^z) \in \mathcal{Y}$, $i = 0, \ldots, n-2$ and $\boldsymbol{u}_{n-1} = (\boldsymbol{u}_{-n+1+t}^y, \boldsymbol{u}_{-n+1+t}^z)_{t\in\mathbb{Z}_-} \in \mathcal{Y}^{\mathbb{Z}_-}$ by

$$\phi(\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{n-2}, \boldsymbol{u}_{n-1}) = \sup_{H\in\mathcal{H}^{RC}}\left\{R(H) - \frac{1}{n}\sum_{i=0}^{n-1} L(H(G^z(\boldsymbol{u}_{-i}^{z,-\infty})), G^y(\boldsymbol{u}_{-i}^{y,-\infty}))\right\}.$$

Fix $k \in \{0, \ldots, n-1\}$ and let $\widetilde{\boldsymbol{u}}$ be an identical copy of the sequence $\boldsymbol{u}$ so that only the k-th entry in $\widetilde{\boldsymbol{u}}$ is different from $\boldsymbol{u}$, that is $\widetilde{\boldsymbol{u}}_{-i} = \boldsymbol{u}_{-i}$ for all $i \neq k$. We now estimate the difference of the function $\phi\colon \mathcal{Y}^{n-1} \times \mathcal{Y}^{\mathbb{Z}_-} \to \mathbb{R}$ evaluated at $\boldsymbol{u}$ and $\widetilde{\boldsymbol{u}}$ as follows:

$$\begin{aligned}
&\phi(\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{n-1}) - \phi(\widetilde{\boldsymbol{u}}_0, \ldots, \widetilde{\boldsymbol{u}}_{n-1}) \\
&\leq \sup_{H\in\mathcal{H}^{RC}} \inf_{\widetilde{H}\in\mathcal{H}^{RC}} \frac{1}{n}\sum_{i=0}^{n-1}\left\{L(\widetilde{H}(G^z(\widetilde{\boldsymbol{u}}_{-i}^{z,-\infty})), G^y(\widetilde{\boldsymbol{u}}_{-i}^{y,-\infty})) - L(H(G^z(\boldsymbol{u}_{-i}^{z,-\infty})), G^y(\boldsymbol{u}_{-i}^{y,-\infty}))\right\} \\
&\qquad - R(\widetilde{H}) + R(H) \\
&\leq \sup_{H\in\mathcal{H}^{RC}} \frac{1}{n}\sum_{i=0}^{n-1}\left\{L(H(G^z(\widetilde{\boldsymbol{u}}_{-i}^{z,-\infty})), G^y(\widetilde{\boldsymbol{u}}_{-i}^{y,-\infty})) - L(H(G^z(\boldsymbol{u}_{-i}^{z,-\infty})), G^y(\boldsymbol{u}_{-i}^{y,-\infty}))\right\} \\
&= \sup_{H\in\mathcal{H}^{RC}} \frac{1}{n}\sum_{i=0}^{k}\left\{L(H(G^z(\widetilde{\boldsymbol{u}}_{-i}^{z,-\infty})), G^y(\widetilde{\boldsymbol{u}}_{-i}^{y,-\infty})) - L(H(G^z(\boldsymbol{u}_{-i}^{z,-\infty})), G^y(\boldsymbol{u}_{-i}^{y,-\infty}))\right\} \\
&\leq \sup_{H\in\mathcal{H}^{RC}} \frac{1}{n}\sum_{i=0}^{k}\left\{L_L(\|H(G^z(\widetilde{\boldsymbol{u}}_{-i}^{z,-\infty})) - H(G^z(\boldsymbol{u}_{-i}^{z,-\infty}))\|_2 + \|G^y(\widetilde{\boldsymbol{u}}_{-i}^{y,-\infty}) - G^y(\boldsymbol{u}_{-i}^{y,-\infty})\|_2)\right\},
\end{aligned} \tag{94}$$

where in the last inequality we used that by assumption the loss function L is L_L-Lipschitz. For the first summand under the supremum in (94) we use the bound (75) and hence write

$$\begin{aligned}
&\sum_{i=0}^{k}\|H(G^z(\widetilde{\boldsymbol{u}}_{-i}^{z,-\infty}))-H(G^z(\boldsymbol{u}_{-i}^{z,-\infty}))\|_2\\
&\quad\le\sum_{i=0}^{k}\overline{L_h}(2r^{k+1-i}M_{\mathcal{F}}+L_R\sum_{j=0}^{k-i}r^j\|G^z(\widetilde{\boldsymbol{u}}_{-j-i}^{z,-\infty})-G^z(\boldsymbol{u}_{-j-i}^{z,-\infty})\|_2)\\
&\quad\le\overline{L_h}\left(2M_{\mathcal{F}}r\frac{1-r^{k+1}}{1-r}+L_R\sum_{i=0}^{k}\sum_{j=i}^{k}r^{j-i}\|G^z(\widetilde{\boldsymbol{u}}_{-j}^{z,-\infty})-G^z(\boldsymbol{u}_{-j}^{z,-\infty})\|_2)\right). \qquad (95)
\end{aligned}$$

Notice that the second summand under the supremum in (94) and the second summand in (95) can be bounded using that (24) holds by Assumption 1 and that by hypothesis both $\widetilde{\boldsymbol{u}}$ and $\boldsymbol{u}$ satisfy $\|\boldsymbol{u}_t^I\|_2\le\overline{M}$ and $\|\widetilde{\boldsymbol{u}}_t^I\|_2\le\overline{M}$ for any $t\in\mathbb{Z}_-$. More specifically, for $I=y,z$ and any $i\in\{0,\dots,k\}$ one obtains

$$\begin{aligned}
\|G^I(\widetilde{\boldsymbol{u}}_{-i}^{I,-\infty})-G^I(\boldsymbol{u}_{-i}^{I,-\infty})\|_2&\le L_I\sum_{l=0}^{\infty}w_l^I\|\boldsymbol{u}_{-l-i}^I-\widetilde{\boldsymbol{u}}_{-l-i}^I\|_2\\
&=L_Iw_{k-i}^I\|\boldsymbol{u}_{-k}^I-\widetilde{\boldsymbol{u}}_{-k}^I\|_2\le 2L_Iw_{k-i}^I\overline{M},
\end{aligned}$$

where we used that $\widetilde{\boldsymbol{u}}_{-l-i}=\boldsymbol{u}_{-l-i}$, for all $l+i\neq k$, $l\in\mathbb{N}$. Combining this expression with (95) we estimate (94) as

$$\begin{aligned}
&\phi(\boldsymbol{u}_0,\dots,\boldsymbol{u}_{n-1})-\phi(\widetilde{\boldsymbol{u}}_0,\dots,\widetilde{\boldsymbol{u}}_{n-1})\\
&\le\sup_{H\in\mathcal{H}^{RC}}\frac{1}{n}\sum_{i=0}^{k}\left\{L_L(\|H(G^z(\widetilde{\boldsymbol{u}}_{-i}^{z,-\infty}))-H(G^z(\boldsymbol{u}_{-i}^{z,-\infty}))\|_2+\|G^y(\widetilde{\boldsymbol{u}}_{-i}^{y,-\infty})-G^y(\boldsymbol{u}_{-i}^{y,-\infty})\|_2)\right\}\\
&\le\frac{2}{n}L_L\left(\overline{L_h}M_{\mathcal{F}}r\frac{1-r^{k+1}}{1-r}+\overline{L_h}L_R\overline{M}L_z\sum_{i=0}^{k}\sum_{j=i}^{k}r^{j-i}w_{k-j}^z+\overline{M}L_y\sum_{i=0}^{k}w_{k-i}^y\right)\\
&=\frac{2}{n}L_L\left(\overline{L_h}M_{\mathcal{F}}r\frac{1-r^{k+1}}{1-r}+\overline{L_h}L_R\overline{M}L_z\sum_{i=0}^{k}\sum_{j=0}^{k-i}r^{j}w_{k-i-j}^z+\overline{M}L_y\sum_{i=0}^{k}w_{i}^y\right)\\
&\le\frac{2}{n}L_L\left(\overline{L_h}M_{\mathcal{F}}\frac{r}{1-r}+\overline{L_h}L_R\overline{M}L_z\left(\sum_{i=0}^{\infty}r^i\right)\left(\sum_{j=0}^{\infty}w_j^z\right)+\overline{M}L_y\sum_{i=0}^{\infty}w_i^y\right)\\
&=\frac{2}{n}L_L\left(\frac{\overline{L_h}}{1-r}\left(M_{\mathcal{F}}r+L_R\overline{M}L_z\|w^z\|_1\right)+\overline{M}L_y\|w^y\|_1\right)=\frac{C_{bd}}{n},
\end{aligned}$$

with the constant C_{bd} as in (93). We now use this bound of differences in McDiarmid's inequality and simply notice that $\Gamma_n=\phi(\boldsymbol{\xi}_0,\dots,\boldsymbol{\xi}_{-n+2},\boldsymbol{\xi}_{-n+1}^{-\infty})$ in the statement, which immediately yields (92), as required. ■

Proposition 20 *Define* $\Gamma_n := \sup_{H\in\mathcal{H}^{RC}}\{R(H) - \widehat{R}_n^{\infty}(H)\}$. *Suppose that Assumptions 4-6 hold and that Assumption 1 is satisfied. Let* $\Phi\colon [0,\infty)\to[0,\infty)$ *be a convex and increasing function that satisfies* $\Phi(0)=0$. *Furthermore, assume that* $\max_{I\in\{y,z\}}\mathbb{E}\left[\Phi(2C_I^{mom}\|\boldsymbol{\xi}_0^I\|_2)\right] < \infty$ *where*

$$C_z^{mom} = L_L\frac{\overline{L_h}L_R}{1-r}L_z\|w^z\|_1, \tag{96}$$

$$C_y^{mom} = L_L L_y\|w^y\|_1, \tag{97}$$

and denote

$$\varphi(\overline{M}) := \sum_{I\in\{y,z\}}\mathbb{E}\left[\Phi(2C_I^{mom}\|\boldsymbol{\xi}_0^I\|_2)\mathbf{1}_{\{\|\boldsymbol{\xi}_0^I\|_2>\overline{M}\}}\right]. \tag{98}$$

Then, there exists a constant $C_0>0$ *such that for any* $\eta>0$, $n\in\mathbb{N}^+$, $\overline{M}>0$ *satisfying*

$$\sum_{I\in\{y,z\}} C_I^{mom}\mathbb{E}[\|\boldsymbol{\xi}_0^I\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_0^I\|_2>\overline{M}\}}] < \frac{\eta}{9} \tag{99}$$

one has

$$\mathbb{P}\left(|\Gamma_n - \mathbb{E}[\Gamma_n]|\geq\eta\right) \leq 2\exp\left(\frac{-2n\eta^2}{9(C_0+2\overline{M}(C_z^{mom}+C_y^{mom}))^2}\right) + \frac{1}{2}\frac{\varphi(\overline{M})}{\Phi(\eta/3)}.$$

The constant C_0 *is explicitly given by* (39).

Proof. Let $\overline{M}>0$ and for any $t\in\mathbb{Z}_-$, $I=y,z$ denote by $\boldsymbol{\xi}_t^{I,\overline{M}}$ the Bernoulli shift innovations whose Euclidean norm is bounded above by $\overline{M}$, that is, $\boldsymbol{\xi}_t^{I,\overline{M}} := \boldsymbol{\xi}_t^I\mathbf{1}_{\{\|\boldsymbol{\xi}_t^I\|_2\leq\overline{M}\}}$. In order to simplify the notation, we define

$$\mathbf{Z}_t^{\overline{M}} := G^z(\ldots,\boldsymbol{\xi}_{t-1}^{z,\overline{M}},\boldsymbol{\xi}_t^{z,\overline{M}}) = G^z((\boldsymbol{\xi}^{z,\overline{M}})_t^{-\infty}), \quad t\in\mathbb{Z}_-, \tag{100}$$

$$\mathbf{Y}_t^{\overline{M}} := G^y(\ldots,\boldsymbol{\xi}_{t-1}^{y,\overline{M}},\boldsymbol{\xi}_t^{y,\overline{M}}) = G^y((\boldsymbol{\xi}^{y,\overline{M}})_t^{-\infty}), \quad t\in\mathbb{Z}_-, \tag{101}$$

$$\widehat{R}_n^{\infty,\overline{M}}(H) := \frac{1}{n}\sum_{i=0}^{n-1} L(H((\mathbf{Z}^{\overline{M}})_{-i}^{-\infty}),\mathbf{Y}_{-i}^{\overline{M}}), \tag{102}$$

$$R^{\overline{M}}(H) := \mathbb{E}[L(H(\mathbf{Z}^{\overline{M}}),\mathbf{Y}_0^{\overline{M}})] \tag{103}$$

and, additionally, denote $\Gamma_n^{\overline{M}} := \sup_{H\in\mathcal{H}^{RC}}\{R^{\overline{M}}(H) - \widehat{R}_n^{\infty,\overline{M}}(H)\}$. Firstly, the triangle inequality yields

$$\begin{aligned}|\Gamma_n - \mathbb{E}[\Gamma_n]| &= |\Gamma_n - \Gamma_n^{\overline{M}} - (\mathbb{E}[\Gamma_n] - \mathbb{E}[\Gamma_n^{\overline{M}}]) + \Gamma_n^{\overline{M}} - \mathbb{E}[\Gamma_n^{\overline{M}}]|\\ &\leq |\Gamma_n - \Gamma_n^{\overline{M}}| + |\mathbb{E}[\Gamma_n - \Gamma_n^{\overline{M}}]| + |\Gamma_n^{\overline{M}} - \mathbb{E}[\Gamma_n^{\overline{M}}]|\\ &\leq |\Gamma_n - \Gamma_n^{\overline{M}}| + \mathbb{E}[|\Gamma_n - \Gamma_n^{\overline{M}}|] + |\Gamma_n^{\overline{M}} - \mathbb{E}[\Gamma_n^{\overline{M}}]|.\end{aligned} \tag{104}$$

For the first summand in expression (104) we write

$$\left|\Gamma_n - \Gamma_n^{\overline{M}}\right| \leq \left|\sup_{H\in\mathcal{H}^{RC}}\left\{R(H) - \widehat{R}_n^{\infty}(H)\right\} - \sup_{H\in\mathcal{H}^{RC}}\left\{R^{\overline{M}}(H) - \widehat{R}_n^{\infty,\overline{M}}(H)\right\}\right|$$

$$
\begin{aligned}
&= \left| \sup_{H\in\mathcal{H}^{RC}} \inf_{\widetilde{H}\in\mathcal{H}^{RC}} \left(\left\{ R(H) - \widehat{R}_n^{\infty}(H) \right\} - \left\{ R^{\overline{M}}(\widetilde{H}) - \widehat{R}_n^{\infty,\overline{M}}(\widetilde{H}) \right\} \right) \right| \\
&\leq \left| \sup_{H\in\mathcal{H}^{RC}} \left\{ R(H) - \widehat{R}_n^{\infty}(H) - R^{\overline{M}}(H) + \widehat{R}_n^{\infty,\overline{M}}(H) \right\} \right| \\
&\leq \sup_{H\in\mathcal{H}^{RC}} \left| R^{\overline{M}}(H) - R(H) \right| + \sup_{H\in\mathcal{H}^{RC}} \left| \widehat{R}_n^{\infty,\overline{M}}(H) - \widehat{R}_n^{\infty}(H) \right|. \qquad (105)
\end{aligned}
$$

Using this result, we can immediately get the following bound for the second summand in expression (104)

$$
\mathbb{E}\left[\left| \Gamma_n - \Gamma_n^{\overline{M}} \right| \right] \leq \sup_{H\in\mathcal{H}^{RC}} \left| R^{\overline{M}}(H) - R(H) \right| + \mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}} \left| \widehat{R}_n^{\infty,\overline{M}}(H) - \widehat{R}_n^{\infty}(H) \right| \right]. \qquad (106)
$$

The first two terms in the right hand side of (104) are thus controlled by (105) and (106). The third term in (104) is of the type required in Proposition 19, that is, the Bernoulli shifts are defined using bounded innovations and hence the term will be controlled in what follows using the result in Proposition 19.

The next step in our proof is to derive bounds for the terms in the right hand sides of (105) and (106). First, we consider the estimate for the term they share, for which we write

$$
\begin{aligned}
&\sup_{H\in\mathcal{H}^{RC}} |R^{\overline{M}}(H) - R(H)| \\
&\quad = \sup_{H\in\mathcal{H}^{RC}} |\mathbb{E}[L(H(\mathbf{Z}^{\overline{M}}), \mathbf{Y}_0^{\overline{M}}) - L(H(\mathbf{Z}), \mathbf{Y}_0)]| \\
&\quad \leq \sup_{H\in\mathcal{H}^{RC}} \mathbb{E}\left[L_L(\|H(\mathbf{Z}^{\overline{M}}) - H(\mathbf{Z})\|_2 + \|\mathbf{Y}_0^{\overline{M}} - \mathbf{Y}_0\|_2) \right] \\
&\quad \leq L_L(\overline{L_h} L_R \sum_{j=0}^{\infty} r^j \mathbb{E}[\|\mathbf{Z}_{-j}^{\overline{M}} - \mathbf{Z}_{-j}\|_2] + \mathbb{E}[\|\mathbf{Y}_0^{\overline{M}} - \mathbf{Y}_0\|_2]) \\
&\quad = L_L \left(\frac{\overline{L_h} L_R}{1-r} \mathbb{E}[\|G^z(\boldsymbol{\xi}^{z,\overline{M}}) - G^z(\boldsymbol{\xi}^z)\|_2] + \mathbb{E}[\|G^y(\boldsymbol{\xi}^y) - G^y(\boldsymbol{\xi}^{y,\overline{M}})\|_2] \right), \qquad (107)
\end{aligned}
$$

where the first and the second (in)equality are obtained using the definition (103) and the assumption that the loss function L is L_L-Lipschitz, the third step follows from the estimate (75) from Lemma 18 and the last step again uses (100)-(101) and the i.i.d. assumption on $\boldsymbol{\xi}$. In order to proceed, notice that since Assumption 1 holds by hypothesis, by (24) one has for any $j \in \mathbb{N}$, $I = y, z$ the following estimate

$$
\begin{aligned}
\|G^I((\boldsymbol{\xi}^I)_{-j}^{-\infty}) - G^I((\boldsymbol{\xi}^{I,\overline{M}})_{-j}^{-\infty})\|_2 &\leq L_I \sum_{l=0}^{\infty} w_l^I \|\boldsymbol{\xi}_{-l-j}^I - \boldsymbol{\xi}_{-l-j}^{I,\overline{M}}\|_2 \\
&= L_I \sum_{l=0}^{\infty} w_l^I \|\boldsymbol{\xi}_{-l-j}^I\|_2 \mathbf{1}_{\{\|\boldsymbol{\xi}_{-l-j}^I\|_2 > \overline{M}\}}. \qquad (108)
\end{aligned}
$$

Combining (107) and (108) and using the i.i.d. assumption on $\boldsymbol{\xi}$ one obtains

$$
\begin{aligned}
&\sup_{H\in\mathcal{H}^{RC}} |R^{\overline{M}}(H) - R(H)| \\
&\quad\le L_L\left(\frac{\overline{L_h}L_R}{1-r}L_z\mathbb{E}[\sum_{l=0}^{\infty} w_l^z\|\boldsymbol{\xi}_{-l}^z\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_{-l}^z\|_2>\overline{M}\}}] + L_y\mathbb{E}[\sum_{l=0}^{\infty} w_l^y\|\boldsymbol{\xi}_{-l}^y\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_{-l}^y\|_2>\overline{M}\}}]\right) \\
&\quad= L_L\left(\frac{\overline{L_h}L_R}{1-r}L_z\|w^z\|_1\mathbb{E}[\|\boldsymbol{\xi}_0^z\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_0^z\|_2>\overline{M}\}}] + L_y\|w^y\|_1\mathbb{E}[\|\boldsymbol{\xi}_0^y\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_0^y\|_2>\overline{M}\}}]\right) \\
&\quad\le C_z^{mom}\mathbb{E}[\|\boldsymbol{\xi}_0^z\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_0^z\|_2>\overline{M}\}}] + C_y^{mom}\mathbb{E}[\|\boldsymbol{\xi}_0^y\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_0^y\|_2>\overline{M}\}}] \\
&\quad= \sum_{I\in\{y,z\}} C_I^{mom}\mathbb{E}[\|\boldsymbol{\xi}_0^I\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_0^I\|_2>\overline{M}\}}] \qquad (109)
\end{aligned}
$$

with C_I^{mom} for $I = y, z$ defined as in (96) and (97).

Next, we analyze the second term in (105). Using the function $\Phi : [0,\infty) \longrightarrow [0,\infty)$ we obtain for $\eta > 0$

$$
\begin{aligned}
&\Phi(\eta)\mathbb{P}\left(\sup_{H\in\mathcal{H}^{RC}}\left|\widehat{R}_n^{\infty,\overline{M}}(H) - \widehat{R}_n^{\infty}(H)\right| \ge \eta\right) \\
&\le \mathbb{E}\left[\Phi\left(\sup_{H\in\mathcal{H}^{RC}}\left|\widehat{R}_n^{\infty,\overline{M}}(H) - \widehat{R}_n^{\infty}(H)\right|\right)\right] \\
&\le \mathbb{E}\Big[\Phi\Big(\frac{L_L}{n}\sum_{i=0}^{n-1}\Big(\overline{L_h}L_R\sum_{j=0}^{\infty} r^j\|G^z((\boldsymbol{\xi}^z)_{-j-i}^{-\infty}) - G^z((\boldsymbol{\xi}^{z,\overline{M}})_{-j-i}^{-\infty})\|_2 \\
&\qquad + \|G^y((\boldsymbol{\xi}^y)_{-i}^{-\infty}) - G^y((\boldsymbol{\xi}^{y,\overline{M}})_{-i}^{-\infty})\|_2\Big)\Big)\Big] \\
&\le \frac{1-r}{n}\sum_{i=0}^{n-1}\sum_{j=0}^{\infty} r^j\mathbb{E}\Big[\Phi\Big(\frac{L_L\overline{L_h}L_R}{1-r}\|G^z((\boldsymbol{\xi}^z)_{-j-i}^{-\infty}) - G^z((\boldsymbol{\xi}^{z,\overline{M}})_{-j-i}^{-\infty})\|_2 \\
&\qquad + L_L\|G^y((\boldsymbol{\xi}^y)_{-i}^{-\infty}) - G^y((\boldsymbol{\xi}^{y,\overline{M}})_{-i}^{-\infty})\|_2\Big)\Big] \\
&\le \frac{1}{2}\mathbb{E}\Big[\Phi\Big(\frac{2L_L\overline{L_h}L_R}{1-r}\|G^z(\boldsymbol{\xi}^z) - G^z(\boldsymbol{\xi}^{z,\overline{M}})\|_2\Big)\Big] + \frac{1}{2}\mathbb{E}\Big[\Phi\Big(2L_L\|G^y(\boldsymbol{\xi}^y) - G^y(\boldsymbol{\xi}^{y,\overline{M}})\|_2\Big)\Big] \\
&\le \frac{1}{2}\mathbb{E}\Big[\Phi\Big(\frac{2L_L\overline{L_h}L_R}{1-r}L_z\sum_{l=0}^{\infty} w_l^z\|\boldsymbol{\xi}_{-l}^z - \boldsymbol{\xi}_{-l}^{z,\overline{M}}\|_2\Big)\Big] + \frac{1}{2}\mathbb{E}\Big[\Phi\Big(2L_LL_y\sum_{l=0}^{\infty} w_l^y\|\boldsymbol{\xi}_{-l}^y - \boldsymbol{\xi}_{-l}^{y,\overline{M}}\|_2\Big)\Big] \\
&= \frac{1}{2}\sum_{I\in\{y,z\}}\mathbb{E}\Big[\Phi\Big(\frac{2C_I^{mom}}{\|w^I\|_1}\sum_{l=0}^{\infty} w_l^I\|\boldsymbol{\xi}_{-l}^I - \boldsymbol{\xi}_{-l}^{I,\overline{M}}\|_2\Big)\Big] \\
&\le \frac{1}{2}\sum_{I\in\{y,z\}}\frac{1}{\|w^I\|_1}\sum_{l=0}^{\infty} w_l^I\mathbb{E}\Big[\Phi\Big(2C_I^{mom}\|\boldsymbol{\xi}_{-l}^I\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_{-l}^I\|_2>\overline{M}\}}\Big)\Big] \\
&= \frac{1}{2}\varphi(\overline{M}), \qquad (110)
\end{aligned}
$$

where C_z^{mom}, C_y^{mom}, and $\varphi(\overline{M})$ are given in (96), (97), and (98), respectively. In these derivations we used Markov's inequality for increasing and non-negative functions in the first

step. The second inequality uses the definition in (102), the estimate (75) from Lemma 18, and the fact that Φ is by hypothesis an increasing function. We subsequently used Jensen's inequality for discrete probability measures, the stationarity assumption and the convexity of the function Φ. Finally, (108), the appropriate choice of constants and once more Jensen's inequality for discrete probability measures and the stationarity assumption as well as $\Phi(0) = 0$ yields the result.

We now notice that this result provides automatically a bound for the second term in (106). In order to see that, one needs to take as the function Φ the identity and then by the second and the last two lines in (110) it holds that

$$\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left|\widehat{R}_n^{\infty,\overline{M}}(H)-\widehat{R}_n^{\infty}(H)\right|\right]\leq\sum_{I\in\{y,z\}}C_I^{mom}\mathbb{E}[\|\boldsymbol{\xi}_0^I\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_0^I\|_2>\overline{M}\}}]. \tag{111}$$

We now consider again expression (104) and taking into account (105), (106), together with the bounds for their ingredients given in (109), (110), and (111) we derive, for any $\eta > 0$ satisfying (99),

$$\begin{aligned}
&\mathbb{P}\left(|\Gamma_n-\mathbb{E}[\Gamma_n]|\geq\eta\right)\\
&\quad\leq\mathbb{P}\left(\frac{2\eta}{9}+\sup_{H\in\mathcal{H}^{RC}}\left|\widehat{R}_n^{\infty,\overline{M}}(H)-\widehat{R}_n^{\infty}(H)\right|+\frac{\eta}{9}+|\Gamma_n^{\overline{M}}-\mathbb{E}[\Gamma_n^{\overline{M}}]|\geq\eta\right)\\
&\quad\leq\mathbb{P}\left(|\Gamma_n^{\overline{M}}-\mathbb{E}[\Gamma_n^{\overline{M}}]|\geq\frac{\eta}{3}\right)+\mathbb{P}\left(\sup_{H\in\mathcal{H}^{RC}}\left|\widehat{R}_n^{\infty,\overline{M}}(H)-\widehat{R}_n^{\infty}(H)\right|\geq\frac{\eta}{3}\right)\\
&\quad\leq 2\exp\left(\frac{-2n\eta^2}{9C_{bd}^2}\right)+\frac{1}{2}\frac{\varphi(\overline{M})}{\Phi(\eta/3)},
\end{aligned}$$

with C_{bd} as in (93). In the first inequality we used the hypothesis in (99) and the bounds (109), (111). In the second inequality we used (110) and (92) in Proposition 19. Finally, noticing that $C_{bd} = 2\overline{M}(C_z^{mom} + C_y^{mom}) + C_0$ with C_0 as in (39) and setting C_z^{mom} and C_y^{mom} as in (96)-(97) immediately yields the claim. ■

Corollary 21 *Suppose that Assumptions 4-6 hold and that Assumption 1 is satisfied. Let $\Phi\colon[0,\infty)\to[0,\infty)$ be a convex and strictly increasing function that satisfies $\Phi(0) = 0$. Furthermore, assume that for all $u > 0$, $\mathbb{E}[\Phi(u\|\boldsymbol{\xi}_0^I\|)^2] < \infty$ for $I = y, z$. Then, for any $\delta \in (0,1)$, $n \in \mathbb{N}^+$,*

$$\mathbb{P}\left(|\Gamma_n-\mathbb{E}[\Gamma_n]|\geq B_\Phi(n,\delta)\right)\leq\frac{\delta}{2},$$

where

$$B_\Phi(n,\delta)=9\max\left(\frac{(C_0+2\Phi^{-1}(n)(C_z^{mom}+C_y^{mom}))\sqrt{\log(\frac{8}{\delta})}}{3\sqrt{2n}},\Phi^{-1}\left(\frac{2C_\Phi}{\delta\sqrt{n}}\right)\right), \tag{112}$$

$$C_\Phi=\sum_{I\in\{y,z\}}\mathbb{E}\left[\Phi(2C_I^{mom}\|\boldsymbol{\xi}_0^I\|_2)^2\right]^{1/2}\mathbb{E}[\Phi(\|\boldsymbol{\xi}_0^I\|_2)]^{1/2}, \tag{113}$$

and C_0, C_z^{mom}, C_y^{mom} are given by (39), (96), *and* (97).

Proof. We start with the function $\varphi(\overline{M})$ given in (98) and obtain that for any $\overline{M} > 0$ it holds that

$$\begin{aligned}
\varphi(\overline{M}) &= \sum_{I\in\{y,z\}} \mathbb{E}\left[\Phi(2C_I^{mom}\|\boldsymbol{\xi}_0^I\|_2)\mathbf{1}_{\{\|\boldsymbol{\xi}_0^I\|_2>\overline{M}\}}\right] \\
&\leq \sum_{I\in\{y,z\}} \mathbb{E}\left[\Phi(2C_I^{mom}\|\boldsymbol{\xi}_0^I\|_2)^2\right]^{1/2} \mathbb{E}\left[(\mathbf{1}_{\{\|\boldsymbol{\xi}_0^I\|_2>\overline{M}\}})^2\right]^{1/2} \\
&= \sum_{I\in\{y,z\}} \mathbb{E}\left[\Phi(2C_I^{mom}\|\boldsymbol{\xi}_0^I\|_2)^2\right]^{1/2} \mathbb{P}(\|\boldsymbol{\xi}_0^I\|_2 > \overline{M})^{1/2} \\
&\leq \frac{C_\Phi}{\Phi(\overline{M})^{1/2}},
\end{aligned} \tag{114}$$

where the first inequality is a consequence of Hölder's inequality, and the last step is obtained by applying Markov's inequality for increasing nonnegative functions and by using the definition in (113). Furthermore, by convexity, Jensen's inequality and since $\Phi(0) = 0$, one has that

$$\begin{aligned}
\Phi\Big(\sum_{I\in\{y,z\}} C_I^{mom}\mathbb{E}[\|\boldsymbol{\xi}_0^I\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_0^I\|_2>\overline{M}\}}]\Big) &\leq \sum_{I\in\{y,z\}} \frac{1}{2}\Phi\left(2C_I^{mom}\mathbb{E}[\|\boldsymbol{\xi}_0^I\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_0^I\|_2>\overline{M}\}}]\right) \\
&\leq \frac{1}{2}\sum_{I\in\{y,z\}} \mathbb{E}\left[\Phi\left(2C_I^{mom}\|\boldsymbol{\xi}_0^I\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_0^I\|_2>\overline{M}\}}\right)\right] \\
&= \frac{1}{2}\varphi(\overline{M}).
\end{aligned} \tag{115}$$

Choosing $\overline{M} = \Phi^{-1}(n)$ in (114) and setting $\eta = B_\Phi(n,\delta)$ defined in (112) one easily verifies that

$$\frac{1}{2}\frac{\varphi(\overline{M})}{\Phi(\eta/9)} \leq \frac{1}{2}\frac{C_\Phi}{\sqrt{n}\Phi(\eta/9)} \leq \frac{\delta}{4}. \tag{116}$$

In particular, this also implies that $\varphi(\overline{M}) < \Phi(\eta/9)$ and so (115) yields

$$\Phi\Big(\sum_{I\in\{y,z\}} C_I^{mom}\mathbb{E}[\|\boldsymbol{\xi}_0^I\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_0^I\|_2>\overline{M}\}}]\Big) < \Phi(\eta/9)$$

and hence

$$\sum_{I\in\{y,z\}} C_I^{mom}\mathbb{E}[\|\boldsymbol{\xi}_0^I\|_2\mathbf{1}_{\{\|\boldsymbol{\xi}_0^I\|_2>\overline{M}\}}] < \frac{\eta}{9}.$$

Thus (99) is satisfied and we may apply Proposition 20 and use $\Phi(\eta/9) \leq \Phi(\eta/3)$ and (116), which yields

$$\mathbb{P}\left(|\Gamma_n - \mathbb{E}[\Gamma_n]| \geq \eta\right) \leq 2\exp\left(\frac{-2n\eta^2}{9(C_0 + 2\Phi^{-1}(n)(C_z^{mom} + C_y^{mom}))^2}\right) + \frac{\delta}{4} \leq \frac{\delta}{2}. \quad \blacksquare$$

5.9.2 Proof of Theorem 14

Proof of part (i). In this situation, the hypotheses of part **(i)** of Corollary 8 are satisfied and so the following bound holds:

$$\mathbb{E}\left[\sup_{H\in\mathcal{H}^{RC}}\left\{R(H)-\widehat{R}_n^{\infty}(H)\right\}\right]\leq\frac{C_1}{n}+\frac{C_2\log(n)}{n}+\frac{C_3\sqrt{\log(n)}}{\sqrt{n}}. \tag{117}$$

Let us denote $\Gamma_n:=\sup_{H\in\mathcal{H}^{RC}}\{R(H)-\widehat{R}_n^{\infty}(H)\}$. We may then apply the triangle inequality, insert the estimate on the difference between the empirical risk and its idealized counterpart obtained in Proposition 5 as well as the estimate on the expected value (117) to obtain that $\mathbb{P}$-a.s.,

$$\begin{aligned}\sup_{H\in\mathcal{H}^{RC}}\{R(H)-\widehat{R}_n(H)\}&=\sup_{H\in\mathcal{H}^{RC}}\{R(H)-\widehat{R}_n(H)+\widehat{R}_n^{\infty}(H)-\widehat{R}_n^{\infty}(H)\}-\mathbb{E}[\Gamma_n]+\mathbb{E}[\Gamma_n]\\&\leq\sup_{H\in\mathcal{H}^{RC}}\{\widehat{R}_n^{\infty}(H)-\widehat{R}_n(H)\}+\Gamma_n-\mathbb{E}[\Gamma_n]+\mathbb{E}[\Gamma_n]\\&\leq\sup_{H\in\mathcal{H}^{RC}}|\widehat{R}_n^{\infty}(H)-\widehat{R}_n(H)|+|\Gamma_n-\mathbb{E}[\Gamma_n]|+\mathbb{E}[\Gamma_n]\\&\leq\frac{(1-r^n)C_0}{n}+|\Gamma_n-\mathbb{E}[\Gamma_n]|+\frac{C_1}{n}+\frac{C_2\log(n)}{n}+\frac{C_3\sqrt{\log(n)}}{\sqrt{n}}.\end{aligned} \tag{118}$$

Part (a): Denote by η the upper bound that we need to prove holds with high probability, that is,

$$\eta:=\frac{(1-r^n)C_0+C_1}{n}+\frac{C_2\log(n)}{n}+\frac{C_3\sqrt{\log(n)}}{\sqrt{n}}+\frac{C_{bd}\sqrt{\log(\frac{4}{\delta})}}{\sqrt{2n}}.$$

Combining the estimate (118) with the exponential concentration inequality Proposition 19 then yields

$$\mathbb{P}\left(\sup_{H\in\mathcal{H}^{RC}}\{R(H)-\widehat{R}_n(H)\}>\eta\right)\leq\mathbb{P}\left(|\Gamma_n-\mathbb{E}[\Gamma_n]|>\frac{C_{bd}\sqrt{\log(\frac{4}{\delta})}}{\sqrt{2n}}\right)\leq\frac{\delta}{2}. \tag{119}$$

By applying the result that we just proved to the loss function $-L$ one obtains that

$$\mathbb{P}\left(\sup_{H\in\mathcal{H}^{RC}}\{\widehat{R}_n(H)-R(H)\}>\eta\right)\leq\frac{\delta}{2}.$$

Using that $|x|=\max(x,-x)$ one can thus combine the two estimates to deduce

$$\begin{aligned}&\mathbb{P}\left(\sup_{H\in\mathcal{H}^{RC}}\left|R(H)-\widehat{R}_n(H)\right|>\eta\right)\\&\quad\leq\mathbb{P}\left(\left\{\sup_{H\in\mathcal{H}^{RC}}\{R(H)-\widehat{R}_n(H)\}>\eta\right\}\cup\left\{\sup_{H\in\mathcal{H}^{RC}}\{\widehat{R}_n(H)-R(H)\}>\eta\right\}\right)\leq\delta.\end{aligned}$$

Part (b): Proceeding analogously as in part **(a)**, denote by η the high-probability upper bound which needs to be established, that is,

$$\eta = \frac{(1-r^n)C_0 + C_1}{n} + \frac{C_2 \log(n)}{n} + \frac{C_3\sqrt{\log(n)}}{\sqrt{n}} + B_\Phi(n,\delta).$$

Combining (118) with Corollary 21 then yields

$$\mathbb{P}\left(\sup_{H\in\mathcal{H}^{RC}}\{R(H) - \widehat{R}_n(H)\} > \eta\right) \leq \mathbb{P}\left(|\Gamma_n - \mathbb{E}[\Gamma_n]| > B_\Phi(n,\delta)\right) \leq \frac{\delta}{2}.$$

The claim then follows precisely as in the proof of part **(a)**.

Proof of part (ii). Firstly, one may use Proposition 5 to obtain $\mathbb{P}$-a.s.,

$$\sup_{H\in\mathcal{H}^{RC}}\{R(H) - \widehat{R}_n(H)\} \leq \sup_{H\in\mathcal{H}^{RC}}\{\widehat{R}_n^\infty(H) - \widehat{R}_n(H)\} + |\Gamma_n| \leq \frac{(1-r^n)C_0}{n} + |\Gamma_n|. \tag{120}$$

Setting

$$\eta = \frac{2}{\delta}\left(\frac{C_1}{n} + \frac{C_2\log(n)}{n} + \frac{C_{3,abs}\sqrt{\log(n)}}{\sqrt{n}}\right) + \frac{(1-r^n)C_0}{n} \tag{121}$$

and applying Markov's inequality, (120) and part **(ii)** of Corollary 8 then yields

$$\begin{aligned}
\mathbb{P}\left(\sup_{H\in\mathcal{H}^{RC}}\{R(H) - \widehat{R}_n(H)\} > \eta\right) &\leq \mathbb{P}\left(|\Gamma_n| > \frac{2}{\delta}\left(\frac{C_1}{n} + \frac{C_2\log(n)}{n} + \frac{C_{3,abs}\sqrt{\log(n)}}{\sqrt{n}}\right)\right) \\
&\leq \mathbb{E}[|\Gamma_n|]\frac{\delta}{2}\left(\frac{C_1}{n} + \frac{C_2\log(n)}{n} + \frac{C_{3,abs}\sqrt{\log(n)}}{\sqrt{n}}\right)^{-1} \\
&\leq \frac{\delta}{2}.
\end{aligned}$$

By applying what we just proved to the loss function $-L$ the claim then follows precisely as in the proof of part **(i)**.

Proof of part (iii). The proof is the same as the proof of part **(ii)**, except that instead of choosing η as in (121) one takes

$$\eta = \frac{(1-r^n)C_0}{n} + \frac{2}{\delta}\left(C_{1,abs}n^{-\frac{1}{2+\alpha^{-1}}} + C_2 n^{-\frac{2}{2+\alpha^{-1}}}\right)$$

and instead of using part **(ii)** of Corollary 8 one applies its part **(iii)** to estimate $\mathbb{E}[|\Gamma_n|]$. ■

5.10 Proof of Proposition 16

Denote by $\Theta := \left\{(\rho_A \mathbf{A}, \rho_C \mathbf{C}, \rho_{\boldsymbol{\zeta}} \boldsymbol{\zeta}) \mid (\rho_A, \rho_C, \rho_{\boldsymbol{\zeta}}) \in (-\frac{a}{\lambda^{\mathbf{A}}}, \frac{a}{\lambda^{\mathbf{A}}}) \times [-c, c] \times [-s, s]\right\}$ the random set of admissible parameters for the echo state network. Since

$$L_\sigma \left(\sum_{l=1}^{N} \sup_{(A,C,\overline{\boldsymbol{\zeta}}) \in \Theta} \|A_{l,\cdot}\|_\infty \right) = \frac{a}{\lambda^{\mathbf{A}}} L_\sigma \left(\sum_{l=1}^{N} \|\mathbf{A}_{l,\cdot}\|_\infty \right) = a \in (0,1),$$

for any realization of $\mathbf{A}, \mathbf{C}, \boldsymbol{\zeta}$ (that is, conditional on $\mathbf{A}, \mathbf{C}, \boldsymbol{\zeta}$) the assumptions of Proposition 11 are satisfied. Thus one may argue as in the proof of part **(ii)** of Theorem 14 to obtain that for any $\eta > 0$,

$$\mathbb{P}\left(\sup_{H \in \boldsymbol{\mathcal{H}}^{RC}} \{R(H) - \widehat{R}_n(H)\} > \frac{(1-r^n)C_0}{n} + \eta \middle| \mathbf{A}, \mathbf{C}, \boldsymbol{\zeta} \right) \leq \frac{\mathbb{E}[|\Gamma_n| \mid \mathbf{A}, \mathbf{C}, \boldsymbol{\zeta}]}{\eta}$$

and then apply part **(ii)** of Corollary 8 to obtain

$$\mathbb{P}\left(\sup_{H \in \boldsymbol{\mathcal{H}}^{RC}} \{R(H) - \widehat{R}_n(H)\} > \frac{(1-r^n)C_0}{n} + \eta \middle| \mathbf{A}, \mathbf{C}, \boldsymbol{\zeta} \right) \leq \frac{1}{\eta} \left(\frac{C_1}{n} + \frac{C_2 \log(n)}{n} + \frac{\mathbf{C}_{3,abs}\sqrt{\log(n)}}{\sqrt{n}} \right),$$

where the constants can be explicitly chosen using (51)-(52). In particular, C_1 and C_2 are given by (51) with C_I as in (26), and $\mathbf{C}_{3,abs}$ can be written using (52) as

$$\mathbf{C}_{3,abs} = 2\mathbf{C}_3 + \frac{4L_L \mathbb{E}\left[\|\mathbf{Y}_0\|_2^2\right]^{1/2}}{\sqrt{\log(\lambda_{max}^{-1})}},$$

with

$$\mathbf{C}_3 = \frac{2\sqrt{m} L_L \mathbf{C_{RC}}}{\sqrt{\log(\lambda_{max}^{-1})}}$$

and

$$\mathbf{C_{RC}} = \frac{\overline{L_h}}{1-a} \left(\lambda^{\mathbf{C}} \mathbb{E}\left[\|\mathbf{Z}_0\|_2^2\right]^{1/2} + \lambda^{\boldsymbol{\zeta}} \right) + L_{h,0}.$$

Taking expectations, one sees that

$$\mathbb{E}[\mathbf{C}_{3,abs}] = C_{3,abs},$$

where $C_{3,abs}$ is as in (52) with C_{RC} given by (73). Thus we obtain

$$\mathbb{P}\left(\sup_{H \in \boldsymbol{\mathcal{H}}^{RC}} \{R(H) - \widehat{R}_n(H)\} > \frac{(1-r^n)C_0}{n} + \eta \right) \leq \frac{1}{\eta} \left(\frac{C_1}{n} + \frac{C_2 \log(n)}{n} + \frac{C_{3,abs}\sqrt{\log(n)}}{\sqrt{n}} \right)$$

and the claim follows by arguing as in the proof of part **(ii)** in Theorem 14 . ■

Acknowledgments

Lukas Gonon and Juan-Pablo Ortega acknowledge partial financial support coming from the Research Commission of the Universität Sankt Gallen, the Swiss National Science Foundation (grants number 175801/1 and 179114), and the French ANR "BIPHOPROC" project (ANR-14-OHRI-0018-02). Lyudmila Grigoryeva acknowledges partial financial support of the Graduate School of Decision Sciences of the Universität Konstanz.

References

T. M. Adams and A. B. Nobel. Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. *Annals of Probability*, 38(4):1345–1367, 2010.

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

P. Alquier and O. Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883–913, 2012.

D. Andrews. First order autoregressive processes and strong mixing. *Cowles Founda-tion Discussion Papers 664*, 1983.

M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations*. 1999.

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, jan 2011.

R. T. Baillie. Long memory processes and fractional integration in econometrics. *Journal of Econometrics*, 73(1):5–59, 1996.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(3):463–482, 2003.

P. L. Bartlett, M. I. Jordan, and J. D. Mcauliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

P. L. Bartlett, D. J. Foster, and M. Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 2017-Decem:6241–6250, 2017.

S. Ben-David and S. Shalev-Shwartz. *Understanding Machine Learning: From Theory to Algorithms*. 2014.

J. Beran. *Statistics for Long-Memory Processes*. CRC Press, 1994.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

P. Bougerol and N. Picard. Strict Stationarity of Generalized Autoregressive Processes. *The Annals of Probability*, 1992.

O. Bousquet and A. Elisseeff. Stability and generalisaiton. *Journal of Machine Learning Reasearch*, 2:499–526, 2002.

S. Boyd and L. Chua. Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems*, 32(11):1150–1161, nov 1985.

A. Brandt. The stochastic equation Yn +1=AnYn + Bn with stationary coefficients. *Advances in Applied Probability*, 18(01):211–220, mar 1986.

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*, 4(1364), 2013.

M. Buehner and P. Young. A tighter bound for the echo state property. *IEEE Transactions on Neural Networks*, 17(3):820–824, 2006.

A. Christmann and I. Steinwart. *Support Vector Machines*. Springer New York, 2008.

B. D. Coleman and V. J. Mizel. On the general theory of fading memory. *Archive for Rational Mechanics and Analysis*, 29(1):18–31, jan 1968.

R. Couillet, G. Wainrib, H. Sevi, and H. T. Ali. The asymptotic performance of linear echo state neural networks. *Journal of Machine Learning Research*, 17(178):1–35, 2016.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

F. Cucker and D.-X. Zhou. *Learning Theory : An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, dec 1989.

J. Dambre, D. Verstraeten, B. Schrauwen, and S. Massar. Information processing capacity of dynamical systems. *Scientific reports*, 2(514), 2012.

J. Dedecker, P. Doukhan, G. Lang, J. R. León, S. Louhichi, and C. Prieur. *Weak Dependence: With Examples and Applications*. Springer Science+Business Media, 2007.

R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 2nd edition, 2014.

R. Engle. *Anticipating Correlations*. Princeton University Press, Princeton, NJ, 2009.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

M. Fabrizio, C. Giorgi, and V. Pata. *A new approach to equations with memory*, volume 198. 2010.

M. Fliess and D. Normand-Cyrot. Vers une approche algébrique des systèmes non linéaires en temps discret. In A. Bensoussan and J. Lions, editors, *Analysis and Optimization of Systems. Lecture Notes in Control and Information Sciences, vol. 28*. Springer Berlin Heidelberg, 1980.

C. Francq and J.-M. Zakoian. *GARCH Models: Structure, Statistical Inference and Financial Applications*. Wiley, 2010.

K.-i. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2:183–192, 1989.

S. Ganguli, D. Huh, and H. Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 105(48):18970–5, dec 2008.

E. Giné and J. Zinn. Some limit theorems for empirical processes. *Annals of Probability*, 12:929–989, 1984.

L. Gonon and J.-P. Ortega. Fading memory echo state networks are universal. *Preprint arXiv: 2010.12047*, pages 1–6, 2020a.

L. Gonon and J.-P. Ortega. Reservoir computing universality with stochastic inputs. *IEEE Transactions on Neural Networks and Learning Systems*, 31(1):100–112, 2020b.

L. Gonon, L. Grigoryeva, and J.-P. Ortega. Approximation error estimates for random neural networks and reservoir systems. *arXiv preprint 2002.05933*, 2020.

L. Grigoryeva and J.-P. Ortega. Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneous state-affine systems. *Journal of Machine Learning Research*, 19(24):1–40, 2018a.

L. Grigoryeva and J.-P. Ortega. Echo state networks are universal. *Neural Networks*, 108: 495–508, 2018b.

L. Grigoryeva and J.-P. Ortega. Differentiable reservoir computing. *Journal of Machine Learning Research*, 20(179):1–62, 2019.

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, and J.-P. Ortega. Nonlinear memory capacity of parallel time-delay reservoir computers in the processing of multidimensional signals. *Neural Computation*, 28:1411–1451, 2016.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 1992.

M. Hermans and B. Schrauwen. Memory in linear recurrent neural networks in continuous time. *Neural networks : the official journal of the International Neural Network Society*, 23(3):341–55, apr 2010.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, second edition, 2013.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

J. R. M. Hosking. Fractional differencing. *Biometrika*, 1981.

T. Hytönen, J. van Neerven, M. Veraar, and L. Weis. *Analysis in Banach Spaces*, volume I. Springer International Publishing, 2016.

D. Ibáñez-Soria, A. Soria-Frisch, J. Garcia-Ojalvo, and G. Ruffini. Characterization of the non-stationary nature of steady-state visual evoked potentials using echo state networks. *PLOS ONE*, 2019.

H. Jaeger. Short term memory in echo state networks. *Fraunhofer Institute for Autonomous Intelligent Systems. Technical Report.*, 152, 2002.

H. Jaeger. The 'echo state' approach to analysing and training recurrent neural networks with an erratum note. Technical report, German National Research Center for Information Technology, 2010.

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

A. Khintchine. Über dyadische Brüche. *Mathematische Zeitschriften*, 18:109–116, 1923.

P. Koiran and E. D. Sontag. Vapnik-Chervonenkis dimension of recurrent neural networks. *Discrete Applied Mathematics*, 1998.

V. Kuznetsov and M. Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

V. Kuznetsov and M. Mohri. Theory and algorithms for forecasting time series. 2018.

F. Laporte, A. Katumba, J. Dambre, and P. Bienstman. Numerical demonstration of neuromorphic computing with photonic crystal cavities. *Optics Express*, 26(7):7955, apr 2018.

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, jan 2012.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer-Verlag, 1991.

R. Legenstein and W. Maass. What makes a dynamical system computationally powerful? In S. Haykin, editor, *New directions in statistical signal processing: from systems to brain*. MIT Press, Cambridge, MA, 2007.

Z. Lu, B. R. Hunt, and E. Ott. Attractor reconstruction by machine learning. *Chaos*, 28 (6), 2018.

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

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.

W. Maass. Liquid state machines: motivation, theory, and applications. In S. S. Barry Cooper and A. Sorbi, editors, *Computability In Context: Computation and Logic in the Real World*, chapter 8, pages 275–296. 2011.

W. Maass and E. D. Sontag. Neural Systems as Nonlinear Filters. *Neural Computation*, 12 (8):1743–1772, aug 2000.

W. Maass, T. Natschläger, and H. Markram. Fading memory and kernel properties of generic cortical microcircuit models. *Journal of Physiology Paris*, 98(4-6 SPEC. ISS.): 315–330, 2004.

W. Maass, P. Joshi, and E. D. Sontag. Computational aspects of feedback in neural circuits. *PLoS Computational Biology*, 3(1):e165, 2007.

G. Manjunath and H. Jaeger. Echo state property linked to an input: exploring a fundamental characteristic of recurrent neural networks. *Neural Computation*, 25(3):671–696, 2013.

S. Marzen. Difference between memory and prediction in linear recurrent networks. *Physical Review E*, 96(3):1–7, 2017.

M. B. Matthews. *On the Uniform Approximation of Nonlinear Discrete-Time Fading-Memory Systems Using Neural Network Models*. PhD thesis, ETH Zürich, 1992.

D. J. McDonald, C. R. Shalizi, and M. Schervish. Nonparametric risk bounds for time-series forecasting. *Journal of Machine Learning Research*, 18:1–40, 2017.

S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.

J. Munkres. *Topology*. Pearson, second edition, 2014.

T. Natschläger, W. Maass, and H. Markram. The "Liquid Computer": a novel strategy for real-time computing on time series. *Special Issue on Foundations of Information Processing of TELEMATIK*, 8(1):39–43, 2002.

Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar. Optoelectronic reservoir computing. *Scientific reports*, 2:287, jan 2012.

J. Pathak, Z. Lu, B. R. Hunt, M. Girvan, and E. Ott. Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data. *Chaos*, 27(12), 2017.

J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott. Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach. *Physical Review Letters*, 120(2):24102, 2018.

T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422, 2004.

A. Rakhlin, K. Sridharan, and A. Tewari. Online learning via sequential complexities. *Journal of Machine Learning Research*, 16:155–186, 2010.

A. Rakhlin, K. Sridharan, and A. Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, 2014.

A. Rodan and P. Tino. Minimum complexity echo state network. *IEEE Transactions on Neural Networks*, 22(1):131–44, jan 2011.

S. Smale and D.-X. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 01(01):17–41, 2003.

E. Sontag. Realization theory of discrete-time nonlinear systems: Part I-The bounded case. *IEEE Transactions on Circuits and Systems*, 26(5):342–356, may 1979a.

E. D. Sontag. Polynomial Response Maps. In *Lecture Notes Control in Control and Information Sciences. Vol. 13*. Springer Verlag, 1979b.

E. D. Sontag. VC dimension of neural networks. *NATO ASI Series F Computer and Systems Sciences*, 168:69–96, 1998.

S. Sternberg. *Dynamical Systems*. Dover, 2010.

K. Vandoorne, J. Dambre, D. Verstraeten, B. Schrauwen, and P. Bienstman. Parallel reservoir computing using optical amplifiers. *IEEE Transactions on Neural Networks*, 22 (9):1469–1481, sep 2011.

K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman. Experimental demonstration of reservoir computing on a silicon photonics chip. *Nature Communications*, 5:78–80, mar 2014.

V. Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems 4 (NIPS 1991)*, pages 831–838, 1991.

V. Vapnik. *Statistical Learning Theory*. Wiley, adaptive a edition, 1998.

V. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Dokl. Akad. Nauk SSSR*, 181(4):781, 1968.

P. Verzelli, C. Alippi, and L. Livi. Echo State Networks with self-normalizing activations on the hyper-sphere. *Scientific Reports*, 9(13887), 2019.

Q. Vinckier, F. Duport, A. Smerieri, K. Vandoorne, P. Bienstman, M. Haelterman, and S. Massar. High-performance photonic reservoir computer based on a coherently driven passive cavity. *Optica*, 2(5):438–446, 2015.

V. Volterra. *Theory of Functionals and of Integral and Integro-Differential Equations.* Dover, 1959.

O. White, D. Lee, and H. Sompolinsky. Short-Term Memory in Orthogonal Neural Networks. *Physical Review Letters*, 92(14):148102, apr 2004.

N. Wiener. *Nonlinear Problems in Random Theory.* The Technology Press of MIT, 1958.

I. B. Yildiz, H. Jaeger, and S. J. Kiebel. Re-visiting the echo state property. *Neural Networks*, 35:1–9, nov 2012.

J. Zhang, Q. Lei, and I. S. Dhillon. Stabilizing gradients for deep neural networks via efficient SVD parameterization. 2018.