Journal of Machine Learning Research 17 (2016) 1-35 Submitted 2/16; Revised 7/16; Published 9/16

The Asymptotic Performance of Linear Echo State Neural Networks

Romain Couillet ROMAIN.COUILLET@CENTRALESUPELEC.FR
CentraleSupélec – LSS – Université ParisSud (Gif-sur-Yvette, France).

Gilles Wainrib GILLES.WAINRIB@ENS.FR
Département Informatique, team DATA, Ecole Normale Supérieure (Paris, France).

Harry Sevi HARRY.SEVI@ENS-LYON.FR
Laboratoire de Physique, Ecole Normale Supérieure de Lyon (Lyon, France).

Hafiz Tiomoko Ali HAFIZ.TIOMOKOALI@CENTRALESUPELEC.FR
CentraleSupélec – LSS – Université ParisSud (Gif-sur-Yvette, France).

Editor: Yoshua Bengio

Abstract

In this article, a study of the mean-square error (MSE) performance of linear echo-state neural networks is performed, both for training and testing tasks. Considering the realistic setting of noise present at the network nodes, we derive deterministic equivalents for the aforementioned MSE in the limit where the number of input data T and network size n both grow large. Specializing then the network connectivity matrix to specific random settings, we further obtain simple formulas that provide new insights on the performance of such networks.

Keywords: recurrent neural networks; echo state networks; random matrix theory; mean square error; linear networks

1. Introduction

Echo State Networks (ESN) are a class of recurrent neural networks (RNN) designed for performing supervised learning tasks, such as time-series prediction (Jaeger, 2001; Jaeger and Haas, 2004) or more generally any supervised learning task involving sequences. The ESN architecture is a special case of the general framework of reservoir computing (Lukoševičius and Jaeger, 2009). The ESN reservoir is a fixed (generally randomly designed) recurrent neural network, driven by a (usually time dependent) input. Since the internal connectivity matrix is not modified during learning, the number of parameters to learn is much smaller than in a classical RNN setting and the system is as such less prone to overfitting. However, the prediction performance of ESN often depends significantly on several hyper-parameters controlling the law of the internal connectivity matrix.

It has in particular been understood that the spectral radius and spectral norm of the connectivity matrix play a key role on the stability of the network (Jaeger, 2001) and that the structure of the connectivity matrix may be adapted to trade memory capacities versus task complexity (Verstraeten et al., 2010; Rodan and Tiňo, 2011; Strauss et al., 2012; Ozturk et al., 2007). Nonetheless, to date, and to the best of the authors' knowledge, the understanding of echo-state networks has progressed mostly through extensive empirical studies and lacks solid theoretical foundations.

In the present article, we consider linear ESN's with a general connectivity matrix and internal network noise. By leveraging tools from the field of random matrix theory, we shall attempt to provide a first *theoretical* study of the performance of ESN's. Beyond the obvious advantage of exploiting theoretical formulas to select the optimal hyper-parameters, this mathematical study reveals key quantities that intimately relate the internal network memory to the input-target relationship, therefore contributing to a better understanding of short-term memory properties of RNNs.

More specifically, we shall consider an n-node ESN trained with an input of size T and shall show that, assuming the internal noise variance η^2 remains large compared to $1/\sqrt{n}$, the training and testing performances of the network can be well approximated by a deterministic quantity which is a function of the training and test data as well as the connectivity matrix. Under the further assumption that the connectivity matrix is random, we shall then obtain closed-form formulas of the aforementioned performances for a large class of connectivity structures. The reach of our study so far only addresses ESN's with linear activation functions, a limitation which we anticipate to work around with more elaborate methods in the future, as discussed in Section 4.

At this point, we wish to highlight the specificity of our approach regarding (i) the introduction of noise perturbing the internal dynamics of the reservoir and (ii) the restriction to linear networks.

The introduction of additive noise in the reservoir is inspired on the one hand by it being a natural assumption in modelling biological neural networks (Ganguli et al., 2008; Toyoizumi and Abbott, 2011) and on the other hand by the observation in (Jaeger, 2001) that ESN's are very sensitive to low variance noise and thus likely unstable in this regime, a problem successfully cured by internal noise addition (Jaeger, 2005).[1] From the neuro-computational perspective, we shall observe tight connections between the ESN performance and the reservoir information processing capacities discussed in (Ganguli et al., 2008). As for the artificial neural network viewpoint, it shall be noticed that the internal noise regularizes the network in a way sharing interesting similarities with the well-known connection between noise at the network output and Tikhonov regularization (Bishop, 1995). It is, as a matter of fact, already mentioned in (Lukoševičius and Jaeger, 2009, Section 8.2) that internal noise behaves as a natural regularization option (similar to what input or output noise would) although this aspect was not deeply investigated. More importantly, while internal noise necessarily leads to random outputs (a not necessarily desirable feature on the onset), we shall show that all these outputs (almost surely) asymptotically have the same performance, thus inducing random but equally useful innovation; this we believe is a more desirable feature than deterministic arbitrary biases in the innovation (see the comments around Remark 12 in Section 2.2).

As for the choice of studying linear activation functions, rarely considered in the practical side of RNNs, it obviously follows first from a mathematical tractability of the problem under study. Nonetheless, while being clearly a strong limitation of our study (recall that the non-linearity is the main driver for the network to perform complex tasks), we believe it brings sufficient insights (at least as far as memory capabilities and minimal performance are concerned) and exploitable results when it comes to parametrizing non-linear network counterparts.

Among other results, the main findings of our study are as follows:

1. According to Jaeger in (Jaeger, 2005) (specifically in the context of reservoirs with output feedback): "When you are training an ESN with output feedback from accurate (mathematical, noise-free) training data, stability of the trained network is often difficult to achieve. A method that works wonders is to inject noise into the dynamical reservoir update during sampling [...]. It is not clearly understood why this works."

1. we retrieve a deterministic implicit expression for the mean-square error (MSE) performance of training and testing for any fixed connectivity matrix $W \in \mathbb{R}^{n\times n}$ which, for every given internal noise variance $\eta^2 > 0$, is all the more accurate that the network size n is large

2. the aforementioned expression reveals fundamental quantities which generalize several known qualitative notions of ESN's, such as the *memory capacity* and the *Fisher memory curve* (Jaeger, 2001; Ganguli et al., 2008);

3. we obtain more tractable closed-form expressions for the same quantities for simple classes of random normal and non-normal matrices; these two classes exhibit a strikingly different asymptotic performance behavior;

4. from the previous analysis, we shall also introduce a novel multi-modal connectivity matrix that adapts to a wider scope of memory ranges and that is reminiscent to the long short-term memory ESNs designed in (Xue et al., 2007);

5. an important interplay between memory and internal noise will be shed light on, by which the questions of noise-induced stability are better understood.

The remainder of the article is organized as follows. In Section 2, we introduce the ESN model and the associated supervised learning problem and we give our main theoretical results in Theorems 2 and 9 (technical proofs are deferred to the Appendix). Then, in Section 3, we apply our theoretical results for various choices of specific connectivity matrices and discuss their consequences in terms of prediction performance. Finally, in Section 4, we discuss our findings and their limitations.

Notations: In the remainder of the article, uppercase characters will stand for matrices, lowercase for scalars or vectors. The transpose operation will be denoted $(\cdot)^\mathsf{T}$. The multivariate Gaussian distribution of mean μ and covariance C will be denoted $\mathcal{N}(\mu, C)$. The notation $V = \{V_{ij}\}_{i=1,j=1}^{n,T}$ denotes the matrix with (i,j)-entry V_{ij} (scalar or matrix), $1 \le i \le n$, $1 \le j \le T$, while $\{V_i\}_{i=1}^n$ is the row-wise concatenation of the V_i's and $\{V_j\}_{j=1}^T$ the column-wise concatenation of the V_j's. We further introduce the notation $(x)^+ \equiv \max(x, 0)$. For random or deterministic matrices X_n and $Y_n \in \mathbb{R}^{n\times n}$, the notation $X_n \leftrightarrow Y_n$ stands for $\frac{1}{n}\operatorname{tr} A_n(X_n - Y_n) \to 0$ and $a_n^\mathsf{T}(X_n - Y_n)b_n \to 0$, almost surely, for every deterministic matrix A_n and vectors a_n, b_n having bounded norm (spectral norm for matrices and Euclidean norm for vectors); for $X_n, Y_n \in \mathbb{R}$ scalar, the notation will simply mean that $X_n - Y_n \to 0$ almost surely. The notation $\rho(X)$ will denote the spectral radius of matrix X, while $\|X\|$ will denote its operator norm (and for vectors, $\|x\|$ is the Euclidean norm). The symbol $\boldsymbol{\delta}_x$ shall stand for Kronecker's delta function, i.e., $\boldsymbol{\delta}_x(y) = 1$ if $y = x$ (or x is true) and zero otherwise.

2. Main Results

We consider here an echo-state neural network constituted of n nodes, with state $x_t \in \mathbb{R}^n$ at time instant t, connectivity matrix $W \neq 0$, and input source sequence $\ldots, u_{-1}, u_0, u_1, \ldots \in \mathbb{R}$. The state evolution is given by the linear recurrent equation

$$x_{t+1} = Wx_t + mu_{t+1} + \eta\varepsilon_{t+1}$$

for all $t \in \mathbb{Z}$, in which $\eta > 0$ and $\varepsilon_t \sim \mathcal{N}(0, I_n)$, while $m \in \mathbb{R}^n$ is the input-to-network connectivity.

Our first objective is to understand the training performance of such a network. To this end, we shall focus on a (training) time window $\{0, \ldots, T-1\}$ and will denote $X =$

$\{x_j\}_{j=0}^{T-1} \in \mathbb{R}^{n\times T}$ as well as $A = MU$, $M \in \mathbb{R}^{n\times T}$, $U \in \mathbb{R}^{T\times T}$, where

$$M \equiv \left\{W^j m\right\}_{j=0}^{T-1}$$
$$U \equiv \frac{1}{\sqrt{T}} \left\{u_{j-i}\right\}_{i,j=1}^{T}.$$

With these notations, we especially have $X = \sqrt{T}(A+Z)$, where $Z = \frac{\eta}{\sqrt{T}}\{\sum_{k=0}^{\infty} W^k \varepsilon_{j-k}\}_{j=0}^{T-1}$. The matrix A can be seen here as the matrix carrying the information about the input vector u while Z serves the purpose of regularization noise.

For X to be properly defined (at least almost surely so), we shall impose the following hypothesis.

Assumption 1 (Spectral Norm) *The spectral norm $\|W\|$ of W satisfies $\|W\| < 1$.*

Note that this constraint is in general quite strong and it is believed (following the insights of previous works (Jaeger, 2001)) that for many model choices of W, it can be lighten to merely requiring that the spectral *radius* $\rho(W)$ be smaller than one. Nonetheless, in the course of the article, we shall often take W to be such that both its spectral norm and spectral radius coincide.

2.1 Training Performance

The training step consists in teaching the network to obtain a specific output sequence $r = \{r_j\}_{j=0}^{T-1}$ out of the network, when fed by a corresponding input vector $u = \{u_j\}_{j=0}^{T-1}$ over the time window. To this end, unlike conventional neural networks, where W is adapted to u and r, ESN's adopt the strategy to solely enforce an output link from the network to a sink (or readout). Letting $\omega = \{\omega_i\}_{i=1}^{n}$ be the network-to-sink connectivity vector, we shall consider here that ω is obtained as the (least-square) minimizer of $\|X^\mathsf{T}\omega - r\|^2$. When $T > n$, we have

$$\omega \equiv \left(XX^\mathsf{T}\right)^{-1} Xr \tag{1}$$

which is almost surely well-defined (since $\eta > 0$) or, when $T \leq n$,

$$\omega \equiv X\left(X^\mathsf{T}X\right)^{-1} r. \tag{2}$$

The per-input mean-square error in training associated with the couple (u, r) for the ESN under study is then defined as

$$E_\eta(u, r) \equiv \frac{1}{T}\left\|r - X^\mathsf{T}\omega\right\|^2 \tag{3}$$

which is identically zero when $T \leq n$.

Our first objective is to study precisely the random quantity $E_\eta(u, r)$ for every given W and noise variance η^2 in the limit where $n \to \infty$. Our scaling hypotheses are as follows.

Assumption 2 (Random Matrix Regime) *The following conditions hold:*

1. $\limsup_n n/T < \infty$
2. $\limsup_n \|AA^\mathsf{T}\| < \infty$.

That is, according to Item 1, we allow n to grow with T. Also, from Item 2, we essentially allow u_t to be of order $O(1)$ (unless u is sparse and then u_t may be as large as $O(\sqrt{T})$) when m remains of bounded Euclidean norm. Under this setting, and along with Assumption 1, we shall thus essentially require all neural connections to be of order $O(n^{-\frac{1}{2}})$ while all input and output data (constituents of u and r) shall be in general of order $O(1)$.

For every square symmetric matrix $B \in \mathbb{R}^{n\times n}$, a central quantity in random matrix theory is the resolvent $(B - zI_n)^{-1}$ defined for every $z \in \mathbb{C} \setminus \mathcal{S}_B$, with $\mathcal{S}_B \subset \mathbb{R}$ the support of the eigenvalues of B. Here, letting $z = -\gamma$ for some $\gamma > 0$, it is particularly convenient to make the following observation.

Lemma 1 (Training MSE and resolvent) *For $\gamma > 0$, let $\tilde{Q}_\gamma \equiv (\frac{1}{T}X^\mathsf{T}X + \gamma I_T)^{-1}$. Then we have, for $E_\eta(u, r)$ defined as in (3),*

$$E_\eta(u, r) = \lim_{\gamma\downarrow 0} \gamma \frac{1}{T} r^\mathsf{T} \tilde{Q}_\gamma r.$$

Our first technical result provides an asymptotically tight approximation for $\tilde{Q}_\gamma$ for every $\gamma > 0$. Recall that, for $X_n, Y_n \in \mathbb{R}^{n\times n}$, the notation $X_n \leftrightarrow Y_n$ means that, for every deterministic and bounded norm matrix A_n or vector a_n, b_n, $\frac{1}{n}\operatorname{tr} A_n(X_n - Y_n) \to 0$ and $a_n^\mathsf{T}(X_n - Y_n)b_n \to 0$, almost surely.

Theorem 2 (Deterministic Equivalent) *Let Assumptions 1–2 hold. For $\gamma > 0$, let also $Q_\gamma \equiv (\frac{1}{T}XX^\mathsf{T} + \gamma I_n)^{-1}$ and $\tilde{Q}_\gamma \equiv (\frac{1}{T}X^\mathsf{T}X + \gamma I_T)^{-1}$. Then, as $n \to \infty$, the following approximations hold:*

$$Q_\gamma \leftrightarrow \bar{Q}_\gamma \equiv \frac{1}{\gamma}\left(I_n + \eta^2 \tilde{R}_\gamma + \frac{1}{\gamma} A \left(I_T + \eta^2 R_\gamma\right)^{-1} A^\mathsf{T}\right)^{-1}$$
$$\tilde{Q}_\gamma \leftrightarrow \bar{\tilde{Q}}_\gamma \equiv \frac{1}{\gamma}\left(I_T + \eta^2 R_\gamma + \frac{1}{\gamma} A^\mathsf{T} \left(I_n + \eta^2 \tilde{R}_\gamma\right)^{-1} A\right)^{-1}$$

where $R_\gamma \in \mathbb{R}^{T\times T}$ and $\tilde{R}_\gamma \in \mathbb{R}^{n\times n}$ are solutions to the system of equations

$$R_\gamma = \left\{\frac{1}{T}\operatorname{tr}\left(S_{i-j}\bar{Q}_\gamma\right)\right\}_{i,j=1}^{T}$$
$$\tilde{R}_\gamma = \sum_{q=-\infty}^{\infty} \frac{1}{T}\operatorname{tr}\left(J^q \bar{\tilde{Q}}_\gamma\right) S_q$$

with $[J^q]_{ij} \equiv \boldsymbol{\delta}_{i+q,j}$ and $S_q \equiv \sum_{k\geq 0} W^{k+(-q)^+}(W^{k+q^+})^\mathsf{T}$.[2]

Remark 3 (On Theorem 2) *Theorem 2 is in fact valid under more general assumptions than in the present setting. In particular, A may be any deterministic matrix satisfying Assumption 2. However, when $A = MU$, an important phenomenon arises, which is that A behaves similar to a low-rank matrix, since, by Assumption 1, only $o(n)$ columns of M have non-vanishing norm. As such, by a low-rank perturbation argument, it can be shown that the term $\bar{Q}_\gamma$ in the expression of R_γ and the term $\bar{\tilde{Q}}_\gamma$ in the expression of $\tilde{R}_\gamma$ can be replaced by $\gamma^{-1}(I_n + \eta^2\tilde{R}_\gamma)^{-1}$ and $\gamma^{-1}(I_T + \eta^2 R_\gamma)^{-1}$, respectively. As such, R_γ and $\tilde{R}_\gamma$ only depend on the matrix W and the parameter η^2, and are thus asymptotically independent of the input data matrix U. Note also in passing that, while $\tilde{R}_\gamma$ is defined with a sum over $q = -\infty$ to ∞, this summation is empty for all $|q| \geq T$.*

2. Note that $\operatorname{tr}(J^q B)$ is merely $\operatorname{tr}(J^q B) = \sum_{i=1+q^+}^{T-q^+} [B_{i,i+q}]$.

In order to evaluate the training mean-square error $E_\eta(u,r)$ from Lemma 1, one must extend Theorem 2 uniformly over γ approaching zero. This can be guaranteed under the following additional assumption.

Assumption 3 (Network size versus training time) *As $n \to \infty$, $n/T \to c \in [0,1) \cup (1,\infty)$.*

Under Assumption 3, two scenarios must be considered. Either $c < 1$ or $c > 1$. In the former case, we can show that, as $\gamma \downarrow 0$, R_γ and $\gamma\tilde{R}_\gamma$ have well defined limits. Besides, it appears that the limit of $\eta^2 R_\gamma$ does not depend on η^2, so that we shall denote $\mathcal{R}$ and $\tilde{\mathcal{R}}$ the limits of $\eta^2 R_\gamma$ and $\gamma\tilde{R}_\gamma$, as $\gamma \downarrow 0$, respectively. Similarly, $\eta^2\bar{Q}_\gamma$ and $\gamma\bar{\tilde{Q}}_\gamma$ converge to well defined limits, denoted respectively $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$. Symmetrically, for $c > 1$, as $\gamma \downarrow 0$, γR_γ and $\eta^2\tilde{R}_\gamma$ have well-behaved limits which we shall also refer to as $\mathcal{R}$ and $\tilde{\mathcal{R}}$; similarly, $\gamma\bar{Q}_\gamma$ and $\eta^2\bar{\tilde{Q}}_\gamma$ converge to non trivial limits again denoted $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$. These results are gathered in the following proposition.

Proposition 4 (Small γ limit of Theorem 2) *Let Assumptions 1–3 hold. For all large n, define $\mathcal{R}$ and $\tilde{\mathcal{R}}$ a pair of solutions of the system*

$$\mathcal{R} = c\left\{\frac{1}{n}\operatorname{tr}\left(S_{i-j}\left(\boldsymbol{\delta}_{c>1}I_n + \tilde{\mathcal{R}}\right)^{-1}\right)\right\}_{i,j=1}^{T}$$

$$\tilde{\mathcal{R}} = \sum_{q=-\infty}^{\infty}\frac{1}{T}\operatorname{tr}\left(J^q(\boldsymbol{\delta}_{c<1}I_T + \mathcal{R})^{-1}\right)S_q.$$

Subsequently define

$$\tilde{\mathcal{Q}} \equiv \left(\boldsymbol{\delta}_{c<1}I_T + \mathcal{R} + \frac{1}{\eta^2}A^\mathsf{T}\left(\boldsymbol{\delta}_{c>1}I_n + \tilde{\mathcal{R}}\right)^{-1}A\right)^{-1}$$

$$\mathcal{Q} \equiv \left(\boldsymbol{\delta}_{c>1}I_n + \tilde{\mathcal{R}} + \frac{1}{\eta^2}A\left(\boldsymbol{\delta}_{c<1}I_T + \mathcal{R}\right)^{-1}A^\mathsf{T}\right)^{-1}.$$

Then, with the definitions of Theorem 2, we have the following results.

1. *If $c < 1$, then in the limit $\gamma \downarrow 0$, $\eta^2 R_\gamma \to \mathcal{R}$, $\gamma\tilde{R}_\gamma \to \tilde{\mathcal{R}}$, $\eta^2\bar{Q}_\gamma \to \mathcal{Q}$, and $\gamma\bar{\tilde{Q}}_\gamma \to \tilde{\mathcal{Q}}$.*
2. *If $c > 1$, then in the limit $\gamma \downarrow 0$, $\gamma R_\gamma \to \mathcal{R}$, $\eta^2\tilde{R}_\gamma \to \tilde{\mathcal{R}}$, $\gamma\bar{Q}_\gamma \to \mathcal{Q}$, and $\eta^2\bar{\tilde{Q}}_\gamma \to \tilde{\mathcal{Q}}$.*

With these notations, we now have the following result.

Corollary 5 (Training MSE for $n < T$) *Let Assumptions 1–3 hold and let $r \in \mathbb{R}^T$ be of $O(\sqrt{T})$ Euclidean norm. Then, with $E_\eta(u,r)$ defined in (3), as $n \to \infty$,*

$$E_\eta(u,r) \leftrightarrow \begin{cases} (1/T)r^\mathsf{T}\tilde{\mathcal{Q}}r & ,\ c < 1 \\ 0 & ,\ c > 1. \end{cases}$$

It is interesting at this point to discuss the a priori involved expression of Proposition 4 and Corollary 5. Let us concentrate on the interesting $c < 1$ case. To start with, observe that $\mathcal{R}$ and $\tilde{\mathcal{R}}$ are deterministic matrices which only depend on W through the S_q matrices so that the only dependence of $E_\eta(u,r)$ on the noise variance η^2 lies explicitly in the expression of $\tilde{\mathcal{Q}}$. Now, making $A^\mathsf{T}\tilde{\mathcal{R}}^{-1}A$ explicit, we have the following telling limiting expression for $E_\eta(u,r)$

$$E_\eta(u,r) \leftrightarrow \frac{1}{T}r^\mathsf{T}\left(I_T + \mathcal{R} + \frac{1}{\eta^2}U^\mathsf{T}\left\{m^\mathsf{T}(W^i)^\mathsf{T}\tilde{\mathcal{R}}^{-1}W^j m\right\}_{i,j=0}^{T-1}U\right)^{-1}r. \tag{4}$$

Recalling that $\tilde{\mathcal{R}}$ is a linear combination of the matrices $S_q = W^{(-q)^+} S_0 W^{(q^+)}$, with $S_0 = \sum_{k\geq 0} W^k (W^k)^\mathsf{T}$, the expression $\frac{1}{\eta^2} m^\mathsf{T} (W^i)^\mathsf{T} \tilde{\mathcal{R}}^{-1} W^j m$ is strongly reminiscent of the Fisher memory curve $f : \mathbb{N} \to \mathbb{R}$ of the ESN, introduced in (Ganguli et al., 2008) and defined by $f(k) = \frac{1}{\eta^2} m^\mathsf{T} (W^k)^\mathsf{T} S_0^{-1} W^k m$. The Fisher memory curve $f(k)$ qualifies the ability of a k-step behind past input to influence the ESN at present time. Correspondingly, it appears here that the ability of the ESN to retrieve the desired expression of r from input u is importantly related to the matrix $\{\frac{1}{\eta^2} m^\mathsf{T} (W^i)^\mathsf{T} \tilde{\mathcal{R}}^{-1} W^j m\}_{i,j=0}^{T-1}$. As a matter of fact, for $c = 0$ (thus for a long training period), note that $\mathcal{R} = 0$ while $\tilde{\mathcal{R}} = S_0$ and we then find in particular

$$E_\eta(u, r) \leftrightarrow \frac{1}{T} r^\mathsf{T} \left(I_T + \frac{1}{\eta^2} U^\mathsf{T} \left\{ m^\mathsf{T} (W^i)^\mathsf{T} S_0^{-1} W^j m \right\}_{i,j=0}^{T-1} U \right)^{-1} r.$$

Pushing further our discussion on $\mathcal{R}$ and $\tilde{\mathcal{R}}$, it is interesting to intuit their respective structures. Observe in particular that $\mathcal{R}_{ij}$ depends only on $i - j$ and thus $\mathcal{R}$ is a Toeplitz matrix. Besides, since $\operatorname{tr} B = \operatorname{tr} B^\mathsf{T}$ for square matrices B, from $S_{i-j}^\mathsf{T} = S_{j-i}$ it comes that $\mathcal{R}_{ij} = \mathcal{R}_{ji}$. Also note that, since $\rho(W^q) = \rho(W)^q$ decays exponentially as $q \to \infty$, it is expected that $\mathcal{R}_{i,i+q}$ decays exponentially fast for large q. As a consequence, $\mathcal{R}$ is merely defined by $o(n)$ first entries of its first row.

From the results of (Gray, 2006) on Toeplitz versus circulant matrices, it then appears that, for every deterministic matrix B, $\frac{1}{T} \operatorname{tr} B \mathcal{R}^{-1}$ is well approximated by $\frac{1}{T} \operatorname{tr} B \mathcal{R}_c^{-1}$ for $\mathcal{R}_c$ a circulant matrix approximation of $\mathcal{R}$. Since circulant matrices are diagonalizable in a Fourier basis, so are their inverses and then, *as far as normalized traces are concerned*, $(I_T + \mathcal{R})^{-1}$ can be seen as approximately Toeplitz with again decaying behavior away from the main diagonals. Although slightly ambiguous, this approximation still makes it that the trace $\frac{1}{T} \operatorname{tr} J^q (I_T + \mathcal{R})^{-1}$ appearing in the expression of $\tilde{\mathcal{R}}$, is well approximated by any value $[(I_T + \mathcal{R})^{-1}]_{i,i+q}$ for i sufficiently far from 1 and T, and decays to zero as q grows large. This, and the fact that S_q also decays exponentially fast in norm allows us to conclude that $\tilde{\mathcal{R}}$ can be seen as a decaying weighted sum of $o(n)$ matrices S_q.

As shall be shown in Section 3, for W taken random with sufficient invariance properties, fundamental differences appear in the structure of $\mathcal{R}$ and $\tilde{\mathcal{R}}$ depending on whether W is taken normal or not. In particular, for W non-normal with left and right independent isotropic eigenvectors and m deterministic or random independent of W, $\mathcal{R}$ is well approximated by a scaled identity matrix and $\{m^\mathsf{T} (W^i)^\mathsf{T} \tilde{\mathcal{R}}^{-1} W^j m\}_{i,j=0}^{T-1}$ well approximated by a diagonal matrix with exponential decay along the diagonal.

Having a clearer understanding of Corollary 5, a few key remarks are in order.

Remark 6 (On the ESN stability to low noise levels) *It is easily seen by differentiation along η^2 that $r^\mathsf{T} \tilde{\mathcal{Q}} r$ is an increasing function of η^2, thus having a minimum as $\eta^2 \downarrow 0$. It is thus tempting to suppose that $E_\eta(u, r)$ converges to this limit in the noiseless case (i.e., for $\eta^2 = 0$). Such a reasoning is however hazardous and incorrect in most cases. Indeed, Corollary 5 only ensures an appropriate approximation of $E_\eta(u, r)$ for given $\eta > 0$ in the limit where $n \to \infty$. Classical random matrix considerations allow one to assert slightly stronger results. In particular, for the approximation of $E_\eta(u, r)$ to hold, one may allow η^2 to depend on n in such a way that $\eta^2 \gg n^{-\frac{1}{2}}$. This indicates that n must be quite large for the ESN behavior at moderate noise levels to be understood through the random matrix method. What seems like a defect of the tool on the onset in fact sheds some light on a deeper feature of ESN's. When η^2 is of the same order of magnitude or smaller than $n^{-\frac{1}{2}}$, Corollary 5 may become invalid due to the* resurgence of randomness *from $\cdots, \varepsilon_{-1}, \varepsilon_0, \varepsilon_1, \ldots$ Precisely, when η^2 gets small and thus the training MSE variance should decay, an opposite effect makes*

the MSE more random and thus possibly no longer tractable; this means in particular that, for any two independent runs of the ESN (with different noise realizations), all other parameters being fixed, the resulting MSE's might be strikingly different, making the network quite unstable. In practice, the opposition of the reduced noise variance η^2 and the resurgence of noise effects lead to various behaviors depending on the task and input data under considerations, ranging from largely increased MSE fluctuations at low η^2 to reduced fluctuations, through stabilisation of the fluctuations. In some specific cases discussed later, it might nonetheless be accepted to let $\eta^2 \to 0$ irrespective of n while keeping the random matrix approximation valid.

Remark 7 (Memory capacity revisited) *For $c < 1$, letting $u_k = \sqrt{T}\boldsymbol{\delta}_t$ and $r_k = \sqrt{T}\boldsymbol{\delta}_{t-\tau}$ (that is, all input energy is gathered in a single entry), for some $t, \tau \in \mathbb{N}$, makes the ESN fill a pure delay task of τ time-steps. In this case, we find that*

$$E_\eta(u,r) \leftrightarrow \left[\left(I_T + \mathcal{R} + \frac{1}{\eta^2} \left\{ m^\mathsf{T} (W^i)^\mathsf{T} \tilde{\mathcal{R}}^{-1} W^j m \right\}_{i,j=0}^{T-1} \right)^{-1} \right]_{\tau+1,\tau+1}.$$

In the particular case where, for all $i \neq j$, $\mathcal{R}_{ij} = o(1)$ and $m^\mathsf{T}(W^i)^\mathsf{T}\tilde{\mathcal{R}}^{-1}W^j m = o(1)$ (see Section 3.1 for a practical application with random non-normal W), by a uniform control argument due to the fast decaying far off-diagonal elements of $\mathcal{R}$ and $\{m^\mathsf{T}(W^i)^\mathsf{T}\tilde{\mathcal{R}}^{-1}W^j m\}$, the training MSE is further (almost surely) well approximated as

$$E_\eta(u,r) \leftrightarrow \frac{\eta^2}{\eta^2(1+\mathcal{R}_{11}) + m^\mathsf{T}(W^\tau)^\mathsf{T}\tilde{\mathcal{R}}^{-1}W^\tau m}.$$

If the quantity $m^\mathsf{T}(W^\tau)^\mathsf{T}\tilde{\mathcal{R}}^{-1}W^\tau m$ remains away from zero as $n \to \infty$, then it is allowed here to say (as opposed to the general case discussed in Remark 6) that $E_\eta(u,r) \to 0$ as $\eta \to 0$ and that $\eta^2/E_\eta(u,r) \sim m^\mathsf{T}(W^\tau)^\mathsf{T}\tilde{\mathcal{R}}^{-1}W^\tau m$, where we recover again a generalized form of the Fisher information curve at delay τ. From this discussion and Remark 6, we propose to define a novel network memory capacity metric $\mathrm{MC}(\tau)$, representing the inverse slope of decay of $E_\eta(\sqrt{T}\boldsymbol{\delta}_t, \sqrt{T}\boldsymbol{\delta}_{t-\tau})$ for small η^2:

$$\mathrm{MC}(\tau) \equiv \lim_{\eta\downarrow 0} \liminf_n \left[\left(\eta^2(I_T + \mathcal{R}) + \left\{ m^\mathsf{T}(W^i)^\mathsf{T}\tilde{\mathcal{R}}^{-1}W^j m \right\}_{i,j=0}^{T-1} \right)^{-1}_{\tau+1,\tau+1} \right]^{-1}.$$

Remark 7 follows up on recent works, here from an MSE performance perspective, that establish links between memory capacity metrics and the Fisher memory curve, as in e.g., (Tiňo and Rodan, 2013). Practical applications of Corollary 5 to specific matrix models for W shall be derived in Section 3. Beforehand, we will study the more involved question of the test MSE performance.

2.2 Test Performance

In this section, we assume $\omega \equiv \omega(X; u, r)$ has been obtained as per (1) or (2), depending on whether $c < 1$ or $c > 1$. We now consider the test performance of the ESN that corresponds to its ability to map an input vector $\hat{u} \in \mathbb{R}^{\hat{T}}$ to an expected output vector $\hat{r} \in \mathbb{R}^{\hat{T}}$ of duration $\hat{T}$ in such a way to fulfill the same task that links u to r. For notational convenience, all test data will be denoted with a hat mark on top.

As opposed to the training mean square error, the testing MSE, defined as

$$\hat{E}_\eta(u,r;\hat{u},\hat{r}) \equiv \frac{1}{\hat{T}} \left\| \hat{r} - \hat{X}^\mathsf{T}\omega \right\|^2 \tag{5}$$

where $\hat{X} = \{\hat{x}_j\}_{j=0}^{\hat{T}-1} \in \mathbb{R}^{n\times\hat{T}}$ is defined by the recurrent equation $\hat{x}_{t+1} = W\hat{x}_t + m\hat{u}_{t+1} + \eta\hat{\varepsilon}_{t+1}$, with $\hat{\varepsilon}_t \sim \mathcal{N}(0, I_n)$ independent of the ε_t's, does not assume a similar simple form as the training MSE. We importantly assume here that a sufficiently long washout period between training and testing is present in the sense that $\hat{x}_0$ is assumed independent of the x_t described in the previous section (see Remark 22 in Appendix A for a discussion on the results generalization when no washout period is assumed). Under these assumptions, we merely have the following result.

Lemma 8 (Testing MSE) *For $\gamma > 0$, $Q_\gamma = (\frac{1}{T}XX^\mathsf{T} + \gamma I_n)^{-1}$, and $\tilde{Q}_\gamma = (\frac{1}{T}X^\mathsf{T}X + \gamma I_T)^{-1}$, we have*

$$\begin{aligned}\hat{E}_\eta(u, r; \hat{u}, \hat{r}) &= \lim_{\gamma\downarrow 0} \frac{1}{\hat{T}}\|\hat{r}\|^2 + \frac{1}{T^2\hat{T}} r^\mathsf{T} X^\mathsf{T} Q_\gamma \hat{X}\hat{X}^\mathsf{T} Q_\gamma X r - \frac{2}{T\hat{T}}\hat{r}^\mathsf{T}\hat{X}^\mathsf{T} Q_\gamma X r \\ &= \lim_{\gamma\downarrow 0} \frac{1}{\hat{T}}\|\hat{r}\|^2 + \frac{1}{T^2\hat{T}} r^\mathsf{T} \tilde{Q}_\gamma X^\mathsf{T}\hat{X}\hat{X}^\mathsf{T} X \tilde{Q}_\gamma r - \frac{2}{T\hat{T}}\hat{r}^\mathsf{T}\hat{X}^\mathsf{T} X \tilde{Q}_\gamma r\end{aligned}$$

with $\hat{E}_\eta(u, r; \hat{u}, \hat{r})$ defined in (5).

If $n < T$, Q_γ is well-defined in the limit $\gamma \downarrow 0$, while if instead $n \geq T$, then one may observe that $X^\mathsf{T} Q_\gamma = \tilde{Q}_\gamma X^\mathsf{T}$ with $\tilde{Q}_\gamma$ having well defined limit as $\gamma \downarrow 0$.

Technically, estimating $\hat{E}$ requires to retrieve, in a similar fashion as for Theorem 2, a deterministic approximation of quantities of the type $Q_\gamma X$ and $X^\mathsf{T} Q_\gamma B Q_\gamma X = \tilde{Q}_\gamma X^\mathsf{T} B X \tilde{Q}_\gamma$ for B a matrix independent of X. We precisely obtain the following result.

Theorem 9 (Second order deterministic equivalent) *Let Assumptions 1–2 hold and let $B \in \mathbb{R}^{n\times n}$ be a deterministic symmetric matrix of bounded spectral norm. Then, recalling the notations of Theorem 2, for every $\gamma > 0$,*

$$\begin{aligned}Q_\gamma \frac{1}{\sqrt{T}} X &\leftrightarrow \bar{Q}_\gamma A (I_n + \eta^2 R_\gamma)^{-1} \\ \frac{1}{T} X^\mathsf{T} Q_\gamma B Q_\gamma X &\leftrightarrow \eta^2\gamma^2 \bar{\tilde{Q}}_\gamma G_\gamma^{[B]} \bar{\tilde{Q}}_\gamma + (I_n + \eta^2 R_\gamma)^{-1} A^\mathsf{T} \bar{Q}_\gamma \left[B + \tilde{G}_\gamma^{[B]}\right] \bar{Q}_\gamma A (I_n + \eta^2 R_\gamma)^{-1}\end{aligned}$$

where $G_\gamma^{[B]} \in \mathbb{R}^{T\times T}$ and $\tilde{G}_\gamma^{[B]} \in \mathbb{R}^{n\times n}$ are solutions to the system of equations

$$\begin{aligned}G_\gamma^{[B]} &= \left\{ \frac{1}{T} \operatorname{tr}\left(S_{i-j} \bar{Q}_\gamma \left[B + \tilde{G}_\gamma^{[B]}\right] \bar{Q}_\gamma \right) \right\}_{i,j=1}^{T} \\ \tilde{G}_\gamma^{[B]} &= \sum_{q=-\infty}^{\infty} \eta^4\gamma^2 \frac{1}{T} \operatorname{tr}\left(J^q \bar{\tilde{Q}}_\gamma G_\gamma^{[B]} \bar{\tilde{Q}}_\gamma \right) S_q.\end{aligned}$$

With these results at hand, we may then determine limiting approximations of the test mean-square error under both $n < T$ and $n > T$ regimes. As in Section 2.1, one may observe here that, under Assumption 3 with, say $c < 1$, $\eta^4 G_\gamma^{[B]}$ and $\tilde{G}_\gamma^{[B]}$ both have well defined limits as $\gamma \downarrow 0$ which we shall subsequently refer to as $\mathcal{G}^{[B]}$ and $\tilde{\mathcal{G}}^{[B]}$, respectively, and the symmetrical result holds for $c > 1$. Precisely, we have the following result.

Proposition 10 (Small γ limit of Theorem 9) *Let Assumptions 1–3 hold and let $B \in \mathbb{R}^{n\times n}$ be a deterministic symmetric matrix of bounded spectral norm. For all large n, define $\mathcal{G}^{[B]}$ and $\tilde{\mathcal{G}}^{[B]}$ a pair of solutions of the system*

$$\begin{aligned}\mathcal{G}^{[B]} &= c \left\{ \frac{1}{n} \operatorname{tr}\left(S_{i-j} \left(\boldsymbol{\delta}_{c>1} I_n + \tilde{\mathcal{R}} \right)^{-1} \left[B + \tilde{\mathcal{G}}^{[B]}\right] \left(\boldsymbol{\delta}_{c>1} I_n + \tilde{\mathcal{R}} \right)^{-1} \right) \right\}_{i,j=1}^{T} \\ \tilde{\mathcal{G}}^{[B]} &= \sum_{q=-\infty}^{\infty} \frac{1}{T} \operatorname{tr}\left(J^q (\boldsymbol{\delta}_{c<1} I_T + \mathcal{R})^{-1} \mathcal{G}^{[B]} (\boldsymbol{\delta}_{c<1} I_T + \mathcal{R})^{-1} \right) S_q.\end{aligned}$$

Then, with the definitions of Theorem 9, we have the following results.

1. *If $c < 1$, then in the limit $\gamma \downarrow 0$, $\eta^4 G_\gamma^{[B]} \to \mathcal{G}^{[B]}$ and $\tilde{G}_\gamma^{[B]} \to \tilde{\mathcal{G}}^{[B]}$.*
2. *If $c > 1$, then in the limit $\gamma \downarrow 0$, $\gamma^2 G_\gamma^{[B]} \to \mathcal{G}^{[B]}$ and $\tilde{G}_\gamma^{[B]} \to \tilde{\mathcal{G}}^{[B]}$.*

Proposition 10 will be exploited on the deterministic matrix $\frac{1}{\hat{T}}\mathrm{E}[\hat{X}\hat{X}^\mathsf{T}] = \eta^2 S_0 + \hat{A}\hat{A}^\mathsf{T}$. Rather than taking $B = \eta^2 S_0 + \hat{A}\hat{A}^\mathsf{T}$, which would induce an implicit dependence of $\mathcal{G}^{[B]}$ and $\tilde{\mathcal{G}}^{[B]}$ on η^2, we shall instead split $\eta^2 S_0 + \hat{A}\hat{A}^\mathsf{T}$ into η^2 times S_0 and $\hat{A}\hat{A}^\mathsf{T}$. Noticing then that $\mathcal{G}^{[\hat{A}\hat{A}^\mathsf{T}]}$ is asymptotically the same as $\mathcal{G}^{[0]}$, with 0 the all zero matrix, we may then obtain an approximation for the test mean square error. Prior to this, we need the following growth control assumptions.

Assumption 4 (Random Matrix Regime for Test Data) *The following conditions hold:*

1. $\limsup_n n/\hat{T} < \infty$
2. $\limsup_n \|\hat{A}\hat{A}^\mathsf{T}\| < \infty$.

Note in passing here that the $\min(T, \hat{T})$ first columns of $\hat{M} \in \mathbb{R}^{n \times \hat{T}}$ in the definition of $\hat{A}$ and $M \in \mathbb{R}^{n \times T}$ in the definition of A are identical. As such, only $\hat{U}$ actually particularizes the data matrix $\hat{A}$.

With this condition, we have the following corollary of Theorem 9.

Corollary 11 (Test MSE) *Let Assumptions 1–4 hold and let $\hat{r} \in \mathbb{R}^{\hat{T}}$ be a vector of Euclidean norm $O(\sqrt{\hat{T}})$. Then, as $n \to \infty$, both for $c < 1$ and $c > 1$, we have, with the notations of Propositions 4–10,*

$$\hat{E}_\eta(u, r; \hat{u}, \hat{r}) \leftrightarrow \left\| \frac{1}{\eta^2\sqrt{T}} \hat{A}^\mathsf{T} \mathcal{Q} A (\boldsymbol{\delta}_{c<1} I_T + \mathcal{R})^{-1} r - \frac{1}{\sqrt{\hat{T}}} \hat{r} \right\|^2 + \frac{1}{T} r^\mathsf{T} \tilde{\mathcal{Q}} \mathcal{G} \mathcal{G} \tilde{\mathcal{Q}} r + \frac{1}{\eta^2 T} r^\mathsf{T} (\boldsymbol{\delta}_{c<1} I_T + \mathcal{R})^{-1} A^\mathsf{T} \mathcal{Q} \left[S_0 + \tilde{\mathcal{G}} \right] \mathcal{Q} A (\boldsymbol{\delta}_{c<1} I_T + \mathcal{R})^{-1} r \tag{6}$$

where $\mathcal{G} \equiv \mathcal{G}^{[S_0]}$ and $\tilde{\mathcal{G}} \equiv \tilde{\mathcal{G}}^{[S_0]}$.

The form of Corollary 11 is more involved than that of Corollary 5 but is nonetheless quite interpretable. To start with, observe that $\mathcal{G}$ and $\tilde{\mathcal{G}}$ are again only function of W and therefore quantify the network connectivity only. Then, note that only the first right-hand side term of the approximation of $\hat{E}_\eta(u, r; \hat{u}, \hat{r})$ depends on $\hat{u}$ and $\hat{r}$. As such, the quality of the learned task relies mostly on this term.

If $c = 0$, for all B, $\mathcal{G}^{[B]} = 0$ and $\tilde{\mathcal{G}}^{[B]} = 0$, so we have here the simplified expression

$$\hat{E}_\eta(u, r; \hat{u}, \hat{r}) \leftrightarrow \left\| \frac{1}{\sqrt{T}} \hat{A}^\mathsf{T} \left(\eta^2 S_0 + AA^\mathsf{T} \right)^{-1} A r - \frac{1}{\sqrt{\hat{T}}} \hat{r} \right\|^2 + \frac{1}{T} r^\mathsf{T} A^\mathsf{T} \left(\eta^2 S_0 + AA^\mathsf{T} \right)^{-2} A r.$$

Some remarks are in order to appreciate these results.

Remark 12 (Noiseless case) *As a follow-up on Remark 6, note that some alternative approaches to ESN normalization assume instead that $\eta = 0$ but that ω is taken to be the* regularized *least-square (or ridge-regression) estimator $\omega = X(X^\mathsf{T}X + \gamma I_T)^{-1} r$ with $\gamma > 0$. In this case, it is easily seen that the corresponding mean-square error performance in training is given by $E^\gamma(u, r) \equiv \gamma^2 \frac{1}{T} r^\mathsf{T} \tilde{Q}_\gamma^2 r$, which is precisely*

$$E^\gamma(u, r) = \frac{1}{T} r^\mathsf{T} \left(I_T + \frac{1}{\gamma} U^\mathsf{T} \left\{ m^\mathsf{T} (W^i)^\mathsf{T} W^j m \right\}_{i,j=0}^{T-1} U \right)^{-2} r.$$

It is interesting to parallel this (exact) expression to the approximation (4) *in which the noise variance* η^2 *plays the role of the regularization* γ*, but (i) where the two additional quantities* $\mathcal{R}$ *and* $\tilde{\mathcal{R}}$ *are present, and (ii) where the power factor of the matrix inverse is* 1 *in place of* 2*. As for the testing performance, we are here comparing Corollary 11 to the noiseless regularized MSE*

$$E^\gamma(u,r;\hat{u},\hat{r}) = \left\| \frac{1}{\sqrt{T}} \hat{A}^\mathsf{T} \left(\gamma I_n + AA^\mathsf{T} \right)^{-1} Ar - \frac{1}{\sqrt{\hat{T}}} \hat{r} \right\|^2 .$$

This is again easily paralleled with the first right-hand side term in (6) *which, for say* $c < 1$*, reads*

$$\left\| \frac{1}{\sqrt{T}} \hat{A}^\mathsf{T} \left(\eta^2 \tilde{\mathcal{R}} + A(I_T + \mathcal{R})^{-1} A^\mathsf{T} \right)^{-1} A(I_T + \mathcal{R})^{-1} r - \frac{1}{\sqrt{\hat{T}}} \hat{r} \right\|^2 .$$

Again, it is clear that η^2 *plays a similar role as that of* γ*, and that the matrices* $\mathcal{R}$ *and* $\tilde{\mathcal{R}}$ *capture the behavior of the in-network noise.*

Remark 12 suggests that internal noise plays a similar role to ridge normalization and that both lead to similar MSE performances. This being said, both regularizations behave strikingly differently in practice. While ridge-regularization provides a deterministic network output for given input vector u, internal noise instead induces *random independent outputs* for any two feeds of the network by the same vector u. Since all such random outputs have similar MSE performance (for sufficiently large network sizes), this may be a preferable choice in practice to avoid deterministic arbitrary mappings.

3. Applications

In this section, we shall further estimate the results of Corollary 5 and Corollary 11 in specific settings for the network connectivity matrix W and the input weights m. By leveraging specific properties of certain stochastic models for W (such as invariance by orthogonal matrix product or by normality), the results of Section 2 will be greatly simplified, by then providing further insights on the network performance.

3.1 Bi-orthogonally invariant W

We first consider the scenario where W is random with distribution invariant to left- and right-multiplication by orthogonal matrices, which we refer to as *bi-orthogonal invariance*. Precisely, in singular-value decomposition form, we shall write $W = U\Omega V^\mathsf{T}$, where U, V, and Ω are independent and U, V are real Haar distributed (that is, orthogonal with bi-orthogonally invariant distribution) and shall impose that the eigenvalues of W remain bounded by $\sigma < 1$ for all large n. Two classical examples of such a scenario are (i) W is itself a scaled Haar matrix, in which case $\Omega = \sqrt{\sigma} I_n$ and the eigenvalues of W all have modulus σ, or (ii) W has independent $\mathcal{N}(0, \sigma^2)$ entries, in which case, according to standard random matrix results, for any $\varepsilon > 0$, the eigenvalues of W have modulus less than $\sigma + \varepsilon$ for all large n almost surely and W is clearly orthogonally invariant by orthogonal invariance of the real multivariate Gaussian distribution.

In this scenario, one can exploit the fact (arising for instance from free probability considerations (Biane, 2003)) that, for all $i \neq j$ fixed, the moments $\frac{1}{n} \operatorname{tr} W^i (W^j)^\mathsf{T}$ vanish as $n \to \infty$. In our setting, i and j may however be growing with n, but then the fact that $\rho(W^i) \leq \sigma^i$ shall easily ensure an exponential decay of these moments. All in all, in the large n setting, only the first few moments $\frac{1}{n} \operatorname{tr} W^i (W^i)^\mathsf{T}$, $i = 1, 2, \ldots$, do not vanish. Although the implication is not immediate, this remark leads naturally to the intuition that

the Toeplitz matrix $\mathcal{R}$ defined in Proposition 4 should be diagonal and thus proportional to the identity matrix.

At this point, we need to differentiate the cases where $c < 1$ and $c > 1$.

3.1.1 Case $c < 1$

Based on the remarks above, we may explicitly solve for $\mathcal{R}$ and $\tilde{\mathcal{R}}$ to find that, in the large n limit

$$\begin{aligned}\mathcal{R} &\leftrightarrow \frac{c}{1-c} I_T \\ \tilde{\mathcal{R}} &\leftrightarrow (1-c) S_0 \\ \mathcal{G}^{[B]} &\leftrightarrow \frac{c}{(1-c)^3} \frac{1}{n} \operatorname{tr}(S_0^{-1} B) I_T \\ \tilde{\mathcal{G}}^{[B]} &\leftrightarrow \frac{c}{1-c} \frac{1}{n} \operatorname{tr}(S_0^{-1} B) S_0.\end{aligned}$$

Replacing in the expressions of both Corollaries 5–11, we obtain the further corollary

Corollary 13 (Orthogonally invariant case, $c < 1$) *Let W be random and left and right independently orthogonally invariant. Then, under Assumptions 1–4 and with $c < 1$, the following hold*

$$\begin{aligned}E_\eta(u, r) &\leftrightarrow (1-c) \frac{1}{T} r^\mathsf{T} \left(I_T + \frac{1}{\eta^2} U^\mathsf{T} D U \right)^{-1} r \\ \hat{E}_\eta(u, r; \hat{u}, \hat{r}) &\leftrightarrow \left\| \frac{1}{\eta^2 \sqrt{T}} \hat{U}^\mathsf{T} \hat{D} U \left(I_T + \frac{1}{\eta^2} U^\mathsf{T} D U \right)^{-1} r - \frac{1}{\sqrt{\hat{T}}} \hat{r} \right\|^2 \\ &\quad + \frac{1}{1-c} \frac{1}{T} r^\mathsf{T} \left(I_T + \frac{1}{\eta^2} U^\mathsf{T} D U \right)^{-1} r - \frac{1}{T} r^\mathsf{T} \left(I_T + \frac{1}{\eta^2} U^\mathsf{T} D U \right)^{-2} r\end{aligned}$$

where we defined $D \equiv \left\{ m^\mathsf{T} (W^i)^\mathsf{T} S_0^{-1} W^j m \right\}_{i,j=0}^{T-1}$ and $\hat{D} \equiv \left\{ m^\mathsf{T} (W^i)^\mathsf{T} S_0^{-1} W^j m \right\}_{i,j=0}^{\hat{T}-1, T-1}$.

We see here that the matrix D plays a crucial role in the ESN performance. First, from its Gram structure and the positive definiteness of S_0, D is symmetric and nonnegative definite. This matrix has an exponential decaying profile down its rows and columns. As such, the dominating coefficients of the matrix $U^\mathsf{T} D U$ lie in its upper-left corner. Recalling that the j-th column of $\sqrt{T} U^\mathsf{T}$ is $\{u_{i-j}\}_{i=1}^T$, $U^\mathsf{T} D U$ is essentially a linear combination of the outer products $\{u_{i-j}\}_{i=1}^T (\{u_{i-j'}\}_{i=1}^T)^\mathsf{T}$ for small j, j', that is of combinations of (outer-products of) short-time delayed versions of the input vector u.

Note that, although we do not provide a rigorous proof of this fact, by the standard universality property of random matrix results, Corollary 13 is equally valid if W is chosen to be a matrix with i.i.d. zero mean and variance σ^2 *non-necessarily Gaussian* entries. In particular, it extends to the case of (properly recentered) matrices W with Bernoulli entries of Bernoulli parameter not scaling with n. In the regime under consideration, the asymptotic performance equivalence suggests that (here non sparse) Bernoulli random matrices have no particular advantage when compared to Gaussian random matrices, which is somewhat opposed to what is sometimes suggested in the ESN literature.[3]

Now, it is interesting to particularize the vector m and study its impact on D. It may be thought that taking m to be one of the dominant eigenvectors of W could drive the

3. However, sparse connectivity matrices prevail over non sparse ones in the literature, in which case our claim no longer holds.

inputs towards interesting memory-capacity levels of W; this aspect is discussed in (Ganguli et al., 2008) where it is found that such an m maximizes the integrated Fisher-memory curve. If such a real eigenvector having eigenvalue close to σ exists, then we would find that $D_{ij} \simeq \sigma^{i+j} m^\mathsf{T} S_0^{-1} m$ and thus D would essentially be a rank-one matrix. As we shall discuss below, this would lead to extremely bad MSE performance in general.

If instead m is chosen deterministic or random independent of W with say $\|m\| = 1$ (or tending to one) for simplicity, then by the trace lemma (Bai and Silverstein, 2009, Lemma B.26), one can show that $m^\mathsf{T}(W^i)^\mathsf{T} S_0^{-1} W^j m \leftrightarrow \frac{1}{n} \operatorname{tr} W^j (W^i)^\mathsf{T} S_0^{-1}$. According to our earlier discussion, this quantity vanishes for all $i \neq j$ as $n \to \infty$, and thus D would now essentially be diagonal. Besides, it is clear that $\operatorname{tr} D \leftrightarrow 1$ and thus D here plays the role of affecting a short-term memorization ability, that can be seen as a total load 1, to the successive delayed versions of u. In particular, from our definition in Remark 7, we have precisely here

$$\mathrm{MC}(\tau) = \frac{1}{1-c} \liminf_n \frac{1}{n} \operatorname{tr} \left(W^\tau (W^\tau)^\mathsf{T} S_0^{-1} \right)$$

which, for the chosen m, is precisely the Fisher memory curve (Ganguli et al., 2008), up to the factor $1 - c$.

Remark 14 (Haar W and independent m) *For $W = \sigma Z$ with Z Haar distributed (orthogonal and orthogonally invariant) and m independent of Z and of unit norm, D is asymptotically diagonal and we find precisely*

$$D_{ii} \leftrightarrow (1 - \sigma^2)\sigma^{2(i-1)}$$

and in particular

$$\mathrm{MC}(\tau) = \frac{1 - \sigma^2}{1 - c} \sigma^{2\tau}.$$

Remark 14 can be extended to design an interesting multiple memory-mode network as follows.

Remark 15 (Multiple memory modes) *Take W to be the block diagonal matrix $W = \operatorname{diag}(W_1, \ldots, W_k)$ where, for $j = 1, \ldots, k$, $W_j = \sigma_j Z_j$, $\sigma_j > 0$, and $Z_j \in \mathbb{R}^{n_j \times n_j}$ is Haar distributed, independent across j. Take then m independent of W with unit norm. Also assume that $n_j/n \to c_j > 0$ as $n \to \infty$ and $\sum_j n_j = n$. Then we find that*

$$D_{ii} \leftrightarrow \frac{\sum_{j=1}^k c_j \sigma_j^{2(i-1)}}{\sum_{j=1}^k c_j (1 - \sigma_j^2)^{-1}}$$

and in particular, with $\mathrm{MC}(\tau)$ defined in Remark 7,

$$\mathrm{MC}(\tau) = \frac{1}{1-c} \frac{\sum_{j=1}^k c_j \sigma_j^{2\tau}}{\sum_{j=1}^k c_j (1 - \sigma_j^2)^{-1}}.$$

A graph of $\mathrm{MC}(\tau)$ for $k = 3$ is depicted in Figure 1, where it clearly appears that the memory curve follows successively each one of the three modes, giving in particular more weight to short-term past inputs at first, and then smoothly providing increasingly more importance to longer term past inputs. This is reminiscent of the long short-term memory framework devised in (Xue et al., 2007).

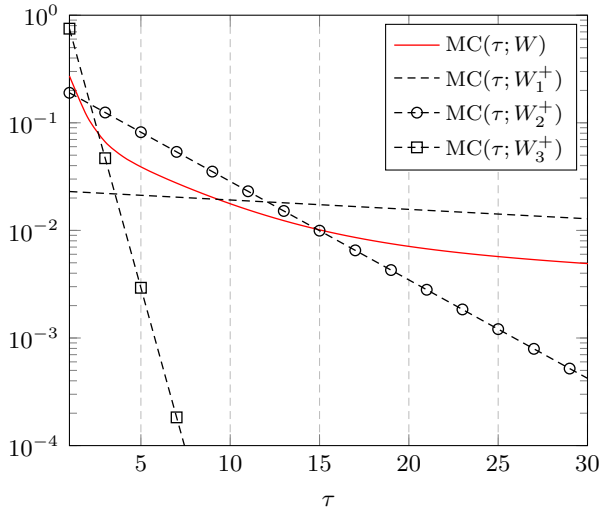

Figure 1: Memory curve for $W = \mathrm{diag}(W_1, W_2, W_3)$, $W_j = \sigma_j Z_j$, $Z_j \in \mathbb{R}^{n_j \times n_j}$ Haar distributed, $\sigma_1 = .99$, $n_1/n = .01$, $\sigma_2 = .9$, $n_2/n = .1$, and $\sigma_3 = .5$, $n_3/n = .89$. The matrices W_i^+ are defined by $W_i^+ = \sigma_i Z_i^+$, with $Z_i^+ \in \mathbb{R}^{n \times n}$ Haar distributed.

It is next interesting to study Corollary 13 more deeply. Let us first assume that the task to be performed, both in training and testing, consists in retrieving a mere linear combination of latest past inputs $u_t, u_{t-1}, \ldots, u_{t-(k-1)}$ for k fixed. Then we may write $r = \sqrt{T} U^\mathsf{T} b$ for some vector $b \in \mathbb{R}^T$ with $b_j = 0$ for all $j \geq k$. We then have

$$E_\eta(u, r) \leftrightarrow (1-c) b^\mathsf{T} U \left(I_T + \frac{1}{\eta^2} U^\mathsf{T} D U \right)^{-1} U^\mathsf{T} b.$$

For D positive diagonal with exponential decaying profile, D^{-1} is extremely ill-conditioned and may only be used with extreme care. However, for k fixed, $D^{-\frac{1}{2}} b$ is well behaved as its norm is bounded by $\|b\| D_{k-1,k-1}^{-\frac{1}{2}}$. We may thus write $b = D^{\frac{1}{2}} (D^{-\frac{1}{2}} b)$ to obtain, after basic algebraic manipulations

$$E_\eta(u, r) \leftrightarrow \eta^2 (1-c) (D^{-\frac{1}{2}} b)^\mathsf{T} \frac{1}{\eta^2} D^{\frac{1}{2}} U U^\mathsf{T} D^{\frac{1}{2}} \left(I_T + \frac{1}{\eta^2} D^{\frac{1}{2}} U U^\mathsf{T} D^{\frac{1}{2}} \right)^{-1} (D^{-\frac{1}{2}} b).$$

Since $\|A(I+A)^{-1}\| \leq 1$ for any symmetric nonnegative definite matrix A, we thus conclude that, for every $\eta, \varepsilon > 0$, $E_\eta(u, r) \leq (1-c)\eta^2 b^\mathsf{T} D^{-1} b + \varepsilon$ for all large n almost surely. Thus, for sufficiently large n, $E_\eta(u, r)$ can be made arbitrarily small in the limit where $\eta \to 0$ and thus the task can be performed accurately. As for $\hat{E}_\eta$, note that, since $\hat{D}$ and D are essentially zero away from the upper left corner and otherwise equal, if $\hat{r} = \sqrt{\hat{T}} \hat{U} \hat{b}$, for $\hat{b} \in \mathbb{R}^{\hat{T}}$ having the same first k entries as b and zeroes next, then we find

$$\begin{aligned} \hat{E}_\eta(u, r; \hat{u}, \hat{r}) \leftrightarrow \frac{\eta^2}{1-c} (D^{-\frac{1}{2}} b)^\mathsf{T} \frac{1}{\eta^2} D^{\frac{1}{2}} U U^\mathsf{T} D^{\frac{1}{2}} \left(I_T + \frac{1}{\eta^2} D^{\frac{1}{2}} U U^\mathsf{T} D^{\frac{1}{2}} \right)^{-1} (D^{-\frac{1}{2}} b) \\ + (D^{-\frac{1}{2}} b)^\mathsf{T} \left(I_T + \frac{1}{\eta^2} D^{\frac{1}{2}} U U^\mathsf{T} D^{\frac{1}{2}} \right)^{-1} D^{\frac{1}{2}} \Delta D^{\frac{1}{2}} \left(I_T + \frac{1}{\eta^2} D^{\frac{1}{2}} U U^\mathsf{T} D^{\frac{1}{2}} \right)^{-1} (D^{-\frac{1}{2}} b) \quad (7) \end{aligned}$$

where $\Delta \equiv [\hat{U}\hat{U}^\mathsf{T}]_{T\times T} - UU^\mathsf{T}$, with the operator $[X]_{T\times T}$ extending (or reducing) X to a $T \times T$ matrix by filling it with zeroes (or discarding last rows and columns). Note here that, for $U = \hat{U}$, $\Delta = 0$ and we find that

$$\hat{E}_\eta(u, r; u, r) \leftrightarrow \frac{1}{(1-c)^2} E_\eta(u, r). \tag{8}$$

When $\Delta \neq 0$, observe first that $\|(I_T + \eta^{-2} D^{\frac{1}{2}} UU^\mathsf{T} D^{\frac{1}{2}})^{-1}\| \leq 1$ and thus $\hat{E}_\eta(u, r; \hat{u}, \hat{r})$ remains bounded. Now, with a more subtle analysis, note that, since the product $BD^{-\frac{1}{2}}b$ for any matrix B only concerns the first k columns of B, the behavior of $\hat{E}_\eta$ as $\eta \to 0$ merely depends on the behavior of the top-left $k \times k$ submatrix of $(I_T + \eta^{-2} D^{\frac{1}{2}} UU^\mathsf{T} D^{\frac{1}{2}})^{-1}$. A block matrix inverse then reveals that the second right-hand side term of (7) goes to zero as $\eta \to 0$ provided that the k-th largest eigenvalue of $D^{\frac{1}{2}} UU^\mathsf{T} D^{\frac{1}{2}}$ remains away from zero as $T \to \infty$. From the structure of U, we thus conclude that, for $\hat{E}_\eta$ to vanish as $\eta \to 0$, it is sufficient for the vector u to be sufficiently "diverse" in its constituents (that is, so that the first columns of U remain linearly independent). An obvious counter-example is when the sequence $\ldots, u_{-1}, u_0, u_1, \ldots \in \mathbb{R}$ is periodic of period less than k. Note that the specific choice of $\hat{u}$ does not alter this behavior.

The discussion above leads to interesting practical considerations that may help improve the design of an ESN.

Remark 16 (Selecting W based on delayed correlations) *Note that, in the aforementioned formulas, the quantity $b^\mathsf{T} D^{-1} b$ with b defined by $r = U^\mathsf{T} b$ appears as a fundamental quantity bounding the training and testing MSE. In practical settings where r is not a pure linear combinations of delayed versions of u, it may nonetheless be useful to obtain an estimate $\hat{b}$ of the closest approximation of r by delays of u, in such a way that $\hat{b}^\mathsf{T} D^{-1} b$ be small. One may for instance let*

$$\hat{b} = (UU^\mathsf{T} + \gamma I_T)^{-1} Ur$$

for some regularization parameter $\gamma \geq 0$ (if needed), and parametrize W so that $\hat{b}^\mathsf{T} D^{-1} \hat{b}$ is minimal. For instance, if $\hat{b}_i = \alpha^{i-1}$ for some $\alpha \in (-1, 1)$, it is easily shown that an optimal choice for $W = \sigma Z$ with Z Haar is to take $\sigma^2 = |\alpha|$. This scenario is illustrated in Figure 2, where the theoretical approximations for the testing and training normalized MSE are depicted for various choices of σ^2. For less obvious values of $\hat{b}$, a more elaborate multi-memory matrix W, as introduced in Remark 15, can be used, with proper setting of the parameters n_i and σ_i.

Remark 17 (Memory Capacity for Stationary Inputs) *Let W be orthogonally invariant and m random, so that D is diagonal in the limit. Further assume the sequence u is an auto-regressive Gaussian process, so that we may write $u = C^{\frac{1}{2}} \tilde{u}$ with $\tilde{u}$ having independent zero mean unit variance Gaussian entries and C a Toeplitz covariance matrix with $C_{ab} = q^{|b-a|}$ for some $q \in [0, 1)$. Then, for the τ-delay memory task, i.e., $r_t = u_{t-\tau}$ with τ fixed, we find that*

$$E_\eta(u, r) \leftrightarrow \eta^2 \frac{1-c}{D_{\tau+1,\tau+1}} \left[1 - \left[\left(I_T + \frac{1}{\eta^2} \left\{ \sqrt{D_{ii}} q^{|i-j|} \sqrt{D_{jj}} \right\}_{i,j=0}^{T-1} \right)^{-1} \right]_{\tau+1,\tau+1} \right].$$

Since $q < 1$, the matrix $\{q^{|i-j|}\}_{ij}$ has its smallest eigenvalue asymptotically far from zero (see e.g., (Gray, 2006) for arguments) so that the right-hand side inner bracket vanishes as $\eta^2 \to 0$ and we thus have, for small η^2

$$\eta^{-2} E_\eta(u, r) \leftrightarrow \frac{1-c}{D_{\tau+1,\tau+1}} + o(\eta^2).$$

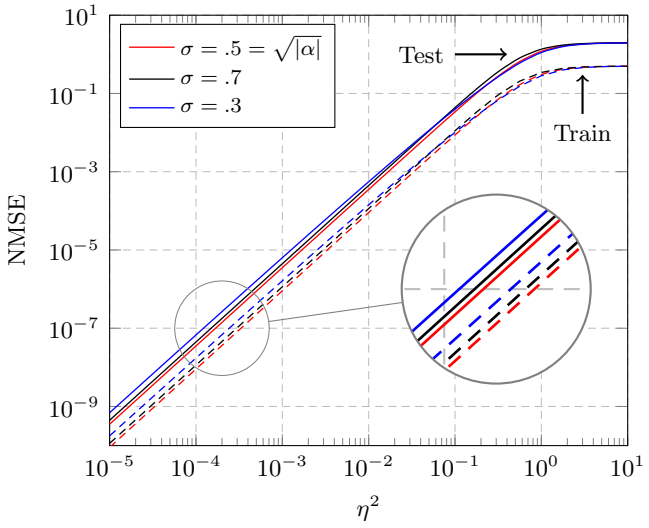

Figure 2: Optimal σ choice for $r_t = \sum_{i\geq 0} u_{t-i} b_i$, $b_i = \alpha^{i-1}$ with $\alpha = -.25$, u i.i.d. zero mean Gaussian, W Haar distributed, $n = 200$, $T = \hat{T} = 400$.

As a consequence, the memory task is performed irrespective of the smoothness of u, so that u can be assumed composed of i.i.d. elements (i.e., $q = 0$). Observe that this leads to the same performance as the memory task considering $u = \sqrt{T}\boldsymbol{\delta}_0$ defining the memory capacity in Remark 7. Of course, if instead $q = 1$, then the matrix in curly brackets would have unit rank and the previous conclusions would fail (in this case u is a constant vector).

Expression (8) also provides us an opportunity to open a short parenthesis on the effect of c on the training and testing MSE. From Corollary 13, it appears that, while E_η is minimal for $c = 1$, $\hat{E}_\eta$ is minimal for $c = 0$. The former observation is clear from the fact that ω is a least-square regressor, but the latter observation is less trivial. As a matter of fact, note that, even if $\hat{U} = U$ and $\hat{r} = r$, in the limit of $\eta > 0$ fixed and $c \to 1$, $\hat{E}_\eta$ becomes arbitrarily large. The reason for this seemingly counter-intuitive effect (after all, we merely ask the ESN to reproduce the exact learned sequence) lies in the fact that ω is built upon the network noise realization during training, while during testing a new noise realization is produced. As such, training an ESN of size almost equal to T produces dramatic effects on testing. However, this has the positive effect of strongly reducing over fitting. Of course, in practical settings, there exists an interplay between η^2 that drives both MSE's to zero as $\eta \to 0$ and c that reduces overfitting as it tends to 1.

Coming back to the approximations of E_η and $\hat{E}_\eta$, note now that if D is a rank-one matrix, then we may write $D = dd^\mathsf{T}$ for some vector $d \in \mathbb{R}^T$ having exponentially vanishing entries. In this case, we find, again after standard algebraic calculus, that

$$E_\eta(u, r) \leftrightarrow (1 - c) \left(\frac{1}{T} \|r\|^2 - \frac{\frac{1}{T} |d^\mathsf{T} U r|^2}{\eta^2 + |d^\mathsf{T} U|^2} \right).$$

Taking as above $r = \sqrt{T} U b$, this is $E_\eta(u, r) \leftrightarrow (1-c) \left(b^\mathsf{T} U U^\mathsf{T} b - \frac{|d^\mathsf{T} U U^\mathsf{T} b|^2}{\eta^2 + d^\mathsf{T} U U^\mathsf{T} d} \right)$. By Cauchy-Schwarz inequality, this quantity, even in the limit $\eta^2 \to 0$, cannot vanish unless $b = d$.

As such, the ESN will only adequately fulfill a single task, which depends on the network configuration itself through d. A similar reasoning can be made on $\hat{E}_\eta$ revealing the same shortcomings.

As a practical example, we provide in Figure 3 Monte Carlo simulations versus theory curves of the training and testing performances of networks of $n = 200$ and $n = 400$ nodes, for training and testing times $T = \hat{T} = 2n$, on the Mackey Glass one-step ahead anticipation task (Glass and Mackey, 1979). The network is chosen to be the multi-memory model introduced in Remark 15 and following the description of Figure 1. The NMSE is defined here as the ratio between the MSE and the output vector squared norm $\|r\|^2/T$ or $\|\hat{r}\|^2/\hat{T}$. Simulations are run for a single W but different noise realizations and comparison is made against theory for either this W or its approximated asymptotic limit. Observe the extremely accurate match between theory and practice, with increasing precision as n, T grow large.

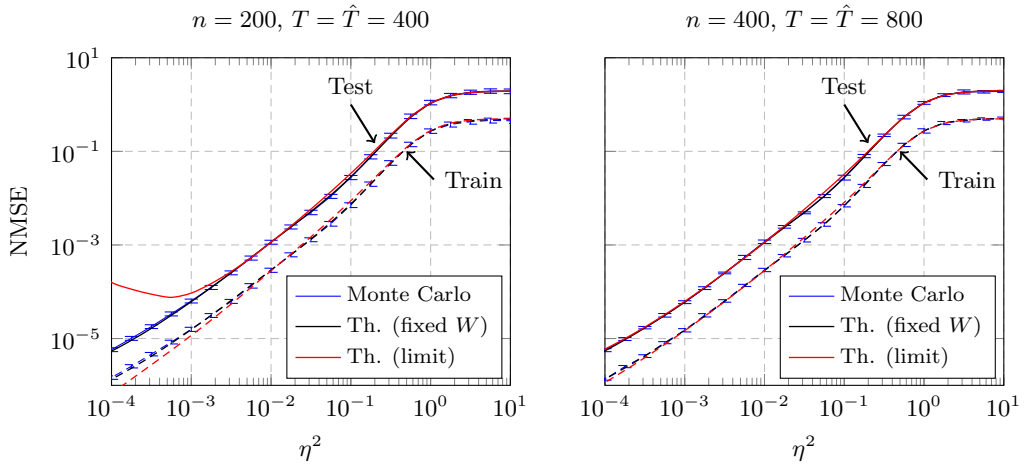

Figure 3: Training and testing (normalized) MSE for the Mackey Glass one-step ahead task, W fixed and defined as in Figure 1, $n = 200$, $T = \hat{T} = 400$ (left) and $n = 400$, $T = \hat{T} = 800$ (right). Comparison between Monte Carlo simulations (Monte Carlo), deterministic approximation assuming W fixed (Th. (fixed W)) as per Corollaries 5 and 11, and assuming W random in the large n limit (Th. (limit)) as per Corollary 13. Error bars indicate one standard deviation of the Monte Carlo simulations.

3.1.2 Case $c > 1$

The case $c > 1$ is slightly more involved as it does not lend itself to a purely explicit expression. Precisely, following the same steps as for $c < 1$, we find that in the large n

limit

$$\begin{aligned}
\mathcal{R} &\leftrightarrow \alpha I_T \\
\tilde{\mathcal{R}} &\leftrightarrow \frac{1}{\alpha} S_0 \\
\mathcal{G}^{[B]} &\leftrightarrow c\alpha^2 \frac{\frac{1}{n} \operatorname{tr} S_0 \left(\alpha I_n + S_0\right)^{-1} B \left(\alpha I_n + S_0\right)^{-1}}{1 - c\frac{1}{n} \operatorname{tr} S_0^2 \left(\alpha I_n + S_0\right)^{-2}} I_T \\
\tilde{\mathcal{G}}^{[B]} &\leftrightarrow c\frac{\frac{1}{n} \operatorname{tr} S_0 \left(\alpha I_n + S_0\right)^{-1} B \left(\alpha I_n + S_0\right)^{-1}}{1 - c\frac{1}{n} \operatorname{tr} S_0^2 \left(\alpha I_n + S_0\right)^{-2}} S_0
\end{aligned}$$

where $\alpha > 0$ is the unique solution to the equation

$$1 = c\frac{1}{n} \operatorname{tr} S_0 \left(\alpha I_n + S_0\right)^{-1}.$$

With these notations, we have the following counterpart to Corollary 13.

Corollary 18 (Orthogonally invariant case, $c > 1$) *Let W be random and left and right independently orthogonally invariant and let $\alpha > 0$ be the unique solution to $1 = c\frac{1}{n} \operatorname{tr} S_0 \left(\alpha I_n + S_0\right)^{-1}$. Then, under Assumptions 1–4 and with $c > 1$, the following holds*

$$\begin{aligned}
\hat{E}_\eta(u, r; \hat{u}, \hat{r}) \leftrightarrow &\left\| \eta^{-2} \hat{U}^\mathsf{T} \hat{D} U \left(I_T + \eta^{-2} U^\mathsf{T} D U\right)^{-1} \frac{r}{\sqrt{T}} - \frac{1}{\sqrt{\hat{T}}} \hat{r} \right\|^2 - \frac{1}{T} r^\mathsf{T} \left(I_T + \eta^{-2} U^\mathsf{T} D U\right)^{-1} r \\
&+ \frac{\frac{1}{T} r^\mathsf{T} \left(I_T + \eta^{-2} U^\mathsf{T} D U\right)^{-1} \left[I_T + \eta^{-2} U^\mathsf{T} D_2 U\right] \left(I_T + \eta^{-2} U^\mathsf{T} D U\right)^{-1} r}{1 - c\frac{1}{n} \operatorname{tr} S_0^2 \left(\alpha I_T + S_0\right)^{-2}}
\end{aligned}$$

where $D \equiv \left\{ m^\mathsf{T} (W^i)^\mathsf{T} (\alpha I_n + S_0)^{-1} W^j m \right\}_{i,j=0}^{T-1}$, $\hat{D} \equiv \left\{ m^\mathsf{T} (W^i)^\mathsf{T} (\alpha I_n + S_0)^{-1} W^j m \right\}_{i,j=0}^{\hat{T}-1,T-1}$, and $D_2 \equiv \left\{ m^\mathsf{T} (W^i)^\mathsf{T} (\alpha I_n + S_0)^{-1} S_0 (\alpha I_n + S_0)^{-1} W^j m \right\}_{i,j=0}^{T-1}$.

Of course here $E_\eta(u, r) = 0$.

Remark 19 (Haar W, random m for $c > 1$) *Although seemingly less tractable, for W following a Haar model, Corollary 18 takes a much simpler form. Indeed, for W and m as defined in Remark 14, we find that $\alpha = (c-1)(1-\sigma^2)^{-1}$ and $S_0 = (1-\sigma^2)^{-1} I_n$ which then leads to*

$$\begin{aligned}
\hat{E}_\eta(u, r; \hat{u}, \hat{r}) \leftrightarrow &\left\| (c\eta^2)^{-1} \hat{U}^\mathsf{T} \hat{D} U \left(I_T + (c\eta^2)^{-1} U^\mathsf{T} D U\right)^{-1} \frac{r}{\sqrt{T}} - \frac{1}{\sqrt{\hat{T}}} \hat{r} \right\|^2 \\
&+ \frac{1}{c-1} \frac{1}{T} r^\mathsf{T} \left(I_T + (c\eta^2)^{-1} U^\mathsf{T} D U\right)^{-1} r
\end{aligned}$$

where D is diagonal with $D_{ii} \equiv (1-\sigma^2)\sigma^{2(i-1)}$.

Aside from obtaining a shorter form expression for D and D_2, the multi memory model of Remark 15 does not lead to an explicit formulation as in Remark 19, but it is nonetheless instructive to observe the performance achieved on the Mackey Glass model from Figure 3, now in the setting where $c > 1$. This is depicted here in Figure 4.

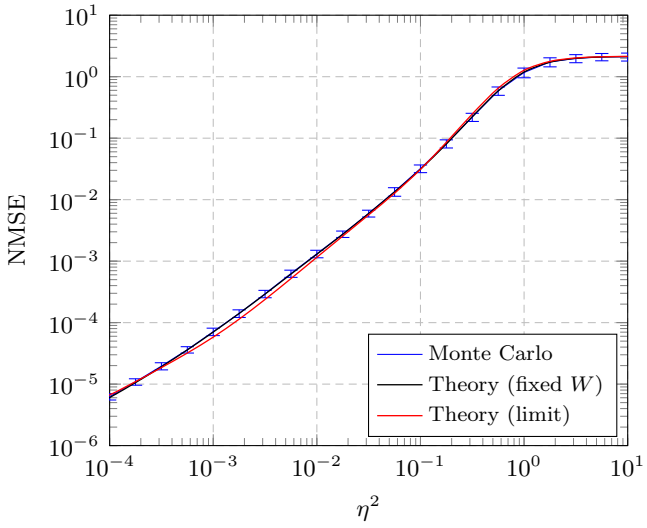

Figure 4: Testing (normalized) MSE for the Mackey Glass one-step ahead task, W fixed and defined as in Figure 1, $n = 400$, $T = \hat{T} = 200$. Error bars indicate one standard deviation of the Monte Carlo simulations.

3.2 Normal W

We now turn to the case of normal matrices. Let then W be normal (i.e., diagonalizable in orthogonal basis) and having an eigenvalue decomposition of the type $W = V\Lambda V^\mathsf{T}$ with V orthogonal and Λ diagonal with largest absolute entry less than one. For simplicity, we shall further assume that, as $n \to \infty$, the normalized counting measure of the diagonal elements of Λ ($n^{-1}\sum_i \boldsymbol{\delta}_{\Lambda_{ii}}$) converges in law to a probability measure μ. We do not make any assumption here on V.

For instance, real Gaussian Wigner matrices W, that is with i.i.d. zero mean variance $\frac{1}{4}\sigma^2$ Gaussian entries on and above the diagonal, and symmetrized below the diagonal, is an example of such a matrix. In this case, μ corresponds (almost surely) to the well-known semi-circular distribution, with density $\mu(d\lambda) = 2(\pi\sigma^2)^{-1}\sqrt{(\sigma^2-\lambda^2)^+}d\lambda$. Another example is when $\mu(d\lambda) = \frac{1}{2}[\boldsymbol{\delta}_\sigma + \boldsymbol{\delta}_{-\sigma}]d\lambda$, so that W is the sum of two (σ-scaled) projection matrices on orthogonal subspaces. In particular here, $W^2 = \sigma^2 I_n$, so that $W^{2k} = \sigma^{2k} I_n$ and $W^{2k+1} = \sigma^{2k} W$, for all $k \geq 0$.

Because of the symmetry property, it is no longer true that $\frac{1}{n}\operatorname{tr} W^i (W^j)^\mathsf{T} = \frac{1}{n}\operatorname{tr} W^{i+j}$ vanished for $i \neq j$, and we then obtain more involved results. To keep this discussion short and since the results take here more involved forms, we shall only deal here with the case $c < 1$ and focus on the training performance. In this case, solving Proposition 4 for $\mathcal{R}$ and $\tilde{\mathcal{R}}$, we have the following result. As $n \to \infty$, $\mathcal{R}$ has a limit (which for simplicity we keep calling $\mathcal{R}$) which is solution to

$$\mathcal{R}_{ab} = c \int \frac{t^{|a-b|}\mu(dt)}{\sum_{q=-\infty}^{\infty} \frac{1}{T}\operatorname{tr}(J^q(I_T + \mathcal{R})^{-1})t^{|q|}} \tag{9}$$

for all $a, b \in \{1, \ldots, T\}$. Remember that $\mathcal{R}$ is Toeplitz with fast decaying values off the diagonal, so that (9) is computationally easy to solve. Similar conclusions can be drawn on the matrices $\mathcal{G}^{[B]}$ and $\tilde{\mathcal{G}}^{[B]}$, that however do not lead to simple expressions.

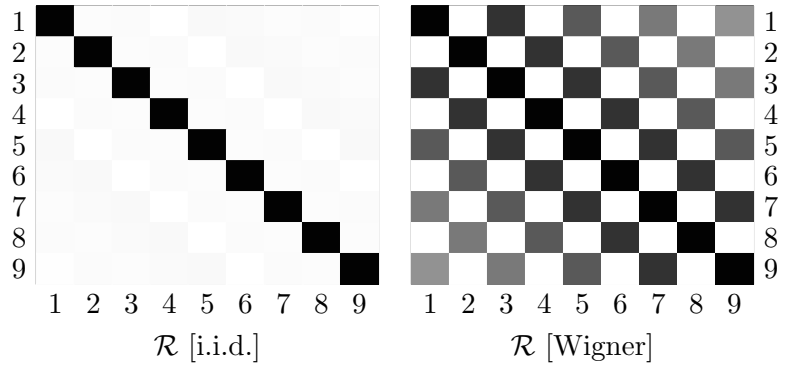

Figure 5: Upper 9×9 part of $\mathcal{R}$ for $c = 1/2$ and $\sigma = 0.9$ for W with i.i.d. zero mean Gaussian entries (left) and W Gaussian Wigner (right). Linear grayscale representation with black being 1 and white being 0.

Remark 20 (Symmetric μ) *An interesting scenario is when μ is symmetric, i.e., $\mu(-t) = \mu(t)$, which is the case of both aforementioned (Wigner and projection matrix) examples. From (9), we find in this case that $[\mathcal{R}]_{ab}$ is zero if $a - b$ is odd and positive if $a - b$ is even. As such, $\mathcal{R}$, takes the form of a checkerboard matrix. Figure 5 provides a representation of $\mathcal{R}$ in both normal and non-normal Gaussian W cases.*

Remark 21 (Projection W) *Let $W = V\Lambda V^\mathsf{T}$ with the normalized counting measure of Λ converging to $\mu(d\lambda) = \frac{1}{2}[\boldsymbol{\delta}_\sigma + \boldsymbol{\delta}_{-\sigma}]d\lambda$ and $c < 1$. Then, $\mathcal{R}_{ab} \leftrightarrow \sigma^{|b-a|} r_0 \boldsymbol{\delta}_{|b-a|\in 2\mathbb{N}}$ and $\tilde{\mathcal{R}} \leftrightarrow -(1-\sigma^2)^{-1} c r_0^{-1} I_n$ where*

$$r_0 = c\left(\sum_{q=-\infty}^{\infty} \frac{1}{T} \operatorname{tr} J^q \left(I_T + \mathcal{R}\right)^{-1}\right)^{-1}.$$

As a consequence, letting m be random, we find that

$$E_\eta(u, r) \leftrightarrow \frac{1}{T} r^\mathsf{T} \left(I_T + r_0 \left\{\sigma^{|j-i|}\boldsymbol{\delta}_{|j-i|\in 2\mathbb{N}}\right\}_{i,j=0}^{T-1} + \frac{r_0(1-\alpha^2)}{\eta^2 c} U^\mathsf{T} \left\{\sigma^{j+i}\boldsymbol{\delta}_{|j-i|\in 2\mathbb{N}}\right\}_{i,j=0}^{T-1} U\right)^{-1} r.$$

Note in particular that the matrix $\{\sigma^{j+i}\boldsymbol{\delta}_{|j-i|\in 2\mathbb{N}}\}_{i,j=0}^{T-1}$ can be decomposed as the sum of two matrices: (i) the rank-one matrix vv^T with $v = (1, 0, \alpha^2, 0, \alpha^4, \ldots)^\mathsf{T}$ and the diagonal matrix $\operatorname{diag}(0, \alpha^2, 0, \alpha^4, \ldots)$. Recalling that rank-one matrices in this position do not allow for efficient training (see the final discussions in Section 3.1, $c < 1$ case), only the diagonal component $\operatorname{diag}(0, \alpha^2, 0, \alpha^4, \ldots)$ really matters here. This diagonal misses half its entries and thus intuitively does not allow for efficient retrieval of odd past steps. This remark generally prefigures a weaker performance of normal matrices with symmetric spectrum than their non-normal counterparts.

The final discussion in Remark 21 motivates a deeper comparative study of the performances of non-normal versus normal connectivity matrices. From (9), we may in particular evaluate the memory curve (as defined here in Remark 7) for W a Wigner random matrix. The performance figures are displayed in Table 1, which show a dramatic decay of the memory curve for the Wigner connectivity matrix as compared to an i.i.d. Gaussian non-normal

matrix. In Figure 6, a practical scenario of a τ-delay task is depicted comparatively for Haar versus Wigner matrices (the input data being extracted from Mackey-Glass processes but the general results hold for any non-trivial input dataset); there we confirm that, for increasing values of the delay τ, the ESN performance strongly decays for Wigner matrices as compared to Haar matrices, as predicted by the theoretical results of Table 1.

τ	i.i.d.	Wigner
0	$5.2 \cdot 10^{-1}$	$4.8 \cdot 10^{-1}$
1	$2.0 \cdot 10^{-1}$	$1.6 \cdot 10^{-2}$
2	$1.0 \cdot 10^{-1}$	$1.3 \cdot 10^{-3}$
3	$6.0 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$
4	$3.9 \cdot 10^{-2}$	$5.7 \cdot 10^{-5}$

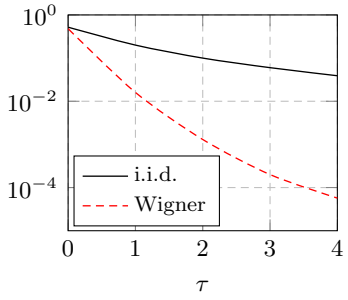

Table 1: Memory curve MC(τ) for i.i.d. versus Wigner matrices, $c = .5$.

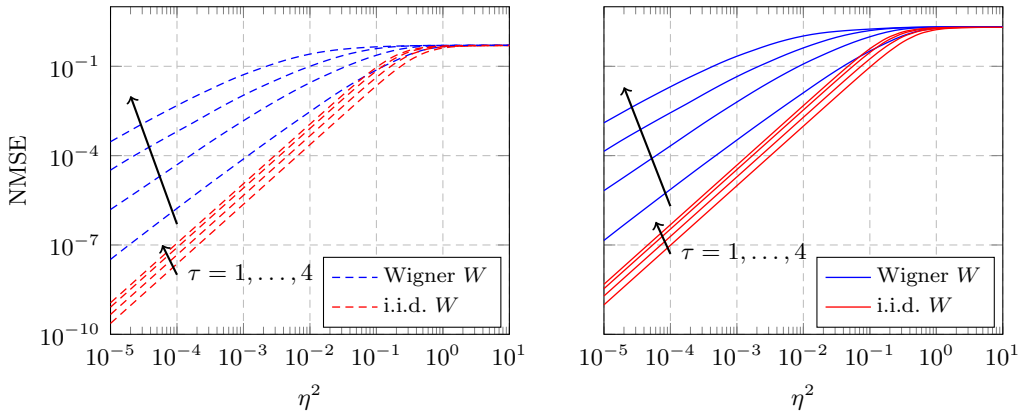

Figure 6: Training (left) and testing (right) performance of a τ-delay task for $\tau \in \{1, \ldots, 4\}$ compared for i.i.d. W versus Wigner W, $\sigma = .9$ and $n = 200$, $T = \hat{T} = 400$ in both cases (here on the Mackey-Glass dataset).

An application example in a less artificial context is devised in Figure 7, where, on a real dataset of daily pollution (PM10) records, we provide the one-day ahead interpolation performance of neural networks assuming m random i.i.d. and either (i) W with i.i.d. Gaussian entries or (ii) W Gaussian Wigner. We observe again a better performance achieved by the ESN with non-normal matrix W which, accordingly with the fact that ESN's rely heavily on past input retrieval, is coherent with the previous remark.

3.3 Further Experiments

In this section, we provide further noticeable results of interest to neural network optimization.

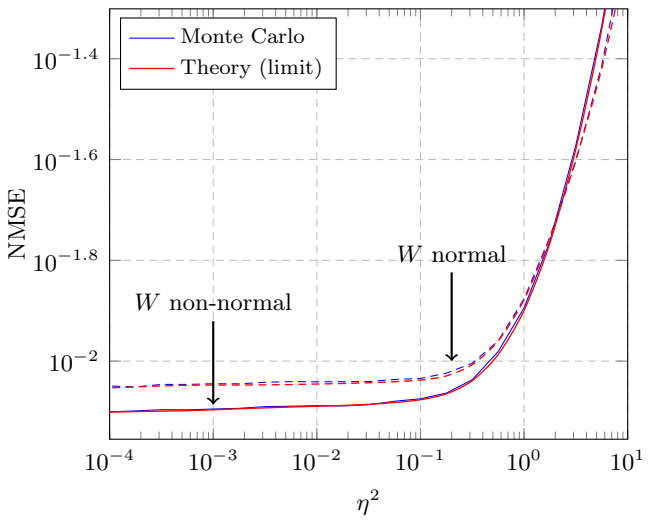

Figure 7: Testing (normalized) MSE for the PM10 one-step ahead task, W i.i.d. Gaussian or Gaussian Wigner ($\sigma = .9$), $n = 200$, $T = \hat{T} = 400$.

To start with, we consider a scenario where the testing dataset is polluted by an additional impulsive white Gaussian noise arising independently with probability p. This is depicted in Figure 8 for the Mackey–Glass one-step ahead task. It is observed here that the in-network noise is valuable in bringing the normalized MSE down to acceptable values. It is in particular seen that the more the noise impulsion probability the larger the variance η^2 should be chosen. A particular realization of the noisy Mackey-Glass output is provided in Figure 9, where it is observed that a visually small noise impulsion in the input vector drives a large fluctuation of the output for a too small-η^2 ESN.

This phenomenon can be theoretically anticipated in simple settings. Let us consider the scenario of Section 3.1 with W orthogonally invariant, where $r = U^\mathsf{T} b$ for a vector $b \in \mathbb{R}^T$ having only its last $T - k$ entries identically zero for some fixed k; let us now assume that $\hat{u} = \hat{u}_0 + \hat{e}$ for some noise vector $\hat{e}$ made of i.i.d. zero mean and variance s^2 entries, and suppose that $\hat{r} = \hat{U}_0$ for $\{\hat{U}_0\}_{ij} = [\hat{u}_0]_{i-j}$. Then, an application of Corollary 13 leads to $\hat{E}_\eta(u, r, \hat{u}, \hat{r})$ asymptotically equal to (7) plus an additional term given by (after calculus)

$$s^2 \left\| \left(\eta^2 I_T + D^{\frac{1}{2}} U U^\mathsf{T} D^{\frac{1}{2}} \right)^{-1} D^{\frac{1}{2}} U U^\mathsf{T} D^{\frac{1}{2}} (D^{-\frac{1}{2}} b) \right\|^2 . \tag{10}$$

From the inequality $\|(\eta^2 I_T + D^{\frac{1}{2}} U U^\mathsf{T} D^{\frac{1}{2}})^{-1}\| \leq \eta^{-2}$ and the fact that $\|D^{-\frac{1}{2}} b\|$ remains bounded, we get that the term (10) can be made arbitrarily small by letting $\eta^2 \to \infty$. Therefore, η^2 induces robustness in this scenario. Since $\eta^2 \to 0$ was shown to be optimal in the scenario where $s^2 = 0$, there must exist an MSE minimizing choice of $\eta^2 \in (0, \infty)$.

In a second experiment, we shall illustrate the "noise resurgence" effect discussed earlier in Remark 6. In Figure 10, we specifically draw the curves of the testing MSE variances for various experiments conducted earlier in the article. It is observed, as discussed in Remark 6 that, somewhat counter-intuitively, smaller η^2 values may lead to increased variances solely due to the in-network noise realization itself (recall that in all our experiments,

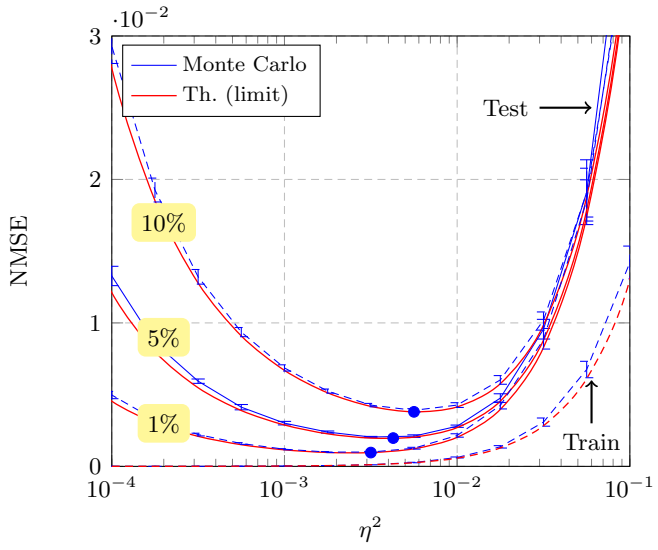

Figure 8: Testing (normalized) MSE for the Mackey-Glass one-step ahead task with 1% or 10% impulsive $\mathcal{N}(0, .01)$ noise pollution in test data inputs, W Haar with $\sigma = .9$, $n = 400$, $T = \hat{T} = 1000$. Circles indicate the NMSE theoretical minima. Error bars indicate one standard deviation of the Monte Carlo simulations.

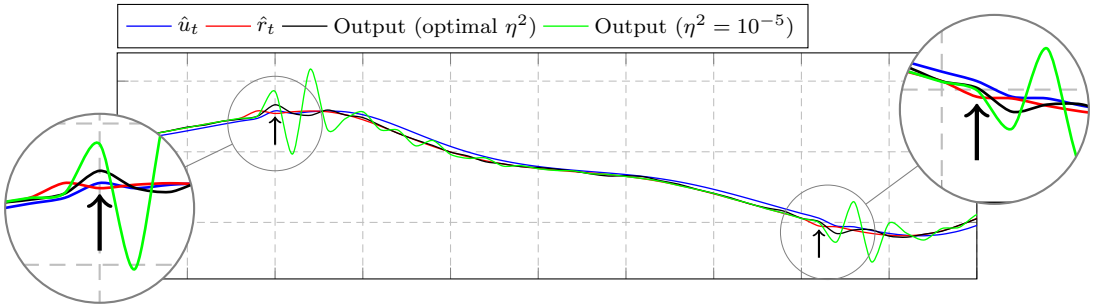

Figure 9: Realization of a 1% $\mathcal{N}(0, .01)$-noisy Mackey-Glass sequence versus network output, W Haar with $\sigma = .9$, $n = 400$, $T = \hat{T} = 1000$. In magnifying lenses, points of added impulsive noise.

the connectivity matrix W and the input-output pairs (u, r) and $(\hat{u}, \hat{r})$ are fixed across all Monte Carlo realizations). It is even more interesting to observe here each of the three possible behaviors: a "natural" MSE variance decay as $\eta^2 \to 0$, a surprising MSE increase, and even an MSE stabilization. Further theoretical analysis to understand those strikingly different behaviors would be appreciable, which would demand more advanced technical considerations.

We complete this section by a last comparative experiment of the performance of the multi-memory matrix W defined in Remark 15 specialized to the setting of Figure 1 (that is, with three rates $\sigma_1 = .99$, $\sigma_2 = .9$, and $\sigma_3 = .5$) versus Haar matrices for the different σ_i

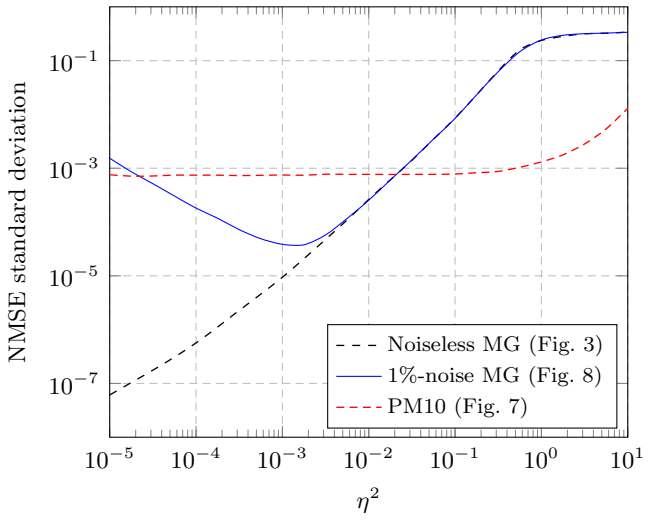

Figure 10: Standard deviation of testing NMSE for different testbeds (exemplifying the resurgence of noise effect). MG in legend stands for Mackey–Glass. In all scenarios, $n = 200$, $T = \hat{T} = 400$.

values, for the Mackey-Glass model. This is depicted in Figure 11, which shows a valuable performance gain versus ill-chosen individual hypotheses of σ and a rather fair match to the best individual σ value.

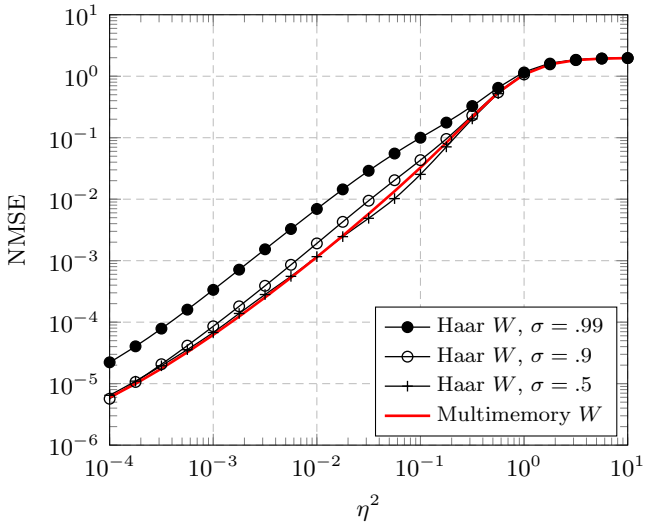

Figure 11: Testing (normalized) MSE for the Mackey Glass one-step ahead task, W (multimemory) versus $W_1^+ = .99Z_1^+$, $W_2^+ = .9Z_2^+$, $W_3^+ = .5Z_3^+$ (with Z_i^+ Haar distributed) all defined as in Figure 1, $n = 400$, $T = \hat{T} = 800$.

4. Concluding Remarks

One of the main outcomes of the present study is a better understanding of the ESN instability to low internal noise variance described by Jaeger in (Jaeger, 2001). We made it clear here that, when the noise variance is sufficiently large compared to the inverse square root of the network size, the ESN tends to have a deterministic behavior (that is, independent of the noise realization) as both time and network size grow large. This deterministic behavior was characterized here through new results from random matrix theory, with the main consequences to ESN's being encapsulated in Corollary 5 and Corollary 11. When the noise variance is however too small, random matrix theory cannot guarantee in general the aforementioned deterministic network behavior in the large system asymptotic. Although difficult to read, the asymptotic performances revolve around a critical matrix that contains the exponential memory decay information and may be use to generalize Ganguli's notion of memory curve (see Remark 7). This generalized memory curve draws improved conclusions on the ESN performance (with sometimes opposite outcomes as compared to the conclusions drawn upon the former memory curve notion).

In the particular case of some standard random matrix models for the neural connectivity matrix, we further simplified the rather involved generic expressions from Corollary 5 and Corollary 11. Of particular interest is the case of bi-orthogonally invariant random connectivity matrices for which the mean square error performances of learning and testing take on the explicit expressions of Corollary 13 or Corollary 18 from which much can be inferred. Among other results, we understood the importance of random input weights for the network performance as compared to input weights that match the leading eigenvectors of the connectivity matrix and we made it clear that the ESN testing performance is asymptotically optimal for arbitrary low noise variances *when* the task to fulfill is a mere linear combination of the last few past inputs. In additional experiments, we also understood the role of a non-trivial noise level as a robustness-to-outliers enhancer.

Beyond their theoretical value, note also that the results of this article may be used in practice to anticipate the behavior of ESN's on real-life datasets, thereby saving one from the painstaking task of running long Monte Carlo simulations. For instance, one may consider retrieving the theoretical MSE outputs corresponding to successive sequences of training and testing inputs so to better tune the ESN parameters. This is all the more precious that the network size and time windows are large since then the formulas of, say Corollary 13, can be retrieved extremely fast. In practice, for random networks, results such as Corollary 13 can be evaluated at a computational cost of $O(T^2)$ operations with minimal optimization (one may exploit the Toeplitz structure in U to further improve computations), when each run of a Monte Carlo simulation requires a prior evaluation of the successive products Wx_t, $t = 1, \ldots, T-1$, to evaluate X, prior to evaluating $(XX^\mathsf{T})^{-1}Xr$; the latter amounts to a total minimum cost of $O(RTn^2)$ with R the number of Monte Carlo iterations.

One frustrating aspect of the work nonetheless remains that, for low noise variances (typically of practical interest), our analysis leads to large mismatches when the network size is kept moderate. This is observed in Figure 3 in particular. There is as such no theoretical control of this regime. This being said, in some scenarios where the limiting singularity at zero noise can be avoided, we showed an accurate fit of our theoretical findings at all noise levels. But the main limitation of the analysis so far lies in its dealing with linear activation functions only. In a currently on-going study, using more advanced notions of random matrix theory, the authors have managed to overcome the non-linearity limitation in retrieving deterministic approximations for the mean-square error performance of a single-layer feedforward neural network with random input layer (sometimes referred

to as extreme learning machines (Huang et al., 2006)). Since the key to this result lies in the independence of the random connectivity matrix entries when seen from each neuron, the extension to multi-layer networks and eventually to recurrent network is naturally envisioned in a future investigation.

Acknowledgments

The work of Couillet and Tiomoko Ali is supported by the ANR RMT4GRAPH Project (ANR-14-CE28-0006).

Appendix A. Proof of Theorem 2

The present and next sections are dedicated to the proofs of the main results Theorems 2–9 of the article. The proofs rely on now well-established tools from random matrix theory, with an additional specificity due to the "infinitely long" time dependence between the columns of the random matrices involved; however, as the time dependence is *effectively* short (of order $o(T^\alpha)$ for any $\alpha > 0$), these matrices can be handled as if dependence was among only a few next and previous columns. We shall not deeply elaborate on all technical arguments for the sake of readability and concision. The reader more interested in the proof techniques and in more advanced time dependence considerations may refer to (Pastur and Ŝerbina, 2011; Hachem et al., 2008) on the Gaussian methods and (Banna and Merlevede, 2013) for stationary processes in random matrix theory.

Before delving into the proof of Theorem 2, let us first prove Lemma 1 which provides the expression of interest for $E_\eta(u,r)$ exploited in Theorem 2. The result is clearly valid when $n/T > 1$ as $X^\mathsf{T}X$ is almost surely non singular. Thus, only the scenario where $n/T < 1$ is of interest. Expanding the expression of ω, first observe that $E_\eta(u,r) = \frac{1}{T}\|r - X^\mathsf{T}\omega\|^2 = \frac{1}{T}r^\mathsf{T}(I_T - X^\mathsf{T}(XX^\mathsf{T})^{-1}X)^2 r$ and that $(I_T - X^\mathsf{T}(XX^\mathsf{T})^{-1}X)^2 = I_T - X^\mathsf{T}(XX^\mathsf{T})^{-1}X$. Introducing $\gamma T > 0$, $E_\eta(u,r) = \lim_{\gamma\downarrow 0}\frac{1}{T}r^\mathsf{T}(I_T - X^\mathsf{T}(XX^\mathsf{T}+\gamma T I_n)^{-1}X)r$. Now, using the identities $(AB+I)^{-1}A = A(BA+I)^{-1}$ and $A(A+bI)^{-1} = I - b(A+bI)^{-1}$ for matrices A, B and scalar b, this is $E_\eta(u,r) = \lim_{\gamma\downarrow 0}\frac{1}{T}r^\mathsf{T}(I_T - X^\mathsf{T}X(X^\mathsf{T}X+\gamma T I_T)^{-1})r = \lim_{\gamma\downarrow 0}\frac{1}{T}\gamma T r^\mathsf{T}(X^\mathsf{T}X+\gamma T I_T)^{-1}r$, from which Lemma 1 follows.

With the result at hand, we are ready to tackle the proof of Theorem 2. In the present section, W is considered a deterministic matrix with operator norm less than unity. We recall that $X = \{x_j\}_{j=0}^{T-1} \in \mathbb{R}^{n\times T}$, for the infinite time series $\ldots, x_{-1}, x_0, x_1, \ldots \in \mathbb{R}^n$, defined recursively through

$$x_{t+1} = Wx_t + mu_{t+1} + \eta\varepsilon_{t+1}$$

with m of bounded norm and $\ldots, u_{-1}, u_0, u_1, \ldots \in \mathbb{R}$ some time series. We additionally denote $A = MU$ where $M = \{W^j m\}_{j=0}^{T-1}$ and $U = T^{-\frac{1}{2}}\{u_{j-i}\}_{i,j=0}^{T-1}$. Also, let $Z = \eta T^{-\frac{1}{2}}\{\sum_{k\geq 0} W^k\varepsilon_{j-k}\}_{j=0}^{T-1}$ the concatenated noise vectors, with $\varepsilon_i \sim \mathcal{N}(0, I_n)$. With these notations and normalization, we have $X = \sqrt{T}(A+Z)$, where A and Z are expected to have operator norm of order $O(1)$ with respect to $n, T \to \infty$ as per Assumption 3 and thus so should $\frac{1}{T}XX^\mathsf{T}$.

For $\gamma > 0$, denoting $Q_\gamma = (\frac{1}{T}XX^\mathsf{T}+\gamma I_n)^{-1}$, our objective is to obtain an approximation of Q_γ in the sense of the equivalence $\leftrightarrow$ using the so-called *Gaussian method* introduced by Pastur in (Pastur and Ŝerbina, 2011). This method consists in two ingredients: (i) an integration by parts formula for Gaussian random variables (also called Stein's lemma) that stipulates that, for $x \sim \mathcal{N}(0,1)$ and a polynomially bounded differentiable f, $\mathrm{E}[xf(x)] =$

$\mathrm{E}[f'(x)]$, and (ii) concentration inequalities or moment based bounds (such as the Nash–Poincaré inequality) to control small terms. The idea here is to expand terms of the type $\mathrm{E}[[\varepsilon_i]_j[Q_\gamma]_{kl}]$ using the Gaussian integration by parts formula in order to retrieve an implicit *but deterministic* expression for Q_γ, up to small random terms. Then, thanks to concentration or moment bounds, the aforementioned small terms are shown to vanish at a sufficient speed to ensure almost sure convergence of Q_γ to the deterministic solution of the implicit equation in the sense of the equivalence $\leftrightarrow$.

We start by noticing that $Q_\gamma = \frac{1}{\gamma} I_n - \frac{1}{\gamma}\frac{1}{T} X X^\mathsf{T} Q_\gamma$, a relation often referred to as the *resolvent identity*. This allows one to write $\mathrm{E}[Q_\gamma]$ as a function of $\mathrm{E}[XX^\mathsf{T}Q_\gamma]$ which lends itself to the integration by parts approach since X is a linear function of the Gaussian variables $[\varepsilon_i]_j$.

In what follows, for readability, we shall denote $Q = Q_\gamma$ (and thus $Q_{ij} = [Q_\gamma]_{ij}$) and $\varepsilon_{ij} = [\varepsilon_i]_j$. Then we have

$$\mathrm{E}[Q_{ij}] = \frac{1}{\gamma}\boldsymbol{\delta}_{ij} - \frac{1}{\gamma}\Big(\underbrace{\mathrm{E}[[ZZ^\mathsf{T}Q]_{ij}]}_{(I)} + \underbrace{\mathrm{E}[[ZA^\mathsf{T}Q]_{ij}]}_{(II)} + \underbrace{\mathrm{E}[[AZ^\mathsf{T}Q]_{ij}]}_{(III)} + \underbrace{\mathrm{E}[[AA^\mathsf{T}Q]_{ij}]}_{(IV)} \Big). \tag{11}$$

Each of the four braced terms needs be treated independently. Note first that term (IV) is simply $\sum_k [AA^\mathsf{T}]_{ik}\mathrm{E}[Q_{kj}]$ and is thus treated similar to $\mathrm{E}[Q_{ij}]$ itself. It then remains to handle terms (I)–(III). Before handling each term, let us first introduce a few elementary results of constant use in what follows. First, by a mere development, we have

$$Z_{ab} = \frac{\eta}{\sqrt{T}} \sum_{k\geq 0}\sum_{p=1}^{n} [W^k]_{ap}\varepsilon_{p,b-k}$$

from which

$$\frac{\partial Z_{ab}}{\partial \varepsilon_{il}} = \frac{\eta}{\sqrt{T}} \sum_{k\geq 0}\sum_{p=1}^{n} \boldsymbol{\delta}_{pi}\boldsymbol{\delta}_{l,b-k}[W^k]_{ap}. \tag{12}$$

Expanding X in the expression of Q and using $\partial Q = -Q(\partial Q^{-1})Q$, we then find

$$\frac{\partial Q_{mj}}{\partial \varepsilon_{il}} = -\frac{\eta}{\sqrt{T}} \sum_{p=1}^{n} \boldsymbol{\delta}_{l\leq p} \left([Q(Z+A)]_{mp} \left[(W^{p-l})^\mathsf{T} Q\right]_{ij} + \left[(Z+A)^\mathsf{T} Q\right]_{pj} \left[QW^{p-l}\right]_{mi} \right). \tag{13}$$

It is important at this point to bring some insight from random matrix theory. If ε_{il} were a *complex* rather than real standard Gaussian random variable, the second term in the right-hand side parenthesis would not have appeared. Since first order deterministic equivalents (which is what we are proceeding to here) are usually valid irrespective of the i.i.d. distribution (real or complex) of the ε_{il}'s, it is expected that this second term will lead to vanishing terms in what follows.

With these preliminary results and this remark in mind, we can tackle the calculus of terms (I)–(III) from (11). Let us first focus on term (I). Developing $\mathrm{E}[[ZZ^\mathsf{T}Q]_{ij}]$ as a function of the ε_{kl}'s and applying the Gaussian integration-by-parts formula, we find

$$\begin{aligned}
\mathrm{E}[[ZZ^\mathsf{T}Q]_{ij}] &= \eta \sum_{l=1}^{T}\sum_{m=1}^{n}\sum_{o=1}^{n}\sum_{k\geq 0} \mathrm{E}\left[\varepsilon_{o,l-k} Z_{m,l} Q_{mj}\right] \left[W^k\right]_{io} \\
&= \eta \sum_{l=1}^{T}\sum_{m=1}^{n}\sum_{o=1}^{n}\sum_{k\geq 0} \left(\mathrm{E}\left[\frac{\partial Z_{m,k}}{\partial \varepsilon_{o,l-k}} Q_{mj}\right] + \mathrm{E}\left[\frac{\partial Q_{mj}}{\partial \varepsilon_{o,l-k}} Z_{ml}\right] \right) \left[W^k\right]_{io}.
\end{aligned}$$

Substituting the derivatives by the forms (12) and (13), we obtain after full development and simplifications

$$\mathrm{E}[[ZZ^\mathsf{T}Q]_{ij}] = \eta^2 \sum_{k\geq 0} \mathrm{E}\left[\left[W^k(W^k)^\mathsf{T}Q\right]_{ij}\right] - \eta^2 \sum_{k\geq 0}\sum_{q=-k}^{T-1}\sum_{l=1}^{T-q^+} \mathrm{E}\left[\left[W^k(W^{k+q})^\mathsf{T}Q\right]_{ij} \frac{1}{T}[Z^\mathsf{T}(Z+A)\tilde{Q}]_{l,q+l}\right] + \mathrm{E}[\zeta_{ij}^{[1]}]$$

where we defined $\tilde{Q} = \tilde{Q}_\gamma = (\frac{1}{T}X^\mathsf{T}X + \gamma I_n)^{-1}$ and where the term $\zeta_{ij}^{[1]}$ arises from the development of the aforementioned second term in the parentheses of (13) and can be shown to satisfy $\zeta^{[1]} \leftrightarrow 0$. Similarly, addressing the term (II) in (11), we find

$$\mathrm{E}[[ZA^\mathsf{T}Q]_{ij}] = -\eta^2 \sum_{k\geq 0}\sum_{q=-k}^{T-1}\sum_{l=1}^{T-q^+} \mathrm{E}\left[\frac{1}{T}\left[A^\mathsf{T}(Z+A)\tilde{Q}\right]_{l,q+l}\left[W^k(W^{k+q})^\mathsf{T}Q\right]_{ij}\right] + \mathrm{E}[\zeta_{ij}^{[2]}]$$

where again we can show that $\zeta^{[2]} \leftrightarrow 0$. Summing the approximations for (I) and (II), from the resolvent identity $(Z+A)^\mathsf{T}(Z+A)\tilde{Q} = I_n - \gamma\tilde{Q}$, we find

$$\mathrm{E}[[Z(Z+A)^\mathsf{T}Q]_{ij}] = \eta^2 \sum_{k\geq 0} \mathrm{E}\left[\left[W^k(W^k)^\mathsf{T}Q\right]_{ij}\right] - \eta^2 \sum_{k\geq 0}\sum_{q=-k}^{T-1}\sum_{l=1}^{T-q^+} \mathrm{E}\left[\left[W^k(W^{k+q})^\mathsf{T}Q\right]_{ij} \frac{1}{T}[I_n - \gamma\tilde{Q}]_{l,q+l}\right] + \mathrm{E}[\zeta_{ij}^{[1]} + \zeta_{ij}^{[2]}].$$

Since $[I_n]_{l,q+l} = \boldsymbol{\delta}_{q=0}$, the first right-hand side term cancels with the part of the second term involved with matrix $\frac{1}{T}I_n$, and we find

$$\mathrm{E}[[Z(Z+A)^\mathsf{T}Q]_{ij}] = \eta^2\gamma \sum_{k\geq 0}\sum_{q=-k}^{T-1}\sum_{l=1}^{T-q^+} \mathrm{E}\left[\left[W^k(W^{k+q})^\mathsf{T}Q\right]_{ij} \frac{1}{T}\tilde{Q}_{l,q+l}\right] + \mathrm{E}[\zeta_{ij}^{[1]} + \zeta_{ij}^{[2]}]. \tag{14}$$

Moving to term (III) in (11), since A is deterministic, we first find the interesting expression

$$\mathrm{E}[[Z^\mathsf{T}Q]_{ij}] = -\eta^2 \sum_{k\geq 0}\sum_{q=-k}^{T-1}\sum_{l=1}^{T-q^+} \mathrm{E}\left[\frac{1}{T}\operatorname{tr}(W^k(W^{k+q})^\mathsf{T}Q)[(Z+A)^\mathsf{T}Q]_{q+i,j}\right] + \mathrm{E}[\zeta_{ij}^{[3]}] \tag{15}$$

with $\zeta^{[3]} \leftrightarrow 0$ from which immediately we get

$$\mathrm{E}[[AZ^\mathsf{T}Q]_{ij}] = -\eta^2 \sum_{k\geq 0}\sum_{q=-k}^{T-1}\sum_{l=1}^{T-q^+} \mathrm{E}\left[\frac{1}{T}\operatorname{tr}(W^k(W^{k+q})^\mathsf{T}Q)A_{il}[(Z+A)^\mathsf{T}Q]_{q+l,j}\right] + \mathrm{E}[[A\zeta^{[3]}]_{ij}]$$

and we of course still have $A\zeta^{[3]} \leftrightarrow 0$.

We must discuss at this point the next key idea of the Gaussian method. In term (III), the right-hand side expectation is taken over the product of the trace $\frac{1}{T}\operatorname{tr}(W^k(W^{k+q})^\mathsf{T}Q)$ and of the quantity $A_{il}[(Z+A)^\mathsf{T}Q]_{q+l,j}$. Writing $\frac{1}{T}\operatorname{tr}(W^k(W^{k+q})^\mathsf{T}Q) = \mathrm{E}[\frac{1}{T}\operatorname{tr}(W^k(W^{k+q})^\mathsf{T}Q)] + (\frac{1}{T}\operatorname{tr}(W^k(W^{k+q})^\mathsf{T}Q) - \mathrm{E}[\frac{1}{T}\operatorname{tr}(W^k(W^{k+q})^\mathsf{T}Q)])$, it can be shown, using Cauchy–Schwarz and the Nash–Poicaré inequalities (Pastur and Ŝerbina, 2011), along with the Borel–Cantelli lemma (Billingsley, 1995), that

$$\sum_{k\geq 0}\sum_{q=-k}^{T-1}\sum_{l=1}^{T-q^+} \left(\frac{1}{T}\operatorname{tr}(W^k(W^{k+q})^\mathsf{T}Q) - \mathrm{E}\left[\frac{1}{T}\operatorname{tr}(W^k(W^{k+q})^\mathsf{T}Q)\right]\right) A_{il}[(Z+A)^\mathsf{T}Q]_{q+l,j} \leftrightarrow 0$$

which unfolds from $\frac{1}{T}\operatorname{tr}(W^k(W^{k+q})^\mathsf{T}Q)$ concentrating around its mean in the large n, T regime, a standard result of random matrix theory. The main non-classical difficulty in showing this result lies here in the fact that the summation over up to T values of the dummy variable q involves both terms in and outside the bracket. Nonetheless, since $\rho(W) < 1$, $\|W^q\|$ vanishes at exponential speed and thus only $O(\log(T))$ values of q are effectively playing a role. The aforementioned Nash–Poicaré inequality argument ensures a control of the residual terms with a $O(1/T^2)$ variance for each q-summand which can then be summed over the non-trivial values of q to bring a total variance bounded by $O(\log(T)/T^2)$, which is summable, and then allows for Borel–Cantelli to be applied.

The same reasoning applies to the main expectation in the expression of $(I) + (II)$, where here the term that concentrates around its mean is $\frac{1}{T}\sum_{l=1}^{T-q^+}\tilde{Q}_{l,q+l}$, which is more easily seen as $\frac{1}{T}\operatorname{tr}(J^q\tilde{Q})$.

The relation (15) in itself is quite instructive. Indeed, with the previous remark on the concentration of $\frac{1}{T}\operatorname{tr}(W^k(W^{k+q})^\mathsf{T}Q)$, we may break the right-hand expectation as well as the term $(Z+A)^\mathsf{T}Q$ into $Z^\mathsf{T}Q + A^\mathsf{T}Q$ to retrieve a connection between left- and right-hand sides. Precisely, we find that

$$\begin{aligned}
&\left[\left(I_T + \eta^2 \sum_{k\geq 0}\sum_{q=-k}^{T-1} \mathrm{E}\left[\frac{1}{T}\operatorname{tr}\left(W^kW^{k+q})^\mathsf{T}Q\right)\right] J^q\right)\mathrm{E}\left[X^\mathsf{T}Q\right]\right]_{ij} \\
&= -\eta^2 \sum_{k\geq 0}\sum_{q=-k}^{T-1} \mathrm{E}\left[\frac{1}{T}\operatorname{tr}(W^k(W^{k+q})^\mathsf{T}Q)\right]\mathrm{E}\left[[J^qA^\mathsf{T}Q]_{i,j}\right] + o(1)
\end{aligned}$$

where we used $[B]_{q+i,j} = [J^qB]_{i,j}$. Remark now that

$$\begin{aligned}
\sum_{k\geq 0}\sum_{q=-k}^{T-1}\frac{1}{T}\operatorname{tr}\left(W^k(W^{k+q})^\mathsf{T}Q\right)J^q &= \sum_{k\geq 0}\left\{\frac{1}{T}\operatorname{tr}\left(W^{k+(b-a)^+}(W^{k+(a-b)^+})^\mathsf{T}Q\right)\right\}_{a,b=1}^T \\
&= \left\{\frac{1}{T}\operatorname{tr}\left(S_{a-b}Q\right)\right\}_{a,b=1}^T.
\end{aligned}$$

Denoting $\bar{R} = \mathrm{E}[\{\frac{1}{T}\operatorname{tr}(S_{a-b}Q)\}_{a,b=1}^T]$ and using concentration arguments (Nash–Poincaré inequality in particular) entails

$$Z^\mathsf{T}Q \leftrightarrow -\eta^2\left(I_T + \eta^2\bar{R}\right)^{-1}\bar{R}A^\mathsf{T}Q. \tag{16}$$

From the definition of the equivalence relation $\leftrightarrow$, this entails

$$AZ^\mathsf{T}Q \leftrightarrow -\eta^2 A\left(I_T + \eta^2\bar{R}\right)^{-1}\bar{R}A^\mathsf{T}Q. \tag{17}$$

Similarly, recalling (14), we have

$$\begin{aligned}
Z(Z+A)^\mathsf{T}Q &\leftrightarrow \eta^2\gamma\sum_{k\geq 0}\sum_{q=-k}^{T-1}\frac{1}{T}\operatorname{tr}(J^q\tilde{Q})W^k(W^{k+q})^\mathsf{T}Q \\
&= \eta^2\gamma\sum_{q=-\infty}^{\infty}\frac{1}{T}\operatorname{tr}(J^q\tilde{Q})S_qQ.
\end{aligned}$$

We may then define $\bar{\bar{R}} = \sum_{q=-\infty}^{\infty}\mathrm{E}[\frac{1}{T}\operatorname{tr}(J^q\tilde{Q})S_q]$. Added to (17) and $AA^\mathsf{T}Q$, this is

$$(Z+A)(Z+A)^\mathsf{T}Q \leftrightarrow -\eta^2\gamma\bar{\bar{R}} - \eta^2A(I_T+\eta^2\bar{R})^{-1}\bar{R}A^\mathsf{T}Q + AA^\mathsf{T}Q.$$

With $AA^\mathsf{T} = A(I_T + \eta^2 \bar{R})^{-1}(I_T + \eta \bar{R})A^\mathsf{T}$ and $(Z+A)(Z+A)^\mathsf{T} Q = I_n - \gamma Q$, this further reads

$$Q \leftrightarrow \frac{1}{\gamma} I_n - \eta^2 \bar{\tilde{R}} Q - \frac{1}{\gamma} A (I_T + \eta^2 \bar{R})^{-1} A^\mathsf{T} Q.$$

which, after gathering the factors of Q together, finally gives the first identity

$$Q \leftrightarrow \frac{1}{\gamma} \left(I_n + \eta^2 \bar{\tilde{R}} + \frac{1}{\gamma} A (I_T + \eta^2 \bar{R})^{-1} A^\mathsf{T} \right)^{-1}. \tag{18}$$

To pursue our investigation, we need to proceed to the same development for the matrix $\tilde{Q}$ which appears in the definition of $\bar{\tilde{R}}$. The idea is to express $\tilde{Q}$ under a form involving Q itself, then closing the loop. The analysis is extremely similar to that of Q and it is not surprising (from the symmetry between Q and $\tilde{Q}$) to finally obtain

$$\tilde{Q} \leftrightarrow \frac{1}{\gamma} \left(I_T + \eta^2 \bar{R} + \frac{1}{\gamma} A^\mathsf{T} (I_n + \eta^2 \bar{\tilde{R}})^{-1} A^\mathsf{T} \right)^{-1}. \tag{19}$$

At this point, however, both $\bar{R}$ and $\bar{\tilde{R}}$ are non explicit quantities that depend on the statistics of Q and $\tilde{Q}$. From (18), we get that, for each a, b,

$$\frac{1}{T} \operatorname{tr}(S_{a-b} Q) \leftrightarrow \frac{1}{\gamma} \frac{1}{T} \operatorname{tr} S_{a-b} \left(I_n + \eta^2 \bar{\tilde{R}} + \frac{1}{\gamma} A (I_T + \eta^2 \bar{R})^{-1} A^\mathsf{T} \right)^{-1}$$

and this relation is shown to be uniform across a, b, as it involves only $O(\log(T))$ non-trivial coefficients. To freely identify $\bar{R}$ with $\{\frac{1}{\gamma T} \operatorname{tr} S_{a-b}(I_n + \eta^2 \bar{\tilde{R}} + \frac{1}{\gamma} A (I_T + \eta^2 \bar{R})^{-1} A^\mathsf{T})^{-1}\}_{a,b=1}^T$, one may ensure that the difference between both matrices vanishes in spectral norm almost surely (here the relation $\leftrightarrow$ may not be enough).[4] Here the result holds true because both $\{\frac{1}{T} \operatorname{tr}(S_{a-b} Q)\}_{a,b=1}^T$ and $\bar{R}$ are Toeplitz matrices with exponentially decaying coefficients away from the main diagonal. Hence, we may essentially see each matrix as the sum of a circulant matrix and of a matrix with $O(\log(T))$ non-vanishing upper-right and lower-left entries (see (Gray, 2006) for such a construction). Circulant matrices being diagonalizable in the Fourier basis with eigenvalues equal to the Fourier transform of the concatenated first column and row, that the difference in spectral norm vanishes boils down to the convergence of the difference between these Fourier transforms, which is easily obtained through the joint entry convergence and exponential decrease. As for the remaining corner entries, being of $\log(T)$ number, we deal here with the difference in spectral norm of small rank matrices, which is obtained by direct uniform convergence. As such, generally speaking, if the entries of a Toeplitz matrix with exponentially vanishing profile converge jointly to given limits, then the limiting Toeplitz matrix is equivalent in the spectral norm sense.

Similarly, to identify $\bar{\tilde{R}}$ with $\sum_q \frac{1}{\gamma T} \operatorname{tr}(J^q (I_T + \eta^2 \bar{R} + \frac{1}{\gamma} A^\mathsf{T} (I_n + \eta^2 \bar{\tilde{R}})^{-1} A^\mathsf{T})^{-1}) S_q$, we need to show the spectral norm difference of these matrices vanishes almost surely. This is here obtained from the uniform convergence across the $O(\log(T))$ first trace coefficients (say for all $|q| \leq C \log(T)$) and from the corresponding exponentially vanishing spectral norm of S_q.

All said, we may then define R_γ, $\bar{R}_\gamma$, $\bar{Q}_\gamma$, and $\bar{\tilde{Q}}_\gamma$ as in Theorem 2 and the results above ensure that $Q_\gamma \leftrightarrow \bar{Q}_\gamma$ and $\tilde{Q}_\gamma \leftrightarrow \bar{\tilde{Q}}_\gamma$.

4. A typical counter-example is the case of $Z \in \mathbb{R}^{n \times T}$ with i.i.d. zero mean and unit variance entries for which $[\frac{1}{T} Z Z^\mathsf{T}]_{ab} \to \boldsymbol{\delta}_{a-b}$ uniformly over a, b while clearly $(\frac{1}{T} Z Z^\mathsf{T} + \gamma I_n)^{-1} \not\leftrightarrow (1+\gamma)^{-1} I_T$.

Remark 22 (Result without washout period) *Theorem 2 assumes an infinite noise time series* $(\ldots, \varepsilon_{-1}, \varepsilon_0, \varepsilon_1, \ldots)$. *One might have alternatively considered a scenario without washout period, that is, with* $x_{-1} = 0$ *and first time instant being* $t = 0$. *In this case, Theorem 2 remains valid but for the following updated expressions of* R_γ *and* $\tilde{R}_\gamma$

$$R_\gamma = \left\{ \sum_{k=0}^{\max(i,j)-1} \frac{1}{T} \operatorname{tr} W^{k+(j-i)^+} (W^{k+(i-j)^+})^\mathsf{T} \bar{Q}_\gamma \right\}_{i,j=1}^{T}$$

$$\tilde{R}_\gamma = \sum_{q=-(T-1)}^{T-1} \frac{1}{T} \operatorname{tr} \left(J^q \bar{\tilde{Q}}_\gamma \right) \sum_{k=0}^{T-1-|q|} W^{k+(-q)^+} (W^{k+q^+})^\mathsf{T}.$$

In particular, R_γ *is no longer Toeplitz. Nonetheless the non-Toeplitz behavior is essentially concentrated in the top-left corner of size* $O(\log(T))$ *since the remainder of the matrix behaves essentially as Toeplitz (for* $i, j \geq C \log(T)$ *for some large enough constant* C*). This modification may alter the behavior of the associated train and test MSE, especially if* r *and* $\hat{r}$ *concentrate their energy in their first entries.*

Appendix B. Proof of Theorem 9

The first part of Theorem 9 is directly obtained from (16) along with $Q_\gamma \leftrightarrow \bar{Q}_\gamma$. Indeed, from these relations, we have

$$\begin{aligned} Q_\gamma \frac{1}{\sqrt{T}} X = Q_\gamma Z + Q_\gamma A &\leftrightarrow -\eta^2 \bar{Q}_\gamma A R_\gamma (I_T + \eta^2 R_\gamma)^{-1} + \bar{Q}_\gamma A \\ &= \bar{Q}_\gamma A (I_T + \eta^2 R_\gamma)^{-1}. \end{aligned}$$

The proof of the second part of Theorem 9 is not as straightforward as it involves twice the matrix Q_γ and thus results from Theorem 2 cannot be immediately applied. To handle this term, first write

$$\frac{1}{T} X^\mathsf{T} Q_\gamma B Q_\gamma X = \underbrace{Z^\mathsf{T} Q_\gamma B Q_\gamma Z}_{(I)} + \underbrace{Z^\mathsf{T} Q_\gamma B Q_\gamma A}_{(II)} + \underbrace{A^\mathsf{T} Q_\gamma B Q_\gamma Z}_{(III)} + \underbrace{A^\mathsf{T} Q_\gamma B Q_\gamma A}_{(IV)}. \tag{20}$$

Since B is assumed symmetric, (III) is the transposed version of (II), so that only one of the two needs be studied.

Similar to Appendix A, we shall from now on simply write Q_γ as Q, $\tilde{Q}_\gamma$ as $\tilde{Q}$, etc.

We start by addressing term (I). We use again the Gaussian tools centered around the Gaussian integration by parts formula. We shall also benefit from the results of Theorem 2. Since B is deterministic, it needs not be included early in calculations so we merely start by evaluating, for given indices i, j, k, l,

$$\begin{aligned} \mathrm{E}\left[[Z^\mathsf{T} Q]_{ij} [QZ]_{kl} \right] &= \sum_{m,m',p,p'=1}^{n} \sum_{q,q' \geq 0} \eta^2 \mathrm{E}\left[\varepsilon_{p,i-q} \varepsilon_{p',l-q'} Q_{mj} Q_{km'} \right] [W^q]_{mp} [W^{q'}]_{m'p'} \\ &= \sum_{m,m',p,p'=1}^{n} \sum_{q,q' \geq 0} \eta^2 \mathrm{E}\left[\frac{\partial (\varepsilon_{p,i-q} Q_{mj} Q_{km'})}{\partial \varepsilon_{p',l-q'}} \right] [W^q]_{mp} [W^{q'}]_{m'p'} \end{aligned}$$

where the second line follows from the Gaussian integration-by-parts formula. Developing the derivative based on (13) and on the fact that $\partial \varepsilon_{ab} / \partial \varepsilon_{cd} = \boldsymbol{\delta}_{ac} \boldsymbol{\delta}_{bd}$, we get after

simplification

$$\begin{aligned}\mathrm{E}\left[[Z^\mathsf{T}Q]_{ij}[QZ]_{kl}\right] &= \eta^2 \sum_{q,q'\geq 0} \mathrm{E}\left[\frac{1}{T}[QW^q(W^{q'})^\mathsf{T}Q]_{jk}\boldsymbol{\delta}_{i-q,l-q'}\right] \\ &\quad - \eta^3 \sum_{q,q'\geq 0}\sum_{s=l-q'}^{T} \mathrm{E}\left[[\frac{1}{\sqrt{T}}\varepsilon^\mathsf{T}(W^q)^\mathsf{T}Q(Z+A)]_{i-q,s}\frac{1}{T}[QW^{q'}(W^{q'+s-l})^\mathsf{T}Q]_{kj}\right] \\ &\quad - \eta^3 \sum_{q,q'\geq 0}\sum_{s=l-q'}^{T} \mathrm{E}\left[[\frac{1}{\sqrt{T}}\varepsilon^\mathsf{T}(W^q)^\mathsf{T}Q]_{i-q,j}[Q(Z+A)]_{k,s}\frac{1}{T}\operatorname{tr}(W^{q'}(W^{q'+s-l})^\mathsf{T}Q)\right] \\ &\quad + \mathrm{E}[\zeta^{[1]}_{ijkl}] \end{aligned} \tag{21}$$

for some $\zeta^{[1]}_{ijkl} \leftrightarrow 0$ (arising from terms consistent with the remark following (13) in Appendix A) and where $\varepsilon = \{\varepsilon_{ij}\}_{ij=1}^{n,T}$. Inserting B_{jk}, summing over j and k, we obtain after simplifications

$$\mathrm{E}\left[[Z^\mathsf{T}QBQZ]_{il}\right] = \eta^2\bar{G}_{il} - \eta^2\mathrm{E}\left[[Z^\mathsf{T}Q(Z+A)\bar{G}]_{il}\right] - \eta^2\mathrm{E}\left[[Z^\mathsf{T}QBQ(Z+A)\bar{R}]_{il}\right] + o(1)$$

where $\bar{R}$ was introduced in Appendix A and we defined $\bar{G}$ the matrix with

$$\bar{G}_{ij} = \sum_{k\geq 0}\mathrm{E}\left[\frac{1}{T}\operatorname{tr}\left(BQW^{k+(j-i)^+}(W^{k+(i-j)^+})^\mathsf{T}Q\right)\right].$$

Gathering the terms in $Z^\mathsf{T}QBQZ$ together along with concentration arguments, we finally obtain

$$Z^\mathsf{T}QBQZ \leftrightarrow \eta^2\bar{G}(I_T+\eta^2\bar{R})^{-1} - \eta^2 Z^\mathsf{T}Q(Z+A)\bar{G}(I_T+\eta^2\bar{R})^{-1} - \eta^2 Z^\mathsf{T}QBQA\bar{R}(I_T+\eta^2\bar{R})^{-1}.$$

In the right-hand side formulation, the second term can be approximated from the results of Theorem 2 as well as the first part of Theorem 9; indeed, note from $(Z+A)^\mathsf{T}Q(Z+A) = \tilde{Q}(Z+A)^\mathsf{T}(Z+A) = I_T - \gamma\tilde{Q}$ that $Z^\mathsf{T}Q(Z+A) = I_T - \gamma\tilde{Q} - A^\mathsf{T}Q(Z+A)$, so that

$$\begin{aligned} Z^\mathsf{T}QBQZ \leftrightarrow{}& \eta^2\gamma\bar{\tilde{Q}}\bar{G}(I_T+\eta^2\bar{R})^{-1} + \eta^2 A^\mathsf{T}\bar{Q}A(I_T+\eta^2\bar{R})^{-1}\bar{G}(I_T+\eta^2\bar{R})^{-1} \\ &- \eta^2 Z^\mathsf{T}QBQA\bar{R}(I_T+\eta^2\bar{R})^{-1}. \end{aligned} \tag{22}$$

In this expression, the last right-hand side term still involves $Z^\mathsf{T}QBQA$, yet to be characterized. This is the objective of the next step, which coincides with the study of the term (II) in (20).

Following the derivation of term (I), terms (II) and (III) are easily obtained (indeed, they somewhat boil down to (21) without the first right-hand side term and without the components $\varepsilon^\mathsf{T}(W^q)^\mathsf{T}$ in the subsequent terms). Precisely, all calculus made, we find that

$$QBQZ \leftrightarrow -\eta^2 Q(Z+A)\bar{G} - \eta^2 QBQ(Z+A)\bar{R}$$

from which

$$QBQZ \leftrightarrow -\eta^2 Q(Z+A)\bar{G}(I_T+\eta^2\bar{R})^{-1} - \eta^2 QBQA\bar{R}(I_T+\eta^2\bar{R})^{-1}.$$

Again, the first right-hand side term is easily expressed by Theorem 2 and the first result of Theorem 9, from which

$$QBQZ \leftrightarrow -\eta^2\bar{Q}A(I_T+\eta^2\bar{R})^{-1}\bar{G}(I_T+\eta^2\bar{R})^{-1} - \eta^2 QBQA\bar{R}(I_T+\eta^2\bar{R})^{-1}. \tag{23}$$

but the second term now involves the quantity QBQ which is our next target. Since studying QBQ entails studying $A^\mathsf{T}QBQA$, this shall provide us with the term (IV) in (20). To address QBQ, it suffices to estimate $\mathrm{E}[Q_{ij}Q_{kl}]$; from the resolvent identity $Q_{ij} = \frac{1}{\gamma}\boldsymbol{\delta}_{ij} - \frac{1}{\gamma}[\frac{1}{T}XX^\mathsf{T}Q]_{ij}$, this is developed as

$$\begin{aligned}\mathrm{E}\left[Q_{ij}Q_{kl}\right] = -\frac{1}{\gamma}\left(\mathrm{E}[[Z^\mathsf{T}ZQ]_{ij}Q_{kl}] + \mathrm{E}[[ZA^\mathsf{T}Q]_{ij}Q_{kl}] + \mathrm{E}[[AZ^\mathsf{T}Q]_{ij}Q_{kl}] + \mathrm{E}[[AA^\mathsf{T}Q]_{ij}Q_{kl}]\right)\\ + \frac{1}{\gamma}\boldsymbol{\delta}_{ij}\mathrm{E}[Q_{kl}].\end{aligned}$$

The deterministic equivalent for $\mathrm{E}[Q_{kl}]$ is already known, and we are then left to evaluate the first four terms, some of which can be retrieved from previous calculus. Developing each term, integrating the previously developed equivalents, while introducing the matrix B and summing, after some tedious calculus, we finally obtain

$$\begin{aligned}QBQ \leftrightarrow \frac{1}{\gamma}B\bar{Q} + \frac{\eta^2}{\gamma}A(I_T+\eta^2\bar{R})^{-1}\bar{G}(I_T+\eta^2\bar{R})^{-1}A^\mathsf{T}\bar{Q}\\ - \eta^2\bar{\tilde{R}}QBQ + \frac{1}{\gamma}\bar{\tilde{G}}\bar{Q} - \frac{1}{\gamma}A(I_T+\eta^2\bar{R})^{-1}A^\mathsf{T}QBQ\end{aligned}$$

where we introduced the notation

$$\bar{\tilde{G}} = \sum_{q=-\infty}^{\infty}\eta^2\mathrm{E}\left[\frac{1}{T}\operatorname{tr}\left(J^q(A+Z)^\mathsf{T}QBQ(Z+A)\right)\right]\sum_{k\geq 0}W^{k+(-q)^+}(W^{k+q^+})^\mathsf{T}.$$

Gathering all terms proportional to QBQ, we finally obtain

$$QBQ \leftrightarrow \bar{Q}(B+\bar{\tilde{G}})\bar{Q} + \eta^2\bar{Q}A(I_T+\eta^2\bar{R})^{-1}\bar{G}(I_T+\eta^2\bar{R})^{-1}A^\mathsf{T}\bar{Q}. \tag{24}$$

Substituting (24) in (23), then substituting the result in (22), we may now completely characterize $\frac{1}{T}X^\mathsf{T}QBQX$ (after simplification) as

$$\frac{1}{T}X^\mathsf{T}QBQX \leftrightarrow \eta^2\gamma^2\bar{\tilde{Q}}\bar{G}\bar{\tilde{Q}} + (I_T+\eta^2\bar{R})^{-1}A^\mathsf{T}\bar{Q}[B+\bar{\tilde{G}}]\bar{Q}A(I_T+\eta^2\bar{R})^{-1}.$$

It remains to evaluate $\mathrm{E}[\frac{1}{T}\operatorname{tr}(J^q(A+Z)^\mathsf{T}QBQ(Z+A))]$ in the expression of $\bar{\tilde{G}}$. For this, we shall exploit the fact that $A = MU$ which, since M has columns of exponentially decreasing norm, can be considered as a matrix of rank "essentially of order $O(\log(T))$"; that is, while being full rank, A can be well approximated in spectral norm by the product $\check{M}\check{U}$ of the first $O(\log(T))$ columns $\check{M}$ of M and first $O(\log(T))$ rows $\check{U}$ of U. This entails that, in the deterministic approximation for $(A+Z)^\mathsf{T}QBQ(Z+A)$, only the terms not involving a product with A or A^T will remain after taking the normalized trace. And thus we get, after development and simplification

$$\left\|\bar{\tilde{G}} - \sum_{q=-\infty}^{\infty}\gamma^2\eta^4\frac{1}{T}\operatorname{tr}\left(J^q\bar{\tilde{Q}}\bar{G}\bar{\tilde{Q}}\right)\sum_{k\geq 0}W^{k+(-q)^+}(W^{k+q^+})^\mathsf{T}\right\| \to 0.$$

It then suffices to use concentration identities and the results of Appendix A to finally substitute $\bar{R}$ with R, $\bar{\tilde{R}}$ with $\tilde{R}$, and $\bar{G}$, $\bar{\tilde{G}}$ with G and $\tilde{G}$, respectively. This concludes the proof of Theorem 9.

Remark 23 (On the speed of convergence) *To better appreciate the interplay between η^2 and n, T, note that all convergences discussed in Appendices A–B involve either quadratic forms of the type $a^\mathsf{T} Q a$ for $Q \in \mathbb{R}^{n\times n}$ a random matrix based on some $\varepsilon \in \mathbb{R}^{n\times T}$, matrix with independent entries, or normalized traces $\frac{1}{n}\operatorname{tr} Q$. It is a standard central limit result in random matrix theory that the former quadratic form $a^\mathsf{T} Q a$ fluctuates at speed $O(n^{-\frac{1}{2}})$, that is, $\operatorname{var}(a^\mathsf{T} Q a) = O(n^{-1})$, and that normalized traces fluctuate at the faster speed $O(n^{-1})$. As such, the results of Theorems 2–9 and Proposition 4–10 can be trusted with high probability within a $O(n^{-\frac{1}{2}})$ error bound.*

With respect to η^2, the bounds between random quantities and deterministic equivalents, say Q and $\bar{Q}$, are proportional to $1/\eta^2$. This is why η^2 is assumed fixed and not decaying in our results. Nonetheless, as both bounds in n and η^2 multiply, it is expected that convergence is maintained in general so long that $n^{-\frac{1}{2}}/\eta^2 \to 0$, i.e., when $\eta^2 \gg n^{-\frac{1}{2}}$.

References

Z. D. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics, New York, NY, USA, second edition, 2009.

Marwa Banna and Florence Merlevede. Limiting spectral distribution of large sample covariance matrices associated with a class of stationary processes. *Journal of Theoretical Probability*, pages 1–39, 2013.

P. Biane. Free probability for probabilists. *Quantum Probability Communications*, 11:55–71, 2003.

P. Billingsley. *Probability and Measure*. John Wiley and Sons, Inc., Hoboken, NJ, third edition, 1995.

C. M. Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7 (1):108–116, 1995.

Surya Ganguli, Dongsung Huh, and Haim Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, 105(48):18970–18975, 2008.

Leon Glass and Michael C. Mackey. A simple model for phase locking of biological oscillators. *Journal of Mathematical Biology*, 7(4):339–352, 1979.

R. M. Gray. Toeplitz and circulant matrices: a review. *Foundations and Trends in Communications and Information Theory*, 2(3), 2006.

W. Hachem, O. Khorunzhy, P. Loubaton, J. Najim, and L. A. Pastur. A new approach for capacity analysis of large dimensional multi-antenna channels. *IEEE Transactions on Information Theory*, 54(9):3987–4004, 2008.

Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1):489–501, 2006.

H. Jaeger. *Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach*. GMD-Forschungszentrum Informationstechnik, 2005.

H. Jaeger and H. Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80, 2004.

Herbert Jaeger. *Short term memory in echo state networks*. GMD-Forschungszentrum Informationstechnik, 2001.

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

M. Ozturk, C. Mustafa, D. Xu, and J. C. Príncipe. Analysis and design of echo state networks. *Neural Computation*, 19(1):111–138, 2007.

L. Pastur and M. Ŝerbina. *Eigenvalue distribution of large random matrices*. American Mathematical Society, 2011.

A. Rodan and P. Tiňo. Minimum complexity echo state network. *Neural Networks, IEEE Transactions on*, 22(1):131–144, 2011.

Tobias Strauss, Welf Wustlich, and Roger Labahn. Design strategies for weight matrices of echo state networks. *Neural computation*, 24(12):3246–3276, 2012.

Peter Tiňo and Ali Rodan. Short term memory in input-driven linear dynamical systems. *Neurocomputing*, 112:58–63, 2013.

T. Toyoizumi and L. F. Abbott. Beyond the edge of chaos: Amplification and temporal integration by recurrent networks in the chaotic regime. *Physical Review E*, 84(5):051908, 2011.

D. Verstraeten, J. Dambre, X. Dutoit, and B. Schrauwen. Memory versus non-linearity in reservoirs. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–8. IEEE, 2010.

Y. Xue, L. Yang, and S. Haykin. Decoupled echo state networks with lateral inhibition. *Neural Networks*, 20(3):365–376, 2007.