A Random Matrix Approach to Recurrent Neural Networks

Abstract

Recurrent neural networks, especially in their linear version, have provided many qualitative insights on their performance under different configurations. This article provides, through a novel random matrix framework, the quantitative counterpart of these performance results, particularly in the case of echo-state networks. Beyond mere insights, our approach conveys a deeper understanding on the core mechanism under play for both training and testing.

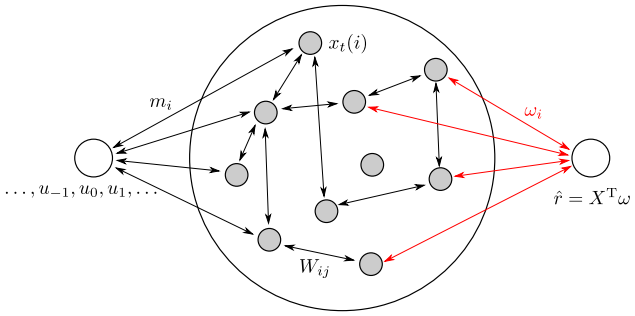

Figure 1. Echo-state neural network.

1. Introduction

Echo-state networks (ESN's) are part of the broader family of recurrent neural networks, specifically dedicated to handling time-series related tasks (such as prediction, non-linear interpolation, etc.) (Jaeger, 2001a;b). Their main feature, as opposed to more conventional neural networks, is to rely on a fixed (but generally randomly chosen) connectivity matrix, the so-called reservoir, and only to enforce network-to-sink edges during the training phase. This reduces overfitting but in turn only allows for short-term memorization capabilities, unlike backward propagated neural networks that instead target long-term memory.

The fact that the reservoir is chosen once and for all, instead of constantly being updated, eases the theoretical analysis of these networks. As such, by means of a succession of works (e.g., (Jaeger, 2001a; Ganguli et al., 2008; Strauss et al., 2012)), many qualitative aspects of ESN's have been fairly well understood. In particular, it has been made clear that both the operator norm and the spectral radius of the connectivity matrix play a central role in the ESN performance, that normal versus non-normal connectivity matrices convey strikingly different behavior, etc. However, to the best of the authors' knowledge, there has never been an attempt to turn these qualitative considerations into concrete quantitative figures.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

The objective of the present article is to provide a first theoretical analysis of the (mean-square error) performance of *linear* ESN's with *internal noise* for both the training and the testing tasks (an illustration of such a network is depicted in Figure 1). To this end, we shall leverage recent tools from the field of random matrix theory (the applications of which are so far almost not existent in neural networks) and shall conveniently work under the assumption that both the reservoir size n and the training (or testing) duration T (or $\hat{T}$) are large and commensurable. The large dimensional framework will induce concentration of measure properties that bring asymptotic determinism in the performance of the random outputs. These results will take closed-form expressions which, if not completely explicit, are nonetheless easily interpreted. Turning the connectivity matrices into (large) random realizations of simple random matrix models, we then further simplify these expressions that ultimately lead to elementary formulas.

Among the noteworthy outcomes of these theoretical results, we shall understand deeply the impact of normal versus non-normal matrices in the training and testing performance as well as the impact of the internal noise as not only an overfitting shield but also as a major driver of robustness to erratic data. We shall also propose a new connectivity matrix design, coined *multi-memory* matrix, the performance of which is fully understood, that helps handling time series with multiple-scale memory properties.

Note in passing that, beyond their machine learning

attractiveness for obvious network stability reasons (although ridge regularized networks without internal noise are often preferred), ESN's with in-network noise appropriately model biological short-term memory in the brain. Our results may then also embrace both the neurophysiology and neurocomputation fields.

In the remainder of the article, we shall first introduce the necessary random matrix toolbox used throughout the article (and so far not conventional in machine learning), before addressing the question of estimating the performance of network training and testing tasks. Applications of our results will be discussed next, before closing on a reflective discussion. The technical results found throughout the paper are proved in an extended version of the present article.

2. The Random Matrix Framework

Before delving into the concrete ESN performance, we shall first consider the elementary objects under study from a random matrix perspective.

We assume here an n-node ESN with connectivity matrix $W \in \mathbb{R}^{n\times n}$, source-to-reservoir vector $m \in \mathbb{R}^n$, states $x_t \in \mathbb{R}^n$, $t = -\infty, \ldots, \infty$, and in-network noise $\eta\varepsilon_t \sim \mathcal{N}(0, \eta^2 I_n)$, fed by a scalar input $u_t \in \mathbb{R}$. The state evolution equation follows:

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

The ESN will be trained for a period T and tested for a period $\hat{T}$, using a least-square regression approach. Denoting $X = [x_0, \ldots, x_{T-1}] \in \mathbb{R}^{n\times T}$, it shall appear in Section 3 that the mean-square error performance in training relies fundamentally on the matrices

$$Q_\gamma \equiv \left(\frac{1}{T}XX^\mathsf{T} + \gamma I_n\right)^{-1}$$
$$\tilde{Q}_\gamma \equiv \left(\frac{1}{T}X^\mathsf{T}X + \gamma I_T\right)^{-1} \tag{2}$$

for $\gamma > 0$. These matrices, respectively called resolvent and co-resolvent of the Gram matrix $\frac{1}{T}XX^\mathsf{T}$ in the operator theory jargon, have been extensively used in random matrix theory for various models of X (Bai & Silverstein, 2009; Pastur & Ŝerbina, 2011) with multiple applications to engineering notably (Couillet & Debbah, 2011). The model of X defined through (1) is not part of those models studied in classical random matrix works, but the technical tools exist to handle it. Precisely, we shall use here the Gaussian framework devised by Pastur (Pastur & Ŝerbina, 2011), based on an integration-by-parts formula for Gaussian random variables and the so-called Nash–Poincaré inequality.

To this end, we first need elementary growth assumptions on the size n and the periods T and $\hat{T}$.

Assumption 1 *Define the matrix $A = MU$ with $M = [m, Wm, \ldots, W^{T-1}m]$ and $U = T^{-\frac{1}{2}}\{u_{j-i}\}_{i,j=0}^{T-1}$. Then, as $n \to \infty$,*

1. *$n/T \to c \in (0, \infty)$ and $n/\hat{T} \to \hat{c} \in [0, \infty)$*
2. *$\limsup_n \|W\| < 1$*
3. *$\limsup_n \|AA^\mathsf{T}\| < \infty$*

with $\|\cdot\|$ the operator norm.

For notational convenience, we define the relation $X_n \leftrightarrow Y_n$ to mean that the (random or deterministic) matrices X_n, Y_n satisfy $a_n^\mathsf{T}(X_n - Y_n)b_n \to 0$, almost surely, for all deterministic unit norm vectors a_n, b_n. Under Assumption 1, applying the aforementioned Gaussian framework, we have the following result.

Theorem 1 (Deterministic Equivalent) *Let Assumption 1 hold. For $\gamma > 0$, and with Q_γ and $\tilde{Q}_\gamma$ defined in (2), as $n \to \infty$,*

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

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

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

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

Theorem 1 precisely states that any *random* bilinear form of the type $a_n^\mathsf{T}Q_\gamma b_n$ for deterministic a_n, b_n, can be well approximated, for all large n, T, by the deterministic quantity $a_n^\mathsf{T}\bar{Q}_\gamma b_n$. As shall be seen in Section 3, this will provide us with a deterministic approximation of the training performance of ESN's. Note in passing that Theorem 1 is strongly reminiscent of the earlier results (Hachem et al., 2006) in a close but different context, and are thus not surprising.

The testing phase performance is more involved and does not solely rely on a single Q_γ-type matrix. This is because this phase involves both the trained dataset X and the newly observed states $\hat{X} = [\hat{x}_0, \ldots, \hat{x}_{\hat{T}-1}]$,

where $\hat{x}_t = x_{t+L}$ for some $L \gg T$ (we assume here a post-training wash-out step for simplicity). Surprisingly enough, we shall not need the advanced statistics of $\hat{X}$ explicitly (as opposed to X as seen previously) but shall rather exploit its independence from X. As such, the quantities at stake here are the matrices

$$\frac{1}{\sqrt{T}} Q_\gamma X, \quad \frac{1}{T} X^\mathsf{T} Q_\gamma B Q_\gamma X$$

for B any symmetric matrix independent of X such that $\limsup_n \|B\| < \infty$. Precisely, we have the following second deterministic equivalents.

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

$$Q_\gamma \frac{1}{\sqrt{T}} X \leftrightarrow \bar{Q}_\gamma A (I_n + \eta^2 R_\gamma)^{-1}$$

$$\frac{1}{T} X^\mathsf{T} Q_\gamma B Q_\gamma X \leftrightarrow \eta^2\gamma^2 \bar{\tilde{Q}}_\gamma G_\gamma^{[B]} \bar{\tilde{Q}}_\gamma + P^\mathsf{T} \bar{Q}_\gamma \left[B + \tilde{G}_\gamma^{[B]} \right] \bar{Q}_\gamma P$$

with $P = A(I_n + \eta^2 R_\gamma)^{-1}$ and where $G_\gamma^{[B]}, \tilde{G}_\gamma^{[B]}$ are solutions to

$$G_\gamma^{[B]} = \left\{ \frac{1}{T} \operatorname{tr}\left(S_{i-j} \bar{Q}_\gamma \left[B + \tilde{G}_\gamma^{[B]} \right] \bar{Q}_\gamma \right) \right\}_{i,j=1}^{T}$$

$$\tilde{G}_\gamma^{[B]} = \sum_{q=-\infty}^{\infty} \eta^4\gamma^2 \frac{1}{T} \operatorname{tr}\left(J^q \bar{\tilde{Q}}_\gamma G_\gamma^{[B]} \bar{\tilde{Q}}_\gamma \right) S_q.$$

Equipped with these technical results, we are now in position to provide our main contribution to the asymptotic performance of ESN as $n, T, \hat{T} \to \infty$.

3. Asymptotic Performance

Recall that the ESN under study is defined by the state equation (1). We shall successively discuss the performance of the training and testing steps of this linear ESN.

3.1. Training Performance

In the training phase, one wishes to map an input sequence $u = [u_0, \ldots, u_{T-1}]^\mathsf{T}$ to a corresponding known output sequence $r = [r_0, \ldots, r_{T-1}]^\mathsf{T}$. To this end, we shall enforce the reservoir-to-sink connections of the network, gathered into a vector $\omega \in \mathbb{R}^n$ and depicted in color in Figure 1, so to minimize the quadratic reconstruction error

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

The solution to this classical problem is to take ω to be the least-square regressor

$$\omega \equiv \begin{cases} (XX^\mathsf{T})^{-1} X r & , \; T > n \\ X(X^\mathsf{T} X)^{-1} r & , \; T \leq n. \end{cases} \tag{3}$$

To such an ω are associated an $E_\eta(u; r)$ which it is convenient to see here as

$$E_\eta(u, r) = \begin{cases} \lim_{\gamma \downarrow 0} \gamma \frac{1}{T} r^\mathsf{T} \tilde{Q}_\gamma r & , \; T > n \\ 0 & , \; T \leq n \end{cases} \tag{4}$$

where $\tilde{Q}_\gamma$ was defined in (2).

Applying Theorem 1 in the limit where $\gamma \to 0$, we have the following first limiting performance result.

Proposition 1 (Training MSE) *Let Assumption 1 hold and let $r \in \mathbb{R}^T$ be a vector of Euclidean norm $O(\sqrt{T})$. Then, with $E_\eta(u, r)$ defined in (4), as $n \to \infty$,*

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

where, for $c < 1$,

$$\tilde{\mathcal{Q}} \equiv \left(I_T + \mathcal{R} + \frac{1}{\eta^2} A^\mathsf{T} \tilde{\mathcal{R}}^{-1} A \right)^{-1}$$

and $\mathcal{R}$, $\tilde{\mathcal{R}}$ are solutions to[1]

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

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

Although seemingly not simple, note that, by writing

$$A^\mathsf{T} \tilde{\mathcal{R}}^{-1} A = U^\mathsf{T} \left\{ m^\mathsf{T} (W^{i-1})^\mathsf{T} \tilde{\mathcal{R}}^{-1} W^{j-1} m \right\}_{i,j=1}^{T} U$$

the matrix $\tilde{\mathcal{Q}}$ involved in the asymptotic expression for $E_\eta(u, r)$ clearly features independently:

- the input data matrix U composed in columns of the successive delayed versions of the vector $T^{-\frac{1}{2}} [u_{-(T-1)}, \ldots, u_{T-1}]^\mathsf{T}$;
- the network structuring matrices $\mathcal{R}$ and $(W^{i-1})^\mathsf{T} \tilde{\mathcal{R}}^{-1} W^{j-1}$;
- the factor η^{-2}, not present in $\mathcal{R}, \tilde{\mathcal{R}}$, which trades off the need for *regularizing* the ill-conditioned matrix $A^\mathsf{T} \tilde{\mathcal{R}}^{-1} A$ (through the matrix M in A) and the need to increase the weight of the information-carrying matrix $A^\mathsf{T} \tilde{\mathcal{R}}^{-1} A$ (through the matrix U in A).

[1] $\mathcal{R}$ and $\tilde{\mathcal{R}}$ are rigorously the limits of R_γ and $\gamma \tilde{R}_\gamma$ from Theorem 1, respectively, as $\gamma \downarrow 0$.

Note in particular that, since $\|W\| < 1$, the matrix $\{m^\mathsf{T}(W^{i-1})^\mathsf{T}\tilde{\mathcal{R}}^{-1}W^{j-1}m\}_{i,j=1}^T$ has an exponentially decaying profile down the rows and columns (essentially decaying with $i+j$). As such, all but the first few columns of $\tilde{\mathcal{R}}^{-\frac{1}{2}}MU$ vanish as n, T grow large, providing us with a first testimony of the ESN short term memory specificity, since only the first columns of U (i.e., the first delays of $\{u_t\}$) are accounted for. The matrix $\tilde{\mathcal{R}}^{-\frac{1}{2}}M$ then plays the important role of tuning the memory decay.

A particularly interesting scenario is when $c = 0$ (i.e., $n/T \to 0$). In this case, it is immediate that $\mathcal{R} = 0$ and $\tilde{\mathcal{R}} = S_0 = \sum_{k\geq 0} W^k(W^k)^\mathsf{T}$, so that

$$E_\eta(u,r) \equiv \frac{\eta^2}{T} r^\mathsf{T}\left(\eta^2 I_T + U^\mathsf{T}M^\mathsf{T}S_0^{-1}MU\right)^{-1} r$$

where $[M^\mathsf{T}S_0^{-1}M]_{kk} = m^\mathsf{T}(W^{k-1})^\mathsf{T}S_0^{-1}W^{k-1}m$ is recognized to be $J(k-1)$ with J the *Fisher memory curve* introduced in (Ganguli et al., 2008). The latter was shown to provide a qualitative measure of the capacity of the ESN to store a k-step delayed information. In Proposition 1, the notion is generalized to account for the finiteness of T with respect to n (i.e., $c > 0$) and, even when $c = 0$, conveys a non trivial importance to the off-diagonal terms $m^\mathsf{T}(W^{i-1})^\mathsf{T}\tilde{\mathcal{R}}^{-1}W^{j-1}m$, $i \neq j$.

3.2. Testing Performance

Having considered the training performance in Section 3.1, we now assume ω given through its definition (3). To avoid unnecessary complications, we shall stick here to the case where $c < 1$ and leave the case $c > 1$ to the extended version of the article. We wish to evaluate the performance in testing new data inputs $\hat{u} = [\hat{u}_0, \ldots, \hat{u}_{\hat{T}-1}]^\mathsf{T}$ which ought to be mapped by the ESN to the desired output $\hat{r} = [\hat{r}_0, \ldots, \hat{r}_{\hat{T}-1}]^\mathsf{T}$. Thus our next quantity of interest is the mean square error

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

where $\hat{X} = [\hat{x}_0, \ldots, \hat{x}_{\hat{T}-1}]$ with $\hat{x}_t = x_{t+L}$ for some L sufficiently large (larger than $2T$, say) to ensure approximate independence between $\{\hat{x}_t\}$ and $\{x_t\}$. That is, we assume a sufficiently long wash-out period between training and testing (as conventionally done). Alternatively, one may merely reinitialize the network after training and generate a sufficiently long dry-run period prior to testing. Similarly, we shall denote next $\hat{A} = \hat{M}\hat{U}$ with $\hat{U}_{ij} = \hat{u}_{j-i}$ and $\hat{M} = [m, \ldots, W^{\hat{T}-1}m]$.

Developing the expression of $\hat{E}_\eta(u,r;\hat{u},\hat{r})$, it is convenient to observe that

$$\hat{E}_\eta(u,r;\hat{u},\hat{r}) = \frac{1}{\hat{T}}\|\hat{r}\|^2 + \lim_{\gamma\downarrow 0}\frac{1}{T^2\hat{T}} r^\mathsf{T}X^\mathsf{T}Q_\gamma\hat{X}\hat{X}^\mathsf{T}Q_\gamma Xr - \lim_{\gamma\downarrow 0}\frac{2}{T\hat{T}}\hat{r}^\mathsf{T}\hat{X}^\mathsf{T}Q_\gamma Xr.$$

We may then straightforwardly apply Theorem 2 in the limit of vanishing γ to retrieve the testing counterpart of Proposition 1 as follows.

Proposition 2 (Testing MSE) *Let Assumption 1 hold with $c < 1$ and let $\hat{r} \in \mathbb{R}^{\hat{T}}$ be a vector of Euclidean norm $O(\sqrt{\hat{T}})$. Then, as $n \to \infty$, with the notations of Proposition 1,*

$$\hat{E}_\eta(u,r;\hat{u},\hat{r}) \leftrightarrow \left\|\frac{1}{\eta^2}\hat{A}^\mathsf{T}\mathcal{Q}\mathcal{P}\frac{r}{\sqrt{T}} - \frac{\hat{r}}{\sqrt{\hat{T}}}\right\|^2 + \frac{1}{T}r^\mathsf{T}\tilde{\mathcal{Q}}\mathcal{G}\tilde{\mathcal{Q}}r + \frac{1}{\eta^2 T}r^\mathsf{T}\mathcal{P}^\mathsf{T}\mathcal{Q}\left[S_0 + \tilde{\mathcal{G}}\right]\mathcal{Q}\mathcal{P}r$$

with $\mathcal{P} = A(I_T + \mathcal{R})^{-1}$ and $(\mathcal{G}, \tilde{\mathcal{G}})$ solution to

$$\mathcal{G} = c\left\{\frac{1}{n}\operatorname{tr}\left(S_{i-j}\tilde{\mathcal{R}}^{-1}\left[S_0 + \tilde{\mathcal{G}}\right]\tilde{\mathcal{R}}^{-1}\right)\right\}_{i,j=1}^T$$

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

Note that the deterministic approximation for $\hat{E}_\eta(u,r;\hat{u},\hat{r})$ in Proposition 2 may be divided into a first term involving $\hat{u}, \hat{r}$ and the next two terms only involving u, r. As such, once training is performed, only the former term may alter the asymptotic performance. Again, the case $c = 0$ leads to simpler expressions, as $\mathcal{G} = 0$ and $\tilde{\mathcal{G}} = 0$ in this case, so that

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

In order to validate the results of Section 3.1 and Section 3.2, we provide in Figure 2 an example of simulated versus asymptotic performance for a prediction task over the popular Mackey–Glass model (Glass & Mackey, 1979). Disregarding for the moment the difference between the two displayed theoretical curves, note importantly that the accuracy of the approximation, while increasing as it should for larger values of $n, T, \hat{T}$, may strongly decrease as $\eta^2 \to 0$. This is an expected outcome (from deeper mathematical

analysis of the proofs of our propositions) which is reminiscent of the ESN instability observed and documented in (Jaeger, 2001b) as the internal noise vanishes. As a matter of fact, it can be shown that one needs $\eta^2 \gg n^{-\frac{1}{2}}$ for a theoretical guarantee that the approximation is accurate.

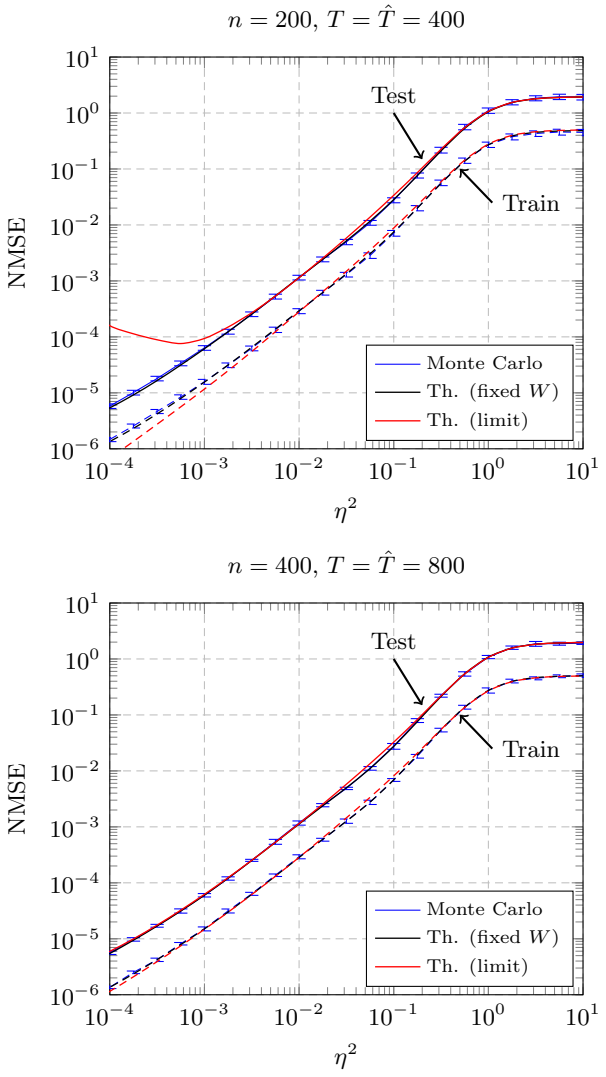

Figure 2. Training and testing (normalized) MSE for the Mackey Glass one-step prediction, W defined as in Figure 3, $n = 200$, $T = \hat{T} = 400$ (top) and $n = 400$, $T = \hat{T} = 800$ (bottom). Comparison between Monte Carlo simulations (Monte Carlo) and theory from Propositions 1–2 (Th. (fixed W)) or Corollary 2 (Th. (limit)).

4. Applications

Let us now move to particular scenarios where Propositions 1–2 either greatly simplify or convey new insights.

4.1. Random W matrices

We consider here the scenario where, instead of being considered deterministic, we take W to be a single realization of an elementary (large dimensional) random matrix.

4.1.1. Real Haar W.

First consider the scenario where $W = \sigma Z$ with Z random orthogonal with statistical invariance by left- and right-multiplication by orthogonal matrices, i.e., Z is a random real Haar matrix. Then one can determine an explicit expression for the matrices $\mathcal{R}$, $\tilde{\mathcal{R}}$, $\mathcal{G}$, and $\tilde{\mathcal{G}}$ involved in Propositions 1–2. In particular, we have the following result (also depicted in Figure 2).

Corollary 1 (Haar W, $c < 1$) *Let $W = \sigma Z$ with Z random real Haar and m be independent of W with $\|m\| = 1$. Then, under Assumption 1 with $c < 1$,*

$$E_\eta(u,r) \leftrightarrow (1-c)\frac{1}{T} r^\mathsf{T} Q r$$

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

where $Q = (I_T + \frac{1}{\eta^2} U^\mathsf{T} D U)^{-1}$, while $D \in \mathbb{R}^{T \times T}$ and $\hat{D} \in \mathbb{R}^{\hat{T} \times \hat{T}}$ are diagonal with

$$D_{ii} = \hat{D}_{ii} = (1-\sigma^2)\sigma^{2(i-1)}.$$

We clearly see through Corollary 1 the impact of σ which weighs through D_{ii} the successive delay vectors $U_{\cdot,i}$ starting from $i = 1$ for zero delay. This is again reminiscent of the works (Ganguli et al., 2008) where the diagonal elements of D were understood qualitatively as a memory curve, with the property that $\sum_{i \geq 1} D_{ii} = 1$, so that the ESN *allocates* a total unit amount of memorization capabilities across the successively delayed versions of u.

This observation inspires the generalization of Corollary 1 to a less obvious, although desirable, structure for W. Indeed, note that, with $W = \sigma Z$ and Z real Haar, memory is allocated according to the exponential decay function $k \mapsto \sigma^k$, thus only allowing for a "single mode" memory, i.e., r_t should be an exponentially decaying function of u_{t-k}. If instead, as is more common, r_t is a more elaborate function of both close past u_{t-k} but also of further past $u_{t-k'}$ values, then it might be appropriate for the ESN not to get restricted to a single $k \mapsto \sigma^k$ decay profile.

As such, we propose to consider the matrix $W = \mathrm{diag}(\sigma_1 Z_1, \ldots, \sigma_\ell Z_\ell)$ (with diag the block-diagonal operator), where the matrices $Z_i \in \mathbb{R}^{n_i \times n_i}$ are independent real Haar matrices of given sizes and $\sigma_i > 0$ assume different values across i. In this case, we have the following natural extension of Corollary 1.

Corollary 2 (Multi-memory W) *Let the assumptions of Corollary 1 hold but for $W = \mathrm{diag}(\sigma_1 Z_1, \ldots, \sigma_\ell Z_\ell)$ with $Z_i \in \mathbb{R}^{n_i \times n_i}$ independent real Haar matrices, $\sum_i n_i = n$, and $\sigma_i > 0$. Then the conclusions of Corollary 1 remain valid but for*

$$D_{ii} = \hat{D}_{ii} = \frac{\sum_{j=1}^{\ell} n_j \sigma_j^{2(i-1)}}{\sum_{j=1}^{\ell} n_j (1-\sigma_j^2)^{-1}}.$$

Figure 3 depicts the function $k \mapsto D_{kk}$ (which again may be thought of as a memory curve) for the "multimemory" matrix W of Corollary 2 versus elementary random Haar matrices with single σ values. Note the evolution of the function slope which successively embraces the memory curves of each individual Haar matrix composing W.

Following the same example, we now compare in Figure 4 the testing performance for the Mackey–Glass one-step prediction task under the multi-memory W versus the composing Haar W_i^+. It is interesting to see here that the multi-memory matrix is almost uniformly more powerful than the composing matrices and that it matches the performance of the best among the latter. This suggests the possibility to use such a matrix structure in scenarios where the experimenter has little knowledge about the particularly adequate choice of σ in a mere Haar model for W.

4.1.2. Normal and non normal i.i.d. W.

We subsequently move to the next natural model for W, that is W composed of i.i.d. zero mean entries with or without Hermitian symmetry. In the latter case, the study is similar to that of the real Haar case and shall lead to the same result as Corollary 1 but for a different profile of the diagonal entries of the matrix D. As for the former symmetrical case (referred to in random matrix theory as the Wigner case), it leads to a more involved (non explicit) expression for $\mathcal{R}$, which assumes a strikingly different structure than when W is non-symmetric (in the later case, $\mathcal{R}$ is proportional to the identity matrix). Visually, we find in the large n limit the structure depicted in Figure 5.

When placed in the context of Proposition 1, the observed checkerboard structure for the Wigner case suggests an inappropriate spread of the reservoir energy

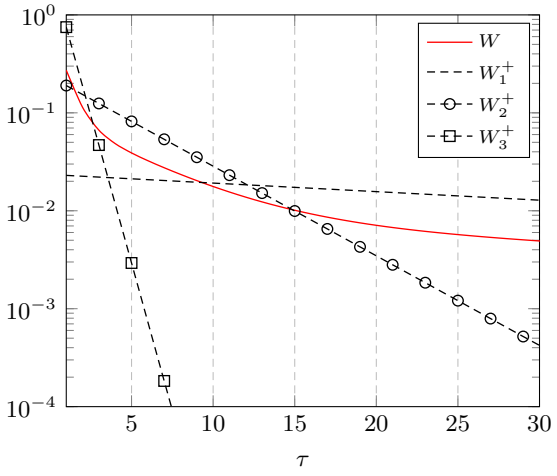

Figure 3. Memory curve $\tau \mapsto D_{\tau,\tau}$ for $W = \mathrm{diag}(W_1, W_2, W_3)$, $W_j = \sigma_j Z_j$, $Z_j \in \mathbb{R}^{n_j \times n_j}$ Haar distributed, $\sigma_1 = .99$, $n_1/n = .01$, $\sigma_2 = .9$, $n_2/n = .1$, and $\sigma_3 = .5$, $n_3/n = .89$. The matrices W_i^+ are defined by $W_i^+ = \sigma_i Z_i^+$, with $Z_i^+ \in \mathbb{R}^{n \times n}$ Haar distributed.

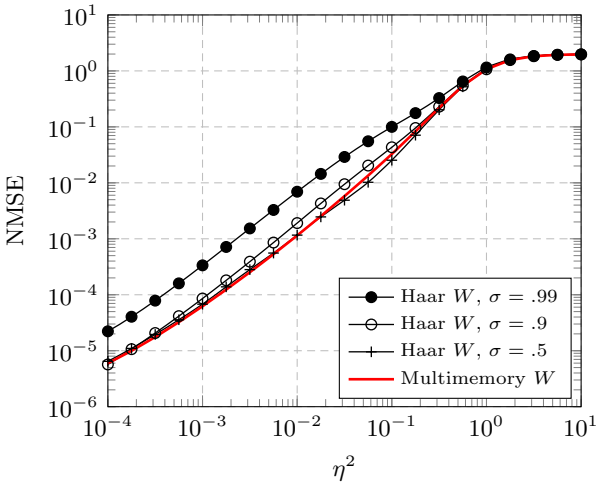

Figure 4. Testing (normalized) MSE for the Mackey Glass one-step ahead task, W (multimemory) versus $W_1^+ = .99Z_1^+$, $W_2^+ = .9Z_2^+$, $W_3^+ = .5Z_3^+$ (with Z_i^+ Haar distributed) all defined as in Figure 3, $n = 400$, $T = \hat{T} = 800$.

when it comes to fulfilling pure delay tasks. This is indeed observed numerically with strong performance losses already induced by elementary memorization tasks. A particular example is depicted in Figure 6 where, again for the Mackey–Glass input dataset *but* for a memorization task (consisting in recalling τ past outputs rather than predicting future outputs). This study and the observed performance results suggest an outstanding performance advantage of non-normal versus normal matrix structures, which might deserve

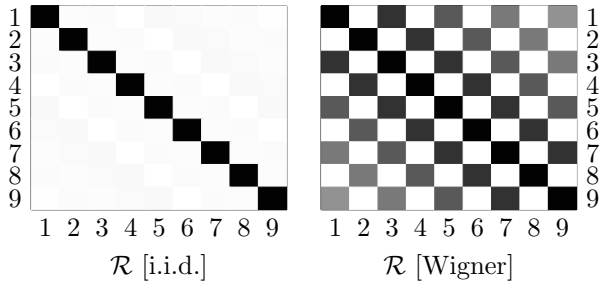

Figure 5. Upper 9×9 part of $\mathcal{R}$ for $c = 1/2$ and $\sigma = 0.9$ for W with i.i.d. zero mean Gaussian entries (left) and W Gaussian Wigner (right). Linear grayscale representation with black being 1 and white being 0.

deeper future investigation.

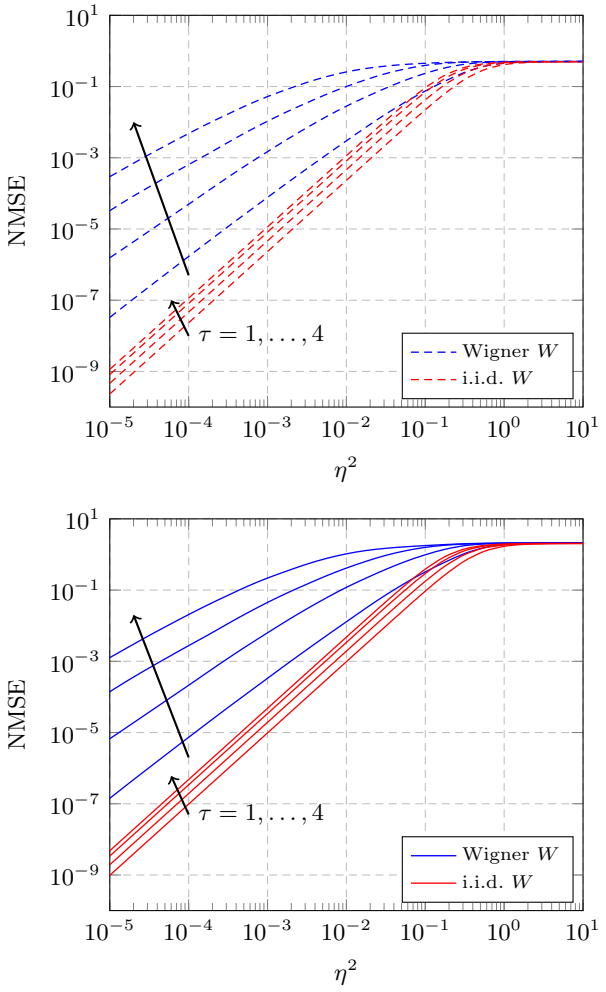

Figure 6. Training (top) and testing (bottom) performance of a τ-delay task for $\tau \in \{1, \ldots, 4\}$ compared for i.i.d. W versus Wigner W, $\sigma = .9$ and $n = 200$, $T = \hat{T} = 400$ in both cases (here on the Mackey-Glass dataset).

4.2. In-network noise and robustness

We complete this section by the investigation of a particular scenario in which we assume that input data may be corrupted by extra noise in the testing dataset. This models the fact that one may often possess a large bank of "clean" data to train an ESN but that the reality of test data can sometimes be somewhat, if not strongly, different. Our toy model here consists in considering that an extra Gaussian noise is added to the data $\hat{u}$ with probability p on each sample $\hat{u}_t$, while the output $\hat{r}$ is still expected to be consistent with the noiseless version of $\hat{u}$.

From a proper theoretical analysis of the result of Proposition 2, conducted precisely in the extended version of the article, we claim that, letting s^2 be the aforementioned additive noise variance and assuming that $\hat{E}_\eta \to 0$ as $\eta \to 0$ and $s = 0$ (a scenario that can be enforced by letting e.g., r be a linear combination of finitely many past inputs of u), then there exists a trade-off by which too small values of η^2 induce an increase in the mean square error (all the more that s^2 is large) while large values of η^2 induce too much internal noise. There thus exists an optimal value for η^2 which minimizes the testing MSE.

This phenomenon is depicted in a concrete scenario in Figure 7, still for the same Mackey–Glass one-step ahead prediction task. A particular realization of a random Mackey–Glass time series is also presented in Figure 8, which clearly highlights the robustness strength of internal noise.

5. Concluding Remarks

The random matrix framework introduced in this article brings new light to the actual performance of *linear* echo-state networks with *internal noise* (and more generally recurrent networks), where past works merely provided insights based on incomplete considerations of the network processing (such as information theoretic metrics of the reservoir information). Our results make it clear what levers should be tuned and optimized upon when designing these networks. Although not presently discussed, an outcome of this study (documented in an extended article) contradicts some beliefs, such as that suggesting that it is appropriate to take m as one of the leading eigenvectors of W; we can prove that this choice necessarily leads to poor mean square error performance. But aside from purely theoretical considerations, our results also allow for a fast evaluation of the ESN performance without requiring extensive Monte Carlo simulations which we believe experimenters should find convenient.

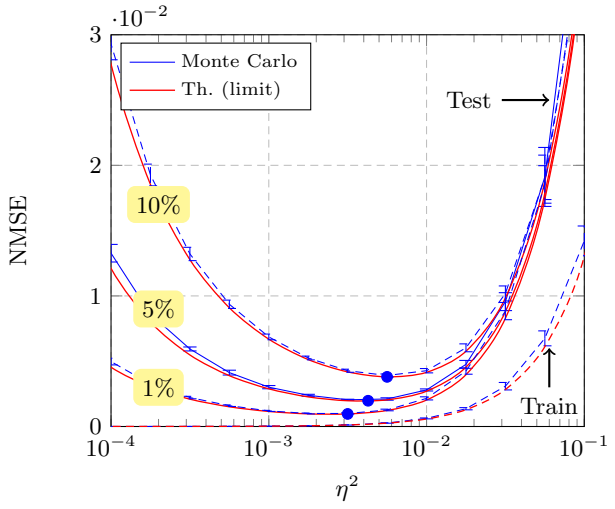

Figure 7. Testing (normalized) MSE for the Mackey-Glass one-step ahead task with 1% or 10% impulsive $\mathcal{N}(0, .01)$ noise pollution in test data inputs, W Haar with $\sigma = .9$, $n = 400$, $T = \hat{T} = 1000$. Circles indicate the NMSE theoretical minima. Error bars indicate one standard deviation of the Monte Carlo simulations.

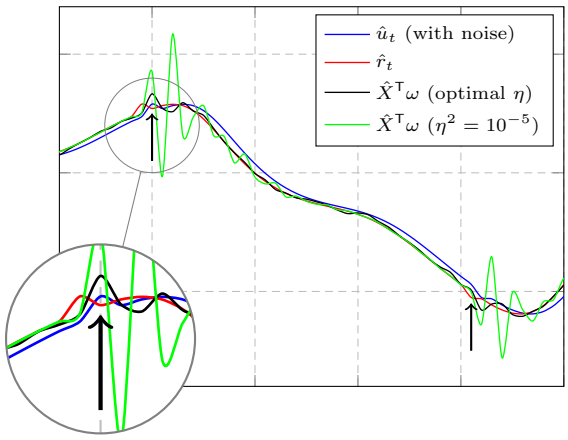

Figure 8. Realization of a 1% $\mathcal{N}(0, .01)$-noisy Mackey-Glass sequence versus network output, W Haar with $\sigma = .9$, $n = 400$, $T = \hat{T} = 1000$. In magnifying lenses, points of added impulsive noise.

We stressed above the words *linear* and *internal noise*, as they are often at the center of debates in the field. Regarding internal noise, while being an appropriate model assumption in biological networks, it is often regarded as artificial in machine learning (where a regularized least square ω is chosen to stabilize the network). Since large networks induce concentration of measure phenomena that stabilize the MSE performance of the network, we forcefully believe that internal noise (leading to random but equally performing outputs) are instead more desirable than deterministic (and thus statically biased) outputs. In the full version of this article, comparisons are performed between both cases. But the utmost limiting aspect of the present work rather lies in the linear character of the state equation (1). It is known, more from experiments and insights than theory, that breaking the linear frontier brings vastly more interesting properties to neural networks, with in particular the possibility for $\|W\|$ to exceed one. A necessary next step of the random matrix framework may be a coupling to mean field considerations which may adequately handle the behavior of non linear versions of echo-state networks.

References

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

Couillet, R. and Debbah, M. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, New York, NY, USA, first edition, 2011. ISBN 978-1107011632.

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

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

Hachem, W., Loubaton, P., and Najim, J. The empirical distribution of the eigenvalues of a Gram matrix with a given variance profile. *Annales de l'Institut H. Poincaré*, 42(6):649–670, 2006.

Jaeger, H. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148:34, 2001a.

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

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

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