TRAINING PERFORMANCE OF ECHO STATE NEURAL NETWORKS

Romain Couillet[1], *Gilles Wainrib*[2], *Harry Sevi*[3], *Hafiz Tiomoko Ali*[1]

[1]CentraleSupélec, University of Paris–Saclay, France
[2]ENS Ulm, Paris, France
[3]ENS Lyon, Lyon, France

ABSTRACT

This article proposes a first theoretical performance analysis of the training phase of large dimensional linear echo-state networks. This analysis is based on advanced methods of random matrix theory. The results provide some new insights on the core features of such networks, thereby helping the practitioner when using them.

Index Terms— Neural networks, ESN, random matrix theory.

1. INTRODUCTION

In the realm of both artificial and biological neural networks, one usually differentiates long- and short-term memory networks. While the former are inherently based on a dedicated rewiring of the network during training phase (i.e., the network edges adapt to the input-to-output training by means of backpropagation) and work in a feed-forward manner, the latter do not change their connectivity matrix but maintain past-input information within the network via self-connection (thereby being recurrent networks rather than feed-forward networks) so that, for energy conservation reasons (inputs are continuously fed into the networks), the past-input memory decays exponentially fast in this case. These networks, less considered than their feed-forward counterparts, have recently been reinstated by Jaeger [1] under the name of *echo-state networks* (ESN), who defends their performance superiority in short-term memory settings [2].

A striking feature of ESN's is that the (fixed) network connectivity matrix can be chosen as a random matrix and that choice is even a very satisfactory one. This observation and intuition from random matrix theory suggests the possibility for a much expected theoretical analysis of such networks. The objective of the present article is to lay down such theoretical bases to propose a first study of the training performance of these networks.

Precisely, we shall consider an echo-state network of n nodes and will prove that, as both the training sequence length T and n grow large simultaneously, the mean-square error

Couillet and Tiomoko Ali's work is supported by the ANR RMT4GRAPH ANR-14-CE28-0006.

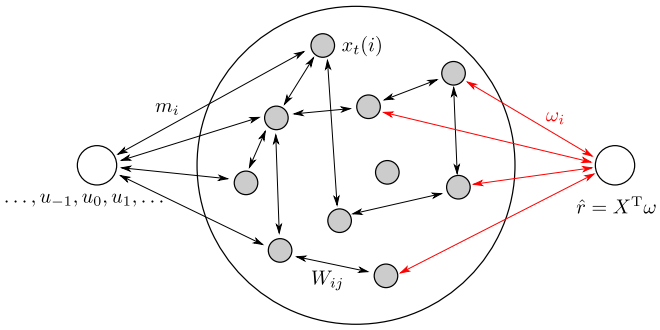

Fig. 1. Echo-state neural network.

performance of the task training procedure tends to be deterministic in the limit, for all well-behaved connectivity matrices $W \in \mathbb{R}^{n\times n}$. Then, specifying W to belong to specific classes of random matrices, we will further show that the aforementioned deterministic behavior takes on a very simple expression that exhibits the salient features of the ESN. Many consequences can be drawn of this analytical formula, some of which will be discussed here.

2. MAIN RESULTS

2.1. Generic W

Consider a neural network of n nodes having at each time t a joint state $x_t \in \mathbb{R}^n$. We assume the following state evolution

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

for some initial (say empty) state at large negative t, with $W \in \mathbb{R}^{n\times n}$ the connectivity matrix, $\ldots, u_{-1}, u_0, u_1, \ldots \in \mathbb{R}$ the network (scalar) inputs, m the input-to-network connectivity vector, and $\eta\varepsilon_t \in \mathbb{R}^n$ an additional in-network noise of amplitude $\eta > 0$. A visual representation of such a network is depicted in Figure 1.

We seek here to evaluate the performance of the so-called *training task* consisting in relating the input sequence $\{u_t\}$ to a pre-determined output sequence $r_0, \ldots, r_{T-1}$ from a given linear combination of the network states $x_0, \ldots, x_{T-1}$ for

a duration T. This is traditionally performed using a least-square regression. Precisely, we shall define

$$\omega \equiv \begin{cases} X\left(X^\mathsf{T}X\right)^{-1} r & , T \leq n \\ \left(XX^\mathsf{T}\right)^{-1} Xr & , T > n \end{cases}$$

where $r = [r_0, \ldots, r_{T-1}]^\mathsf{T}$ and $X = [x_0, \ldots, x_{T-1}] \in \mathbb{R}^{n \times T}$, and consider $X^\mathsf{T}\omega$ as the least-square estimate of r with mean-square estimation error

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

This quantity is clearly identically zero if $T \leq n$ and we shall thus only consider the non-trivial case where $T > n$.

Our objective is to understand the behavior of $E_\eta(u, r)$ defined in (2) in the limit where $n, T \to \infty$, for an arbitrary (and then for a random) matrix W. The underlying idea behind our approach is that the noise ε_t, which regularizes the otherwise unstable network, will tend to concentrate as n, T grow large, thereby leading to a (much desired) more deterministic behavior of the network as a whole. Considering X as a random matrix through these ε_t's, we shall provide an asymptotically consistent estimate of $E_\eta(u, r)$.

To ensure proper conditioning in the large n, T limit, the following assumption is needed.

Assumption 1 (Large n, T limit) *As $n, T \to \infty$,*

1. $\lim_n \frac{n}{T} = c \in [0, 1)$
2. $\limsup_n \|W\| < 1$
3. $\limsup_n \|AA^\mathsf{T}\| < \infty$

where $A = MU$, $M = [m, Wm, \ldots, W^{T-1}m] \in \mathbb{R}^{n \times T}$ and $U = \{u_{i-j}\}_{1 \leq i,j \leq T}$.

Item (2) is a (often too stringent) sufficient condition to ensure network stability, while Item (3) is merely a not-too demanding technical conditioning. The aforementioned matrix U will play a key role in what follows. Note that its columns constitute delayed versions of the sequence $u_{-(T-1)}, \ldots, u_{T-1}$ and thus can be seen as successive snapshots of the (T-step behind) network memory as time elapses.

With Assumption 1 in place, we have our first main result.

Theorem 1 (Asymptotic MSE) *Let Assumptions 1 hold and let $r \in \mathbb{R}^T$ be of $O(\sqrt{T})$ Euclidean norm. Then, with $E_\eta(u, r)$ defined in* (2)*, as $n \to \infty$,*

$$\left| E_\eta(u, r) - \frac{1}{T} r^\mathsf{T} \tilde{\mathcal{Q}} r \right| \to 0$$

almost surely, where $\tilde{\mathcal{Q}} \equiv (I_T + \mathcal{R} + \eta^{-2} A^\mathsf{T} \tilde{\mathcal{R}}^{-1} A)^{-1}$, with $(\mathcal{R}, \tilde{\mathcal{R}})$ a solution to the implicit system

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

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

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

Sketch of Proof 1 *The idea behind the proof consists in applying the so-called deterministic equivalent method (see e.g., [3, Chapter 6]) of random matrix theory on the random matrix $\frac{1}{T} X^\mathsf{T} X$. That is, defining $\tilde{Q}_\gamma = (\frac{1}{T} X^\mathsf{T} X - \gamma I_T)^{-1}$ for all $\gamma > 0$, the resolvent of $\frac{1}{T} X^\mathsf{T} X$, the method consists in determining a matrix $\bar{\tilde{Q}}_\gamma$ such that, as $n, p \to \infty$, $a^\mathsf{T}(\tilde{Q}_\gamma - \bar{\tilde{Q}}_\gamma) b$ almost surely vanish for all deterministic and bounded norm a, b vectors. This is performed via the Gaussian integration-by-parts and Nash–Poincaré inequality framework devised by Pastur in [4, Chapter 2]. Once this is achieved, we show that the aforementioned convergence still holds when $\gamma = 0$, which is valid as long as $c < 1$. Calling $\tilde{\mathcal{Q}}$ the limit of $\bar{\tilde{Q}}_\gamma$ as $\gamma \to 0$ provides our final result. The complete details of the proof are provided in the article [5].*

Although seemingly intractable, the form taken by $\tilde{\mathcal{Q}}$ is quite interesting in itself. First, note that $\mathcal{R}$ and $\tilde{\mathcal{R}}$ only depend on W so that the terms $\mathcal{R}$ and $M^\mathsf{T}\tilde{\mathcal{R}}^{-1}M$ contain all the information about the W found in $\tilde{\mathcal{Q}}$. The noise level η^2 then trades the need for regularization of $\tilde{\mathcal{Q}}$ and the need for emphasizing the part $A^\mathsf{T}\tilde{\mathcal{R}}^{-1}A$ versus $I_T + \mathcal{R}$; the former matrix indeed has its columns living in the span of the columns of U with weights imposed by the network matrix $M^\mathsf{T}\tilde{\mathcal{R}}^{-1}M$. Thus, we expect the memory performance of the ESN to relate strongly to both η^2 and the effect of $\mathcal{R}$ and $\tilde{\mathcal{R}}$.

An interesting corollary is found when $c = 0$ (i.e., $n/T \to 0$) in which case one easily shows that $\mathcal{R} = 0$ and $\tilde{\mathcal{R}} = S_0$. And this thus brings the almost sure convergence

$$\left| E_\eta(u, r) - \frac{1}{T} r^\mathsf{T} \left(I_T + \frac{1}{\eta^2} U^\mathsf{T} D U \right)^{-1} r \right| \to 0$$

with D the matrix with entry $D_{ij} = m^\mathsf{T}(W^{i-1})^\mathsf{T} S_0^{-1} W^{j-1} m$ (recall that $S_0 = \sum_{k \geq 0} W^k (W^k)^\mathsf{T}$). We recognize here the diagonal entries of D to be the quantity

$$D_{ii} = m^\mathsf{T}(W^{i-1})^\mathsf{T} \left(\sum_{k \geq 0} W^k (W^k)^\mathsf{T} \right)^{-1} W^{i-1} m$$

known in the ESN literature as (the value i of) the *Fisher memory curve* [2, 6], an abstract mesure of the ability of the ESN to maintain an i-step old input in memory.

Having an insight on the couple $(\mathcal{R}, \tilde{\mathcal{R}})$ for non trivial values of c is however quite involved, and it is interesting to consider specific cases where this expression simplifies. In particular, one may choose W to be a given snapshot of a random matrix model. Since n, p are assumed large, by means of random matrix identities, this one realization will have an almost sure deterministic behavior in the limit, thereby leading to explicit approximations for $E_\eta(u, r)$.

We shall consider next two classical random matrix models for W modelling directed and undirected random graphs.

2.2. Non-Hermitian random W

The first model assumes a non-Hermitian structure for W but with statistical invariance when (left- or right-) multiplied by orthogonal matrices. In this case, $\mathcal{R}$ merely becomes $\frac{c}{1-c} I_T$ while $\tilde{\mathcal{R}}$ is essentially $(1-c)S_0$ and we have the following corollary of Theorem 1.

Corollary 1 (Non-Hermitian Random W) *Let the Assumptions of Theorem 1 hold and take W to be random with left and right orthogonal invariance, and m of unit norm independent of W. Then*

$$\left| E_\eta(u, r) - (1-c)\frac{1}{T} r^\mathsf{T} \left(I_T + \frac{1}{\eta^2} U^\mathsf{T} D U \right)^{-1} r \right| \to 0$$

almost surely, where D is diagonal with

$$D_{ii} \equiv \frac{1}{n} \operatorname{tr} W^{i-1} (W^{i-1})^\mathsf{T} S_0^{-1}.$$

In particular, if $W = \sigma Z$ with Z a random orthogonal and orthogonally invariant matrix,[1] *then $D_{ii} = (1-\sigma^2)\sigma^{2(i-1)}$.*

Corollary 1 provides much expected insights on the behavior of ESN's in the case of non-Hermitian random W. In particular, from the diagonal structure of D and the fact that the columns of U are delayed versions of the input sequence $\{u_t\}$, the performance of the ESN will depend on the accuracy of the representation of r as a linear combination of weighted delays of $\{u_t\}$. Note interestingly that these delays satisfy $\sum_{i \geq 1} D_{ii} = 1$, so that the ESN effectively distributes its memory capacity along the successive delays (in an exponentially decaying manner). The specific choice of the singular values of W impact these delays.

Figure 2 provides a performance comparison between Monte Carlo simulations and our theoretical results for the training of an interpolation task of chaotic data (here the Mackey–Glass model [7]). An accurate match is observed between theory and practice with increased precision as n, T grow large, consistently with our results. Despite the apparent accuracy on this example, it is nonetheless important to stress that Theorem 1 is only valid for fixed $\eta > 0$ and growing n, T. In particular, the approximation for small η's may dramatically fail in some specific cases that we simulated.

[1] This is often referred to as a *Haar* matrix.

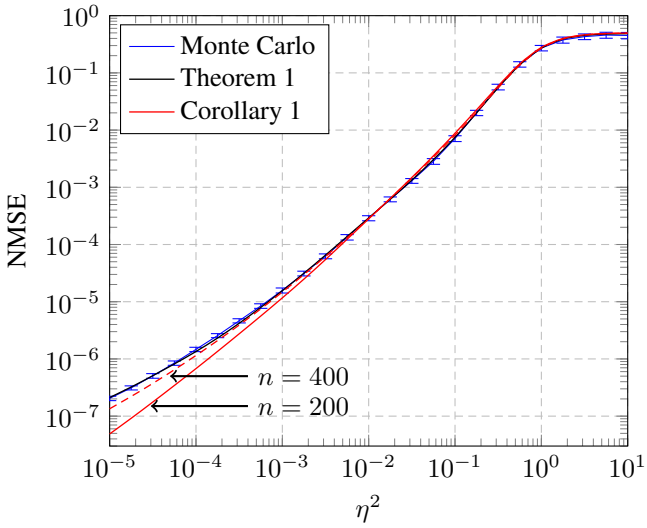

Fig. 2. (Normalized) MSE for the Mackey Glass one-step ahead task, $W = \sigma Z$ for Z Haar, $n = 200$, $T = 400$. Error bars indicate one standard deviation of noise realizations.

2.3. Hermitian random W

A second natural example is to assume that the ESN is now an *undirected graph*, so that W is symmetric. In this case, the matrix $\mathcal{R}$ is no longer diagonal but is found to be solution of a simple fixed-point equation. Precisely, assuming W Hermitian orthogonally invariant with normalized empirical eigenvalue distribution (i.e., the measure $n^{-1} \sum_i \boldsymbol{\delta}_{\lambda_i(W)}$ with $\lambda_i(W)$ the eigenvalues of W) converging to μ, we have

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

for all $a, b \in \{1, \ldots, T\}$. Since $\mathcal{R}$ is a Toeplitz matrix with exponentially decaying first column entries, this equation only involves a few parameters and is easily solved. An interesting case is when μ is symmetrical (i.e., $\mu(-t) = \mu(t)$) in which case all values of $\mathcal{R}_{ab}$ for $a - b$ odd vanish. Figure 3 provides a visual comparison of $\mathcal{R}$ when W is a Gaussian matrix with i.i.d. entries with or without symmetry.

As opposed to the case of non-Hermitian W, the ESN output is not merely mapped to a delayed version of past inputs but a more intricate combination of them. This intuitively induces performance losses when it comes to fulfilling pure delay task. This is confirmed in Figure 4, where ESN's with or without Hermitian symmetry are trained on a τ-delay memory task on the Mackey–Glass dataset. While the performance decay incurred by increasing τ for the non-symmetric case is moderate, it is instead dramatic for symmetric matrices. As such, undirected random graphs are seen to perform poorly on related simple memory tasks.

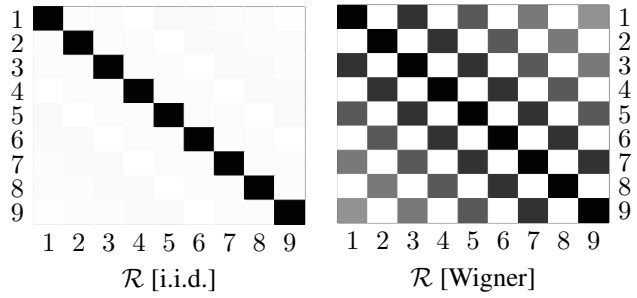

Fig. 3. Upper 9×9 part of $\mathcal{R}$ for $c = 1/2$ and $\sigma = 0.9$ for W with i.i.d. zero mean Gaussian entries [left] and W Gaussian symmetric (Wigner) [right].

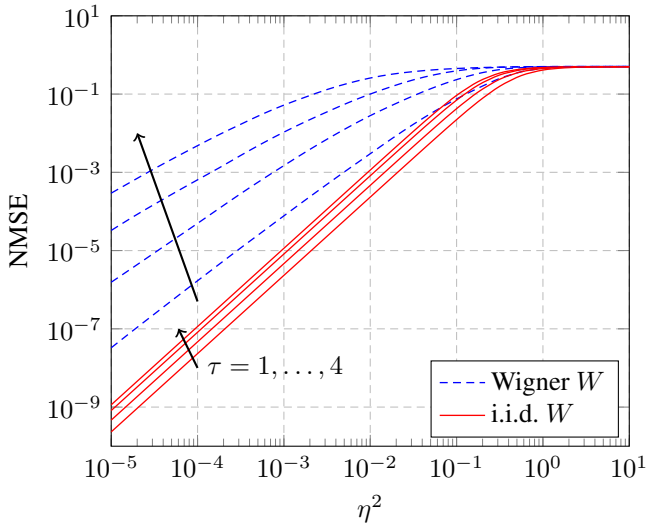

Fig. 4. Performance of a τ-delay task for $\tau \in \{1, \ldots, 4\}$ compared for i.i.d. W versus Wigner W, $\sigma = .9$ and $n = 200$, $T = 400$ (here on the Mackey-Glass dataset).

3. CONCLUDING REMARKS

Several interesting remarks can be drawn from the theoretical results in Section 2. Consider for instance the scenario where $W = \sigma Z$ for Z a Haar matrix and in which the sought-for output r is defined as $r = \sqrt{T} U^{\mathsf{T}} b$ for some vector $b \in \mathbb{R}^k$ and k small. Then, from Corollary 1,

$$\frac{E_\eta(u,r)}{1-c} \simeq b^{\mathsf{T}} U \left(I_T + \sum_{i \geq 0} \frac{(1-\sigma^2)\sigma^{2(i-1)}}{\eta^2} U_{\cdot,i}^{\mathsf{T}} U_{\cdot,i} \right)^{-1} U^{\mathsf{T}} b$$

which can be shown to converge to zero as $\eta \to 0$ and may, for every η, be numerically minimized over σ. In particular, letting $b_i = \alpha^{i-1}$ for some $\alpha \in (-1, 1)$, one can show that $\sigma^2 = |\alpha|$ minimizes $E_\eta(u, r)$. Thus, the network should (as one would expect) align as much as possible to the dependence of r_t in u_{t-i} for each i. For more general than exponential decaying relations between r_t and u_{t-i}, and in particular to account for heterogeneous memory dynamics, one may appropriately consider $W = \mathrm{diag}(\sigma_1 Z_1, \ldots, \sigma_L Z_L)$ for different values of σ_ℓ and (different sized) independent Haar Z_ℓ, parametrizable upon the application context.

For k not small however, or for more general models of r as a function of u (especially non linear models), the convergence of $E_\eta(u, r)$ to zero as $\eta \to 0$ is not guaranteed. In this regime of small η's, our analysis however theoretically breaks so that the instability shortcomings already observed by Jaeger in [2] cannot be resolved at this point and remain an open question. Nonetheless, it clearly appears through our results that large enough η's (it can be proved enough to have $\eta^2 \gg n^{-\frac{1}{2}}$) induce system stability.

Generalizations of the present work encompass the extension to the testing (as opposed to training) performance of ESN's (see [5] for preliminary results) along with the challenging consideration of *non-linear* activation functions, i.e., generalizing (1) to $x_{t+1} = S(Wx_t + mu_{t+1} + \eta\varepsilon_{t+1})$ for some (pointwise) sigmoid function S. The theoretical difficulties incurred by the latter may be circumvented by ideas from mean field dynamics.

4. REFERENCES

[1] H. Jaeger, "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*, vol. 148, pp. 34, 2001.

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

[3] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*, Cambridge University Press, New York, NY, USA, first edition, 2011.

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

[5] R. Couillet, Gilles Wainrib, Harry Sevi, and Hafiz Tiomoko Ali, "The asymptotic performance of linear echo state neural networks," *(submitted to) Journal on Machine Learning Research*, 2016.

[6] Surya Ganguli, Dongsung Huh, and Haim Sompolinsky, "Memory traces in dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 105, no. 48, pp. 18970–18975, 2008.

[7] Leon Glass and Michael C. Mackey, "A simple model for phase locking of biological oscillators," *Journal of Mathematical Biology*, vol. 7, no. 4, pp. 339–352, 1979.