

Orthogonality in Additive Reservoir Computing

Andrea Ceni and Claudio Gallicchio *

Department of Computer Science, University of Pisa
Largo Bruno Pontecorvo 3 - 56127 Pisa, Italy

Abstract. Reservoir computing (RC) is a state-of-the-art approach for efficient training in temporal domains. In this paper, we explore new RC architectures that generalise the popular leaky echo state network model (leaky-ESN) introducing an additive orthogonal term outside the nonlinear part of the ESN equation. We investigate the benefits of employing orthogonal matrices in ESNs both inside the nonlinearity and outside of it. We show empirically how to boost the memory capacity towards the theoretical maximum value while still preserving the power of nonlinear computations. Ergo, we optimise the compromise between computing with memory and computing with nonlinearity. The proposed model demonstrates to outperform both leaky-ESN and orthogonal reservoir ESN models on tasks requiring nonlinear computations with memory.

1 Introduction

Recurrent neural networks (RNNs) are a powerful deep learning tool for sequential data, successfully exploited in classification of time series, speech recognition, natural language, and many others. Training RNNs via stochastic gradient descent methods involves a great computational effort. Moreover, learning long-term dependencies with RNNs is especially difficult, due to a fundamental problem known in the literature as the vanishing-exploding gradient issue of RNNs [1]. Reservoir computing (RC) is an alternative paradigm for training RNNs, which elegantly circumvents the computational burden and the vanishing-exploding gradient issue at once. A large *reservoir* of recurrent nonlinear units is randomly initialised and left untrained. Input is fed into the reservoir, which in turn develops an “echo” of nonlinear activations from which we readout the past history of the input signal. Consequently, only an output layer needs to be trained, as long as the reservoir is large and heterogeneous enough to encode the information provided by the input. Echo state networks (ESNs) are a class of RC models that demonstrated over the years to be a powerful nonlinear computational model [2, 3, 4]. The efficiency of ESNs made them appealing to neuromorphic implementations [5], and low-power devices embedding [6], two crucial assets for building a truly pervasive AI.

In this paper, we push further the boundaries of ESNs, exploring new architectural variants that generalise well-established ESN models, such as the leaky-ESN [7], and orthogonal reservoir ESNs [8, 9]. In particular, we study reservoir dynamical equations where both the nonlinear term and the additive “leakage”

*This work has been partially supported by the project TEACHING, under the European Union’s Horizon 2020 Research and Innovation program (G.A. ID: 871385), and by the My-BreathingHeart project (Bando Ricerca COVID-19 Toscana CUP n. J64G20000380001).

term can include orthogonal state transformations. The isometric property of orthogonal matrices makes them extraordinarily useful for propagating information along deep architectures, and thus for learning long-term dependencies in RNNs. This intuition led to a flurry of works focusing on unitary learning, i.e. learning algorithms constrained into the manifold of unitary matrices (see, e.g., [10] and references therein). In RC, ESNs with orthogonal reservoirs have been shown to produce comparable performance to standard ESNs, while enhancing their memory capacity (MC) [8]. However, the possibility of exploiting orthogonal transformations outside the nonlinear term of the reservoir state update equation still remains unexplored and motivates our analysis in this paper. A related RC architecture is given by the leaky-ESN [7], where the nonlinear part is added on top of the internal activations of the previous time step. Instead of re-using the previous internal activations exactly as they are, the novel idea of this paper is to filter it through an untrained orthogonal matrix. In this sense, the leaky-ESN is a peculiar case where the orthogonal matrix is the identity matrix. This idea turns out as simple as it is effective. We show empirically that this approach leads to significantly outperform both leaky-ESN and orthogonal reservoir ESN models, enhancing the MC towards the theoretical maximum, and optimising the trade-off between long short-term memory and nonlinear computation.

2 Additive reservoirs based on orthogonal matrices

We start our analysis by recalling the equations of a leaky-ESN with linear readout, which reads as follows:

$$x[t] = \alpha \tanh(\rho \mathbf{W}_r x[t-1] + \omega \mathbf{W}_{in} u[t]) + (1 - \alpha)x[t-1], \quad (1)$$

$$z[t] = \mathbf{W}_o x[t]. \quad (2)$$

The internal state $x[t]$, input $u[t]$, and output $z[t]$ are, respectively, N_r -dimensional, N_i -dimensional and N_o -dimensional vectors of real values. Matrices \mathbf{W}_r , \mathbf{W}_{in} are randomly instantiated and left untouched. In this paper, we initialise \mathbf{W}_{in} with i.i.d. random uniformly distributed entries in $(-1, 1)$, and \mathbf{W}_r with i.i.d. normally distributed entries with zero mean and standard deviation $\frac{1}{\sqrt{N_r}}$. Rooted in the circular law from random matrix theory, this initialisation scheme for \mathbf{W}_r ensures, in the limit of an infinitely large reservoir, that the spectral radius of \mathbf{W}_r is 1. Therefore, the hyperparameter ρ can be thought as the spectral radius of the effective recurrent matrix. While, given a training set $\{u[t], y[t]\}_{t=1}^T$, the readout matrix \mathbf{W}_o is trained via ridge regression [4] by means of the following formula $\mathbf{W}_o = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \mu\mathbf{I})^{-1}$, where \mathbf{X} is a matrix of dimension $N_r \times T$, containing all the internal states $x[t]$ of the ESN driven by the input $u[t]$ for $k = 1, \dots, T$, \mathbf{Y} a matrix of dimension $N_o \times T$, containing all the target values $y[t]$, \mathbf{I} is the identity matrix of dimension $N_r \times N_r$, and μ is the regularisation parameter. In the reservoir state transition (1), the input scaling ω is an hyperparameter entitled to rescale the weight of the current input into the reservoir dynamics. The hyperparameter ρ is a positive real value controlling the amount

of nonlinearity into the reservoir and the contribution of the past activations. ESNs work under the fundamental assumption of the echo state property (ESP), a condition ensuring a unique stable input-driven response [2, 11]. In few words, the ESP guarantees that the internal state $x[t]$ is uniquely determined by the entire past history of the input signal. Often in the literature it is set $\rho < 1$, and in most cases this condition is correlated with the ESP. However, it worth to mention that such a choice is in general neither a sufficient nor a necessary condition for the ESP, but rather ρ should be optimised in synergy with the input scaling ω in order to ensure the ESP. The tuning between these two hyperparameters should place the model in a sweet spot between the edge of unstable dynamics and slightly stable contractive dynamics where the ESP holds. The hyperparameter $\alpha \in (0, 1]$ is typically exploited to tune the internal temporal characteristics of the network according to a given task [7].

In this paper we aim to explore the benefits of having a non-identical filter for the linear additive part of the leaky-ESN model. In particular, we propose the following modification of equation (1):

$$x[t] = \alpha \tanh(\rho_A \mathbf{W}_A x[t-1] + \omega \mathbf{W}_{in} u[t]) + (1 - \alpha) \rho_B \mathbf{W}_B x[t-1], \quad (3)$$

where \mathbf{W}_A and \mathbf{W}_B respectively modulate the recurrence in the nonlinear and in the additive terms. We consider the four combinations of $\mathbf{W}_A, \mathbf{W}_B$ to be either orthogonal or random with i.i.d. entries normally distributed with zero mean and standard deviation of $\frac{1}{\sqrt{N_r}}$, where N_r is the reservoir size. Random orthogonal matrices are obtained by first generating a random matrix W with i.i.d. uniformly distributed entries in $(-1, 1)$, and then applying a QR decomposition, $W = QR$. Therefore, the resulting Q matrix is a random orthogonal matrix. Note however that this procedure does not ensure to explore uniformly the manifold of orthogonal matrices [12]. In both cases of random or random orthogonal initialization, the values of ρ_A and ρ_B in (3) regulate the spectral radii of the \mathbf{W}_A -recurrent part and \mathbf{W}_B -recurrent part, respectively. Note that the model in (3) is a generalisation of the leaky-ESN model, since we can recover it with the choice of \mathbf{W}_B as the identity matrix with $\rho_B = 1$. Analogously, the model in (3) is also generalising orthogonal reservoir ESNs (ortho-ESN), which can be obtained via setting an orthogonal \mathbf{W}_A with $\alpha = 1$. We denote the four variants of the proposed model as: RandA-RandB, where both $\mathbf{W}_A, \mathbf{W}_B$ are random matrices; OrthoA-RandB, where \mathbf{W}_A is random orthogonal, and \mathbf{W}_B is random; RandA-OrthoB, where \mathbf{W}_A is random, and \mathbf{W}_B is random orthogonal; OrthoA-OrthoB, where both $\mathbf{W}_A, \mathbf{W}_B$ are random orthogonal matrices.

In the following, we compare the performances of these four models between each other, and against the leaky-ESN and the ortho-ESN models.

3 Experiments

Memory capacity. We consider a fully connected reservoir of $N_r = 100$ neurons. The input $u[t]$ is an i.i.d. signal uniform in $[-0.8, 0.8]$ of discrete-time length $t = 1, \dots, 6000$. We split 5000 time steps for training (excluding the very

first 100 to washout the initial transient), and 1000 for test. The MC is defined as $MC = \sum_{k=1}^{\infty} MC_k$, where MC_k is the squared correlation coefficient between the output $z_k[t]$ and the target $y_k[t] = u[t - k]$ computed along the test session. The MC sum is computed up to $k = 200$. Each one of the four models has been run for 10 different initialisations for each delay k , and the computed MC has been averaged over these trials. For the calculation, we set spectral radii ρ_A, ρ_B to 0.9 whenever $\mathbf{W}_A, \mathbf{W}_B$ are random (or kept to 1 whenever they are orthogonal), and an input scaling of 0.1. No regularisation has been applied, that is $\mu = 0$. This setting of hyperparameters has been tested as good for ESNs in previous works [13]. Keeping fixed those, the leak rate α has been varied in $\{A \cdot 10^{-s} : A = 1, \dots, 10, s = 1, 2, 3\}$. The results are plotted as blue dots in Figure 1. The shadow of these curves represents a range of plus and minus 3 times the empirical standard deviation computed on the 10 trials. From these simulations, two main insights arise. An orthogonal reservoir matrix \mathbf{W}_A makes the memory capacity increase for α values approaching to 1, reaching a maximum value of 89.8 ± 1.2 , regardless of the matrix \mathbf{W}_B . On the other hand, an orthogonal matrix \mathbf{W}_B makes the memory capacity peak for α values around 0.05, reaching a global maximum value of 98.25 ± 0.22 , regardless of the reservoir matrix \mathbf{W}_A . Notably, the theoretical maximum value for a reservoir of N_r neurons has been proved to be N_r [14], i.e. 100. This maximum is essentially reached by the models with an orthogonal matrix \mathbf{W}_B . Spikes in the standard deviations for small values of α in RandB variants are due to some randomly generated \mathbf{W}_B with a spectral radius slightly larger than one.

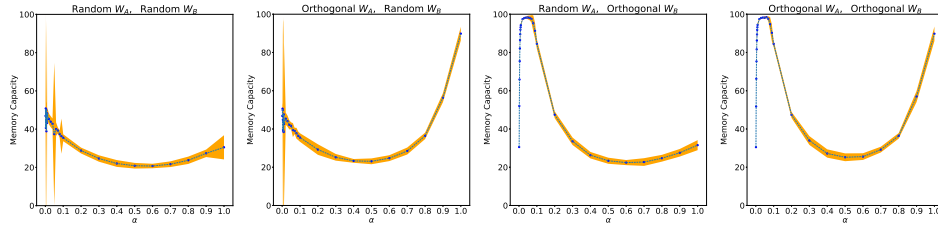


Fig. 1: Memory Capacity averaged over 10 trials for the proposed ESN variants.

Nonlinear computations with memory. Here we test the ability of the reservoir system to both retrieve the input signal from the past and at the same time make nonlinear computations based on it. The input signal $u[t]$ is i.i.d. random uniform in $[-0.8, 0.8]$, and the task is to output the target signal of the form $y[t] = \sin(\nu * u[t - \tau])$, where ν quantifies the *nonlinearity strength*, and τ measures the *memory depth* [15]. We consider 4 combinations of $(\tau, \log(\nu))$ in order to span all the major behaviours, i.e. the 4 combinations of values $\tau = 1$ (small delay), $\tau = 30$ (large delay), $\log(\nu) = -2$ (weak nonlinearity), and $\log(\nu) = 1$ (strong nonlinearity), where \log is the natural logarithm. We consider fully connected reservoirs of 100 units. We compare the four variants of the proposed architecture against the leaky-ESN and the ortho-ESN. We optimise the hyperparameters of all models via random search in the following ranges.

Input scaling $\omega \in (0.2, 6)$; spectral radii $\rho, \rho_A, \rho_B \in (0.1, 3)$; α generated via the formula $A \cdot 10^{-s}$, with A uniformly random in $(0.1, 1)$, and s uniformly random in $\{0, 1\}$, so that they vary in $(10^{-2}, 1)$. The only exception is the ortho-ESN model, where it is kept $\alpha = 1$. The input signal has length 7000. We use the first 5000 time steps for training (excluding the very first 100 to washout the initial transient), the steps from 5001 to 6000 for validation, and the remaining steps for testing. For all models, we run 1000 different trials to find the optimal hyperparameter setting. Finally, we run 10 different initialisations with the best hyperparameter setting found, train on the first 6000 time steps, and test the trained models on the remaining time steps from 6001 to 7000. In Table 1 we report the mean and standard deviation of these 10 computed NRMSEs on test.

Table 1: Mean and standard deviations of test NRMSE values over 10 trials.

Test NRMSE w $(\tau, \log(\nu))$	(1, -2)	(1, 1)	(30, -2)	(30, 1)
Leaky-ESN	$(3.45 \pm 0.79) \cdot 10^{-4}$	$(2.88 \pm 0.62) \cdot 10^{-3}$	$(9.52 \pm 0.11) \cdot 10^{-1}$	$(9.62 \pm 0.29) \cdot 10^{-1}$
Ortho-ESN	$(4.91 \pm 2.59) \cdot 10^{-5}$	$(1.01 \pm 0.13) \cdot 10^{-3}$	$(3.83 \pm 0.12) \cdot 10^{-1}$	$(4.90 \pm 0.09) \cdot 10^{-1}$
RandA-RandB	$(6.38 \pm 2.14) \cdot 10^{-5}$	$(3.38 \pm 2.10) \cdot 10^{-4}$	$(1.64 \pm 0.45) \cdot 10^{-1}$	$(5.03 \pm 0.32) \cdot 10^{-1}$
OrthoA-RandB	$(3.11 \pm 1.55) \cdot 10^{-5}$	$(2.37 \pm 1.16) \cdot 10^{-4}$	$(3.83 \pm 0.47) \cdot 10^{-1}$	$(4.75 \pm 0.40) \cdot 10^{-1}$
RandA-OrthoB	$(1.60 \pm 0.22) \cdot 10^{-5}$	$(1.83 \pm 0.21) \cdot 10^{-4}$	$(1.76 \pm 0.24) \cdot 10^{-2}$	$(2.11 \pm 0.12) \cdot 10^{-1}$
OrthoA-OrthoB	$(2.19 \pm 1.12) \cdot 10^{-5}$	$(2.65 \pm 2.20) \cdot 10^{-4}$	$(4.69 \pm 0.34) \cdot 10^{-2}$	$(2.19 \pm 0.14) \cdot 10^{-1}$

From Table 1 we see that in the case of small delay $\tau = 1$ all the proposed models outperform the benchmarks (leaky-ESN and ortho-ESN), apart from the case of $(\tau, \log(\nu)) = (1, -2)$ where RandA-RandB gives comparable performance to ortho-ESN. For large delay $\tau = 30$, the variants with a random \mathbf{W}_B are prone to instabilities, so we set the additional condition $\rho(\mathbf{W}_B) \leq 1$, to avoid huge values of the mean and standard deviation of NRMSE due to occasional exploding dynamics. On the contrary, the variants with an orthogonal \mathbf{W}_B exhibit reliable results. In general, the variants with orthogonal \mathbf{W}_B provide the best performances. Particularly evident is the case of $(\tau, \log(\nu)) = (30, -2)$, where the improvement is of almost two orders of magnitude w.r.t. the leaky-ESN, and one order w.r.t. the ortho-ESN. Moreover, even in the challenging case of strong nonlinearity $\log(\nu) = 1$, and large delay $\tau = 30$, the resulting performance is more than doubled w.r.t. ortho-ESN, and quadrupled w.r.t. leaky-ESN.

4 Conclusions

In this paper, we have introduced a generalisation of the leaky-ESN model. We explored the role of orthogonality outside of the nonlinearity in the additive term of the leaky-ESN's equation. Experimental results revealed the striking advantages obtainable from this simple variant of leaky-ESN model by (i) essentially reaching the theoretical maximum value of memory capacity, and (ii) accomplishing nonlinear computations without degrading the memory. The proposed model is tested on a task that require nonlinear computations over the past

history of the input. The results show that this new architecture substantially outperforms both the leaky-ESN and the orthogonal reservoir ESN models.

This work suggests several future directions of research that we intend to explore in depth, among which to (i) investigate the proposed architecture with structured orthogonal matrices (e.g. permutations) in order to provide more efficient reservoir computers exploitable for neuromorphic, IoT low-power devices and health-monitoring applications, (ii) provide an exhaustive theoretical analysis of the merits of orthogonality in additive state transition functions, and finally (iii) consider the proposed model outside reservoir computing in the context of fully trainable recurrent neural networks.

References

- [1] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. PMLR, 2013.
- [2] Herbert 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*, 148(34):13, 2001.
- [3] Herbert Jaeger and Harald Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *science*, 2004.
- [4] Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.
- [5] Gouhei Tanaka, Toshiyuki Yamane, Jean Benoit Héroux, Ryosho Nakane, Naoki Kanazawa, Seiji Takeda, Hidetoshi Numata, Daiju Nakano, and Akira Hirose. Recent advances in physical reservoir computing: A review. *Neural Networks*, 115:100–123, 2019.
- [6] Davide Bacciu, Paolo Barsocchi, Stefano Chessa, Claudio Gallicchio, and Alessio Micheli. An experimental characterization of reservoir computing in ambient assisted living applications. *Neural Computing and Applications*, 24(6):1451–1464, 2014.
- [7] Herbert Jaeger, Mantas Lukoševičius, Dan Popovici, and Udo Siewert. Optimization and applications of echo state networks with leaky-integrator neurons. *Neural networks*, 20(3):335–352, 2007.
- [8] Ali Rodan and Peter Tino. Minimum complexity echo state network. *IEEE transactions on neural networks*, 22(1):131–144, 2010.
- [9] Olivia L White, Daniel D Lee, and Haim Sompolinsky. Short-term memory in orthogonal neural networks. *Physical review letters*, 92(14):148102, 2004.
- [10] Mario Lezcano-Casado and David Martinez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In *International Conference on Machine Learning*, pages 3794–3803. PMLR, 2019.
- [11] Andrea Ceni, Peter Ashwin, Lorenzo Livi, and Claire Postlethwaite. The echo index and multistability in input-driven recurrent neural networks. *Physica D: Nonlinear Phenomena*, 412:132609, 2020.
- [12] Francesco Mezzadri. How to generate random matrices from the classical compact groups. *arXiv preprint math-ph/0609050*, 2006.
- [13] Claudio Gallicchio, Alessio Micheli, and Luca Pedrelli. Deep reservoir computing: A critical experimental analysis. *Neurocomputing*, 268:87–99, 2017.
- [14] Herbert Jaeger. Short term memory in echo state networks. gmd-report 152. In *GMD-German National Research Institute for Computer Science (2002)*, <http://www.faculty.jacobs-university.de/hjaeger/pubs/STMEchoStatesTechRep.pdf>. Citeseer, 2002.
- [15] Masanobu Inubushi and Kazuyuki Yoshimura. Reservoir computing beyond memory-nonlinearity trade-off. *Scientific reports*, 7(1):1–10, 2017.