Critical echo state network dynamics by means of Fisher information maximization

Filippo Maria Bianchi*, Lorenzo Livi †, Robert Jenssen *, Cesare Alippi ‡,
* Machine Learning Group, University of Tromsø, Norway, Email: filippo.m.bianchi@uit.no, robert.jenssen@uit.no
† Department of Computer Science, College of Engineering, Mathematics and Physical Sciences,
University of Exeter, UK, Email: l.livi@exeter.ac.uk
‡ Department of Electronics, Information, and Bioengineering,
Politecnico di Milano, Italy, Email: cesare.alippi@polimi.it

***Abstract*—The computational capability of an Echo State Network (ESN), expressed in terms of low prediction error and high short-term memory capacity, is maximized on the so-called "edge of criticality". In this paper we present a novel, unsupervised approach to identify this edge and, accordingly, we determine hyperparameters configuration that maximize network performance. The proposed method is independent of the task the network is required to solve and stems from recent theoretical results consolidating the link between Fisher information and critical phase transitions. We show how to identify optimal ESN hyperparameters by relying only on the Fisher information matrix (FIM) estimated from the activations of hidden neurons. In order to take into account the particular input signal driving the network dynamics, we adopt a recently proposed non-parametric FIM estimator. Experimental results on a set of standard benchmarks are provided and discussed, demonstrating the validity of the proposed method.**

I. Introduction

In the last years, ESNs have emerged as a powerful class of recurrent neural networks (RNNs), achieving outstanding result in prediction of real-valued time series [1], [2], [3]. Although ESNs are typically randomly initialized, the network designer has access to a set of hyperparameters to control the network behavior. For instance, the spectral radius of the reservoir weight matrix directly affects ESN dynamics and, therefore, its computational capability.

However, such hyperparameters are difficult to set and with limited portability in different applications; parameters tuning is usually accomplished through a long trial-and-error approach [4], [5], [6], [7], [8], relying on blackbox cross-validation techniques [9]. This limits their use beyond field experts with domain knowledge, up to the point of hampering the potential benefit of such methods. Furthermore, cross-validation is a supervised method that requires to evaluate the performance on a validation set. This might be an issue in real-life applications, where data are scarce and supervised information not always available. Moreover, the model has to be re-evaluated for each hyperparameter configuration, leading to long training time if learning procedures are complex.

As a novel solution, in this paper we propose an unsupervised method, which exploits the Fisher information matrix (FIM) properties of a system undergoing a continuous phase transition, for identifying an optimal ESN hyperparameter configuration. Our approach is founded on a theoretical result, which demonstrates that Fisher information is maximized for systems operating at criticality [10]. We assume that the ESN dynamics can be characterized by a continuous phase transition and that its operating state is controlled by the considered hyperparameter configurations. Hence, we define the edge of criticality of an ESN as the collection of hyperparameters that leads the system to a state where Fisher information is maximized. Given an input signal, the proposed method identifies a configuration of the network where it achieves high computational capability, disregarding of the specific task the network is required to solve. The proposed criterion is theoretically motivated and further highlights the existence of a possible interplay between the field of reservoir computing, complex systems, and critical phenomena.

II. Phase transitions and the edge of criticality

ESNs, as well as other classes of RNNs, generate complex dynamics characterized by sharp transitions between ordered and chaotic regimes. Highest information processing capabilities, in terms of memory capacity (storage of past events) and performance on the modeling/prediction task at hand, are usually achieved on the edge of this transition [11]. This general behavior is in agreement with the widely-discussed, yet still controversial, "criticality hypothesis" associated with the functioning of many biological (complex) systems [12], [13]. In fact, these systems tend to self-organize and operate in a critical regime, being highly responsive to external stimuli and hence capable of generating any dynamics requested by the specific task [12].

Determination of system configurations lying on such edge of criticality is an important research endeavor [14], which is addressed, for instance, through appropriate sensitivity analyses. In this sense, Fisher information (or FIM, in the multivariate case) [15] provides a well-established framework. Fisher information is tightly linked with statistical mechanics and more specifically with the field of (continuous) phase transitions, which describe transformations affecting the qualitative behavior of a system. FIM components can be directly linked with the rate of change of order parameters, which are used to distinguish different phases of a controlled (thermodynamic) system [10]. The mathematical relationship between Fisher information and order parameters is useful to develop a statistical description of continuous, second-order phase transitions and, consequently, of any complex system approaching and/or operating at criticalility. In the case of continuous phase transitions, the first-order derivatives of the order parameters are discontinuous and divergent in at least one dimension. This implies that Fisher information diverges at criticality for infinite-dimensional systems, while it is maximized in the finite-size system case [10]. This fact provides a quantitative, well-justified tool to detect the onset of criticality.

Our objective is to provide a first principle method based on the notion of critical phase transition. Such a method can be used to determine in an unsupervised way the configurations that bring ESNs on the edge of criticality. This concept is illustrated in Fig. 1. The control parameters influencing the system behavior are, in our case, identified with ESN hyperparameters. By providing a connection between statistical mechanics and ESNs, we demonstrate that the same approach adopted to identify continuous phase transitions in control parameter space can be used to detect the onset of criticality in ESNs, where the computation capabilities are maximized for a large set of practical tasks.

To the best of our knowledge, approaches based on FIM are missing in the ESN literature. As such, the proposed method constitutes a novel contribution in the field. Furthermore, we believe that the interplay between concepts typically used in complex systems and RNNs would provide several new insights, which could lead to theoretical advances and disclose new applications in both research fields.

III. Echo state networks

An ESN is characterized by a *reservoir*, a large recurrent layer of non-linear units with randomly generated weights, which acts as a kernel mapping inputs to a high-dimensional space [16]. A linear, memory-less *readout*, is then trained with a regularized least-square optimization to solve a specific task. The state-update and the output of an ESN are, respectively, ruled by

$$\begin{aligned} \mathbf{h}[k] =& \psi(\mathbf{W}_r^r\mathbf{h}[k-1] + \mathbf{W}_i^r\mathbf{x}[k] + \mathbf{W}_o^r\mathbf{y}[k-1]), \\ \mathbf{y}[k] =& \mathbf{W}_i^o\mathbf{x}[k] + \mathbf{W}_r^o\mathbf{h}[k]. \end{aligned} \tag{1}$$

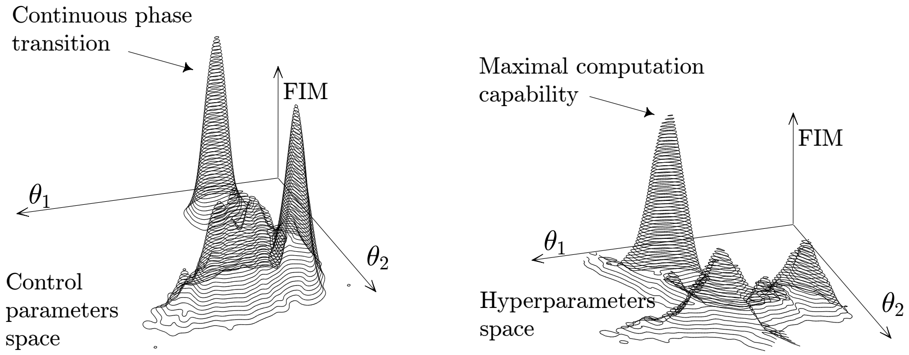
The reservoir contains N_r neurons, whose activation function $\psi(\cdot)$ is typically implemented as a hyperbolic tangent. At time instant k, the network is driven by the input signal $\mathbf{x}[k] \in \mathbb{R}^{N_i}$, it produces output $\mathbf{y}[k] \in \mathbb{R}^{N_o}$ and its state is represented by $\mathbf{h}[k] \in \mathbb{R}^{N_r}$. The weight matrices $\mathbf{W}_r^r \in \mathbb{R}^{N_r \times N_r}$ (reservoir connections), $\mathbf{W}_i^r \in \mathbb{R}^{N_i \times N_r}$ (input-to-reservoir connections), and $\mathbf{W}_o^r \in \mathbb{R}^{N_o \times N_r}$ (output-to-reservoir feedback) are usually initialized with random values drawn from a uniform distribution in $[-1, 1]$. $\mathbf{W}_i^o$ and $\mathbf{W}_r^o$, instead, are optimized for the task at hand. A visual representation of the ESN architecture is reported in Fig. 2

The behavior of the network can be controlled by tuning a set of scalar hyperparameters. Usually, one considers θ_{IS}, the scaling of the input weights $\mathbf{W}_i^r$, affecting the non-linearity introduced by the neurons; θ_{SR}, the spectral radius of $\mathbf{W}_r^r$, which influences both stability and computational capability of the network by shifting the transfer function poles [17]; θ_{RC}, which determines the sparsity of connectivity in $\mathbf{W}_r^r$, i.e., the number of weights set to 0; θ_{FB}, which affects $\mathbf{W}_r^o$, that is, the importance of output feedback connections. In this study, we set $\theta_{FB} = 0$ with a consequent simplification of ESN state-update (1).

Asymptotic stability is guaranteed by the so-called echo state property, which requires the reservoir to exhibit a fading memory of past inputs [18]. In practice, the degree of stability is often assessed by analyzing the Maximal Local Lyapunov Exponent (MLLE), computed from the Jacobian of the reservoir, which is easily derivable from Eq. 1. The MLLE approximates the separation rate in phase space of trajectories having very similar initial states [2]. In autonomous systems, MLLE ¡ 0 indicates stability, while MLLE ¿ 0 is characteristic of chaotic systems. The transition point, MLLE = 0, provides thus a criterion for detecting the onset of criticality in dynamic systems. Another indicator used to predict network performance is the minimal singular value of the Jacobian (shortened as mSVJ), which provides accurate information regarding the ESN dynamics. The collection of hyperparameter configurations that maximize mSVJ generates a dynamical system that is far from singularity, it has many degrees of freedom, a good excitability, and separates well the input signals in phase space [19]. These indicators have been used in the literature to define unsupervised methods for tuning hyperparameters and we consider them as a comparative baseline in the experimental section to evaluate the proposed criterion based on FIM.

IV. Identification of the critical ESN configurations

In the following, we describe the proposed method based on the determinant of the FIM for identifying a configuration of ESN hyperparameters lying on the critical region. The details of the FIM estimation procedure are provided in Sec. V. We take into account three important hyperparameters that affect an ESN dynamics, namely $\boldsymbol{\theta} = [\theta_{IS}, \theta_{SR}, \theta_{RC}]^T \in \Theta \subset \mathbb{R}^3$. It is worth underlying that the continuous parameter space Θ is actually quantized according to some user-defined resolution (although this is not a necessary assumption for the proposed methodology). This choice allows to disentangle the problems of defining and finding the edge of criticality. In fact, in order to



(a) Thermodynamic systems

(b) Echo State Networks

Fig. 1. The approach based on FIM maximization used to identify a continuous phase transition can be adopted also to characterize the dynamics in a ESN. In this context, the ESN hyperparameters (e.g., spectral radius and input scaling) play the same role of the control parameters in a thermodynamic system (e.g., temperature affects the magnetization). Accordingly, in ESN hyperparameter space FIM is maximized where the computation capability is highest. Note that the densities plotted in the two figures are not related; we report them as an example to show the role played by FIM in the two different contexts.

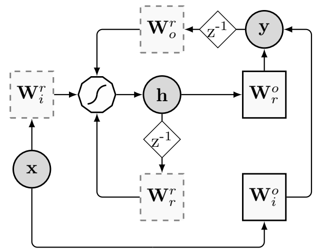


Fig. 2. Schematic depiction of an ESN. Circles represent input ($\mathbf{x}$), network state ($\mathbf{h}$) and output ($\mathbf{y}$), respectively. Solid squares $\mathbf{W}_r^o$ and $\mathbf{W}_i^o$, are the trainable matrices of the readout, while dashed squares, $\mathbf{W}_r^r$, $\mathbf{W}_o^r$, and $\mathbf{W}_i^r$, are randomly initialized matrices. The polygon represents the non-linear transformation performed by neurons and z^{-1} is the unit delay operator.

find the edge of criticality and thus objectively validate the proposed method, in this paper the focus is on the definition of the edge and we just implement a straightforward grid search on the hyperparameter space. A schematic description of the main stages of the procedure is shown in Fig. 3.

Given an input time-series $\boldsymbol{x}[1], \cdots, \boldsymbol{x}[K]$ and an initial configuration of the hyperparameters $\boldsymbol{\theta}_0$, the FIM is estimated from the series of reservoir neuron activations $\mathcal{S}_{\boldsymbol{\theta}_i} = \{\mathbf{h}[k]\}_{k=1}^{K}$. The edge of criticality, denoted as $\mathcal{K} \subset \Theta$, is then determined by relying on the determinant of $\hat{\mathbf{F}}$, the estimated FIM. FIM defines a metric tensor for the smooth manifold of parametric PDFs embedded in Θ and can be proved [20] that $\mathcal{K}$ corresponds to a region of Θ characterized by the highest concentration of distinguishable parametric PDFs .

Since the determinant $\det(\hat{\mathbf{F}}(\boldsymbol{\theta}))$ is monotonically related to such volume element and since FIM is a positive definite matrix, $\mathcal{K}$ can be defined as the set of hyperparameters $\boldsymbol{\theta}^*$ for which:

$$\mathcal{K} = \left\{ \boldsymbol{\theta}^* | \boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \Theta} \det(\hat{\mathbf{F}}(\boldsymbol{\theta})) \right\}. \tag{2}$$

The pseudo-code describing the proposed procedure is reported in Algorithm 1. The effect of the variation of the hyperparameters $\boldsymbol{\theta}$ on the resulting ESN state cannot be expressed analytically without making further assumptions, as the reservoir topology or the (unknown) input signal affects the ESN dynamics. Therefore, we rely here on the non-parametric FIM estimator (Sec. V) to calculate $\hat{\mathbf{F}}(\boldsymbol{\theta})$. Given a hyperparameter configuration $\boldsymbol{\theta}_i$, $\hat{\mathbf{F}}(\boldsymbol{\theta}_i)$ is estimated by analyzing the sequence $\mathcal{S}_{\boldsymbol{\theta}_i} = (\mathbf{h}[1], ..., \mathbf{h}[K])$ of reservoir neuron activations generated as the input $(\mathbf{x}[1], \cdots, \mathbf{x}[K])$ is processed. Additional sequences of activations $\mathcal{S}_{\bar{\boldsymbol{\theta}}_i}$ are generated by perturbing M times the current network configuration $\boldsymbol{\theta}_i$ with a small noise drawn from $\mathcal{N}(0, \sigma \mathbf{I}_{d \times d})$ (see line 7). The PDF associated to the internal states of the ESN, necessary to compute the FIM, arises from such stochastic perturbations of state sequence. In fact, since there is no stochasticity in the ESN state update, it would be impossible to evaluate a distribution of the states when the network is driven by a deterministic signal. Note that σ is an important parameter, which controls the magnitude of the perturbation.

To obtain a more robust estimate of the FIM, we perform a number of independent trials by repeating the estimation procedure T different times (see line 3). At each repetition, a new ESN is randomly initialized. At the end, the determinant is calculated on the resulting average FIM (see line 16).

V. Fisher information matrix and the non-parametric estimation

Here we provide details on FIM and the approach adopted for its estimation. To compute FIM, the analytical form of the underlying PDF generating the data is required. However, in many experimental settings this is often unknown, as well as the relation between the control parameters $\boldsymbol{\theta}$ and the resulting $p_{\boldsymbol{\theta}}(\cdot)$. Therefore, estimators are usually adopted. These, however, struggle if the domain of the unknown PDF is high-dimensional, such as the

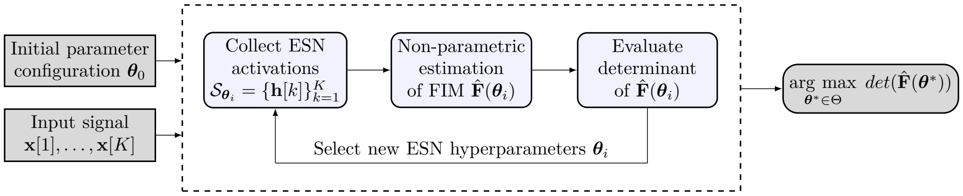

Initial parameter configuration $\boldsymbol{\theta}_0$
Input signal $\mathbf{x}[1], \ldots, \mathbf{x}[K]$
Collect ESN activations $\mathcal{S}_{\boldsymbol{\theta}_i} = \{\mathbf{h}[k]\}_{k=1}^{K}$
Non-parametric estimation of FIM $\hat{\mathbf{F}}(\boldsymbol{\theta}_i)$
Evaluate determinant of $\hat{\mathbf{F}}(\boldsymbol{\theta}_i)$
$\arg\max_{\boldsymbol{\theta}^* \in \Theta} det(\hat{\mathbf{F}}(\boldsymbol{\theta}^*))$
Select new ESN hyperparameters $\boldsymbol{\theta}_i$

Fig. 3. Schematic, high-level description of the proposed procedure.

Algorithm 1 Procedure for determining an ESN configuration on the edge of criticality.

```
Input: ESN architecture, input $\mathbf{x}$ of $K$ samples, quantized parameter space $\Theta$, standard deviation $\sigma$ for the perturbations, number of trials $T$ and perturbations $M$.
Output: A configuration $\boldsymbol{\theta}^* \in \mathcal{K}$
Select an initial parameter configuration, $\boldsymbol{\theta}_0 \in \Theta$; maximum $\nu = 0$
loop
    for $t = 1$ to $T$ do
        Randomly initialize the ESN weight matrices
        Configure ESN with $\boldsymbol{\theta}_i$ and process input $\mathbf{x}$
        Collect the related activations $\mathcal{S}_{\boldsymbol{\theta}_i} = \{\mathbf{h}[k]\}_{k=1}^{K}$
        for $j = 1$ to $M$ do
            Generate a perturbation vector $\mathbf{r}_j \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{d \times d})$
            Randomly initialize the ESN weight matrices
            Configure ESN with perturbed version $\bar{\boldsymbol{\theta}}_i^{(j)} = \boldsymbol{\theta}_i + \mathbf{r}_j$ and process input $\mathbf{x}$
            Collect the related activations $\mathcal{S}_{\bar{\boldsymbol{\theta}}_i^{(j)}} = \{\mathbf{h}[k]\}_{k=1}^{K}$
        end for
        Define $\mathcal{S}_{\bar{\boldsymbol{\theta}}_i} = \bigcup_{j=1}^{M} \mathcal{S}_{\bar{\boldsymbol{\theta}}_i^{(j)}}$
        Estimate the FIM $\hat{\mathbf{F}}^{(t)}(\boldsymbol{\theta}_i)$ of trial $t$ using $\mathcal{S}_{\boldsymbol{\theta}_i}$ and $\mathcal{S}_{\bar{\boldsymbol{\theta}}_i}$ with the non-parametric estimator introduced in Sec. V
    end for
    Compute the average FIM, $\hat{\mathbf{F}}(\boldsymbol{\theta}_i)$, using all $\hat{\mathbf{F}}^{(t)}(\boldsymbol{\theta}_i), t = 1, ..., T$
    if $\det(\hat{\mathbf{F}}(\boldsymbol{\theta}_i)) > \nu$ then
        Update $\nu = \det(\hat{\mathbf{F}}(\boldsymbol{\theta}_i))$ and $\boldsymbol{\theta}^* = \boldsymbol{\theta}_i$
    end if
    if Stop criterion is met (e.g. maximum number of iterations) then
        return $\boldsymbol{\theta}^*$
    else
        Select a new $\boldsymbol{\theta}_i \in \Theta$ based on a suitable search scheme
    end if
end loop
```

sequences of ESN states $\mathcal{S}_{\boldsymbol{\theta}_i}$ taken into account in our case. To address this issue, which has never be treated in the ESN literature, we evaluate FIM with a non-parametric estimator recently proposed in [21], which operates directly by relying on data/observations. We choose this particular estimator since, being based on a graph representation of the data (minimum spanning tree), it is suitable for dealing with high-dimensional distributions.

FIM is a symmetric positive semi-definite (PD) matrix, whose elements are

$$F_{ij}(p_{\boldsymbol{\theta}}(\cdot)) = \int p_{\boldsymbol{\theta}}(\mathbf{u}) \frac{\partial \ln p_{\boldsymbol{\theta}}(\mathbf{u})}{\partial \theta_i} \cdot \frac{\partial \ln p_{\boldsymbol{\theta}}(\mathbf{u})}{\partial \theta_j} d\mathbf{u}, \tag{3}$$

where $p_{\boldsymbol{\theta}}(\cdot)$ is a parametric probability density function (PDF), which depends on d parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_d]^T \in \Theta \subset \mathbb{R}^d$; Θ is the hyperparameter space. In (3), $\ln p_{\boldsymbol{\theta}}(\cdot)$ is the log-likelihood function. To ease notation, we denote $\mathbf{F}(p_{\boldsymbol{\theta}}(\cdot))$ as $\mathbf{F}(\boldsymbol{\theta})$. The $d(d+1)/2$ distinct entries in FIM encode the sensitivity of the PDF w.r.t. parameters $\boldsymbol{\theta}$.

The estimator adopted ([21]) is based on novel f-divergence measure,

$$\begin{aligned} D_\alpha(p, q) = {} & \frac{1}{4\alpha(1-\alpha)} \cdot \\ & \cdot \int_{\mathcal{D}} \frac{(\alpha p(\mathbf{u})(1-\alpha) q(\mathbf{u}))^2}{\alpha p(\mathbf{u})(1-\alpha) q(\mathbf{u})} d\mathbf{u} - (2\alpha - 1)^2, \end{aligned} \tag{4}$$

where $\alpha \in (0, 1)$ and $p(\cdot)$, $q(\cdot)$ are PDFs both supported on $\mathcal{D}$. Eq. 4 can be computed without estimating the PDFs by means of the Friedman-Rafsky test. The test uses two datasets, $\mathcal{S}_p$ and $\mathcal{S}_q$, containing samples extracted from $p(\cdot)$ and $q(\cdot)$, respectively. As the number of samples $n = |\mathcal{S}_p|$ and $m = |\mathcal{S}_q|$ grows, we have

$$1 - \mathcal{C}(\mathcal{S}_p, \mathcal{S}_q) \frac{n+m}{2nm} \xrightarrow{a.s.} D_\alpha(p, q), \tag{5}$$

being $\mathcal{C}(\mathcal{S}_p, \mathcal{S}_q)$ the outcome of Friedman-Rafsky test.

The FIM can be estimated with a proper f-divergence measure, calculated between the parametric PDF of interest and a perturbed version of it. By expanding Eq. 4 up to the second order one obtains:

$$D_\alpha(p_{\boldsymbol{\theta}}, p_{\hat{\boldsymbol{\theta}}}) \simeq \frac{1}{2} \mathbf{r}^T \mathbf{F}(\boldsymbol{\theta}) \mathbf{r}, \tag{6}$$

where $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \mathbf{r}$, being $\mathbf{r} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{d \times d})$ a small, normally distributed perturbation vector. In the following, we omit $\boldsymbol{\theta}$ and we refer to the estimated FIM as $\hat{\mathbf{F}}$. By considering Eq. 6, FIM can be approximated using the least-square method:

$$\hat{\mathbf{F}}_{\text{hvec}} = (\mathbf{R}^T \mathbf{R})^{-1} \mathbf{R}^T \mathbf{v}_{\boldsymbol{\theta}},$$

where $\mathbf{v}_{\boldsymbol{\theta}} = [v_{\boldsymbol{\theta}}(\mathbf{r}_1), ..., v_{\boldsymbol{\theta}}(\mathbf{r}_M)]^T$, with $v_{\boldsymbol{\theta}}(\mathbf{r}_i) = 2D_\alpha(p_{\boldsymbol{\theta}}, p_{\hat{\boldsymbol{\theta}}_i}), i = 1, ..., M$, and $D_\alpha(\cdot, \cdot)$ is computed according to Eq. 5. $\mathbf{R}$ is a matrix whose columns are the M perturbations vectors $\mathbf{r}_i$ and $\hat{\mathbf{F}}_{\text{hvec}}$ is the *half-vector* representation of $\hat{\mathbf{F}}$, which is defined as $\left[\hat{f}_{11}, \ldots, \hat{f}_{dd}, \hat{f}_{12}, \ldots, \hat{f}_{d(d-1)}\right]^T$.

VI. Experiments

To support our methodological developments, in this section we discuss the results of experiments performed on different tasks. In particular, we evaluate whether the FIM determinant is maximized in the same regions of the hyperparameter space Θ where ESN achieves the highest performance. We refer with ϕ as the critical region $\mathcal{K}$ in Θ where the FIM determinant is maximized. We compare our results with the MLLE criterion, which identifies $\mathcal{K}$ as the region where MLLE crosses zero; we denote such region with λ. Similarly, we adopt also the criterion that defines $\mathcal{K}$ as a region in Θ where mSVJ is maximum; we denote such region as η.

The hyperparameters are selected in a discretized space through a grid search, which considers 10 different values for each parameter. Specifically, we search θ_{SR} in $\{0.1, \ldots, 1.6\}$, θ_{IS} in $\{0.15, \ldots, 0.9\}$, and θ_{RC} in $\{0.1, \ldots, 0.7\}$, evaluating a total of 1000 configurations. As we considered a hyperparameter space Θ with 3 dimensions, the related edge of criticality is a 2-dimensional manifold embedded in Θ. For each hyperparameter configuration, in Algorithm 1 we generate $M = 80$ perturbations and we perform $T = 10$ trials to compute the ensemble average of the FIM. In each trial, we sample new (and independent) input and reservoir connection weights (W_i^r and W_r^r). The readout layer is trained by using a ridge least-square regression, with a regularization parameter set to 0.05. In every test we use a reservoir with $N_r = 75$ neurons; a standard washout procedure is adopted [22], which discards the first 100 states in order to get rid of the ESN transient.

We perform 4 different experiments, described in the following. In Fig. 4, we report the critical regions in Θ identified in each test by the three indicators based on maximization of FIM determinant, zero-crossing of MLLE and maximization of mSVJ; the light gray manifold corresponds to the regions in Θ where the performance of the network is maximized and the dark gray manifolds represent ϕ, λ and η respectively. In Tab. I, we report the numerical values of the correlation between the light gray manifolds and the dark gray ones.

A. Memory capacity

This test quantifies the capability of ESN to remember past sequences of an i.i.d. input. Given a time delay $\delta > 0$, the ESN is trained to reproduce the input $\mathbf{x}[k-\delta]$, after having seen the input up to time k. Memory Capacity (MC) is measured as the squared correlation coefficient between the desired output, which is the input signal delayed by different δ time steps, and the observed network output $\mathbf{y}[k]$:

$$\mathrm{MC} = \sum_{\delta=1}^{\delta_{\max}} \frac{\mathrm{cov}^2\left(\mathbf{x}[k-\delta], \mathbf{y}[k]\right)}{\mathrm{var}\left(\mathbf{x}[k-\delta]\right)\mathrm{var}\left(\mathbf{y}[k]\right)}. \tag{7}$$

MC is computed by training several readouts, one for each delay $\delta \in \{1, 10, \ldots, 100\}$, while keeping fixed input and reservoir layers.

As we can see from the 3 graphics in Figs. 4(a), the critical regions identified by each unsupervised method follow with a good accuracy the region in Θ where MC is maximized. The degrees of correlation for the MC task are described in Tab. I. Surprisingly, λ shows a very high correlation (81%) preforming better than η (65%) for this task. The correlation between ϕ and the region with maximum MC is also very high (75%), showing that both ϕ and λ can be used as reliable indicators to identify the optimal configurations that enhance the short-term memory capacity of the ESN.

B. Prediction accuracy

In this test, we evaluate if ϕ, λ and η are consistent with the accuracy on the prediction task. We define the prediction accuracy as $\gamma = \max\{1 - \mathrm{NRMSE}, 0\}$, were NRMSE is the Normalized Root Mean Squared Error achieved by the ESN. The prediction accuracy is evaluated on three prediction tasks of increasing complexity. For each of them, we set the forecast step $\tau_f > 0$ equal to the smallest time delay that guarantees input measurements to be decorrelated, which corresponds to the first zero of the autocorrelation function of the time-series.

In the first test, the ESN is trained to predict a *sinusoidal input* (SIN) using a forecast step equal to 1/4 of its period. As we can see from the graphics in Fig. 4(b), each measure is consistent with γ, the region where prediction performance are maximized. From Tab. I we can observe that ϕ achieves the best results (58 % correlation), but also the remaining measures have positive and similar degrees of correlation with γ ($\mathrm{corr}(\lambda, \gamma) = 52\%$ and $\mathrm{corr}(\eta, \gamma) = 56\%$, respectively).

The input signal in the successive test is generated by the *Mackey-Glass system*, described by the following differential equation:

$$\frac{dx}{dk} = \frac{\alpha x(k - \tau_{\mathrm{MG}})}{1 + x(k - \tau_{\mathrm{MG}})^{10}} - \beta x(k).$$

We generated a time-series using $\tau_{\mathrm{MG}} = 17, \alpha = 0.2, \beta = 0.1$, initial condition $x(0) = 1.2$, 0.1 as integration step and we trained the system to predict $\tau_f = 6$ step ahead. As we can see from Fig. 4(c) and the results in the table, for this test both ϕ (71% correlation) and λ (66% correlation) provide much better results than η (38% correlation) for identifying the optimal configuration.

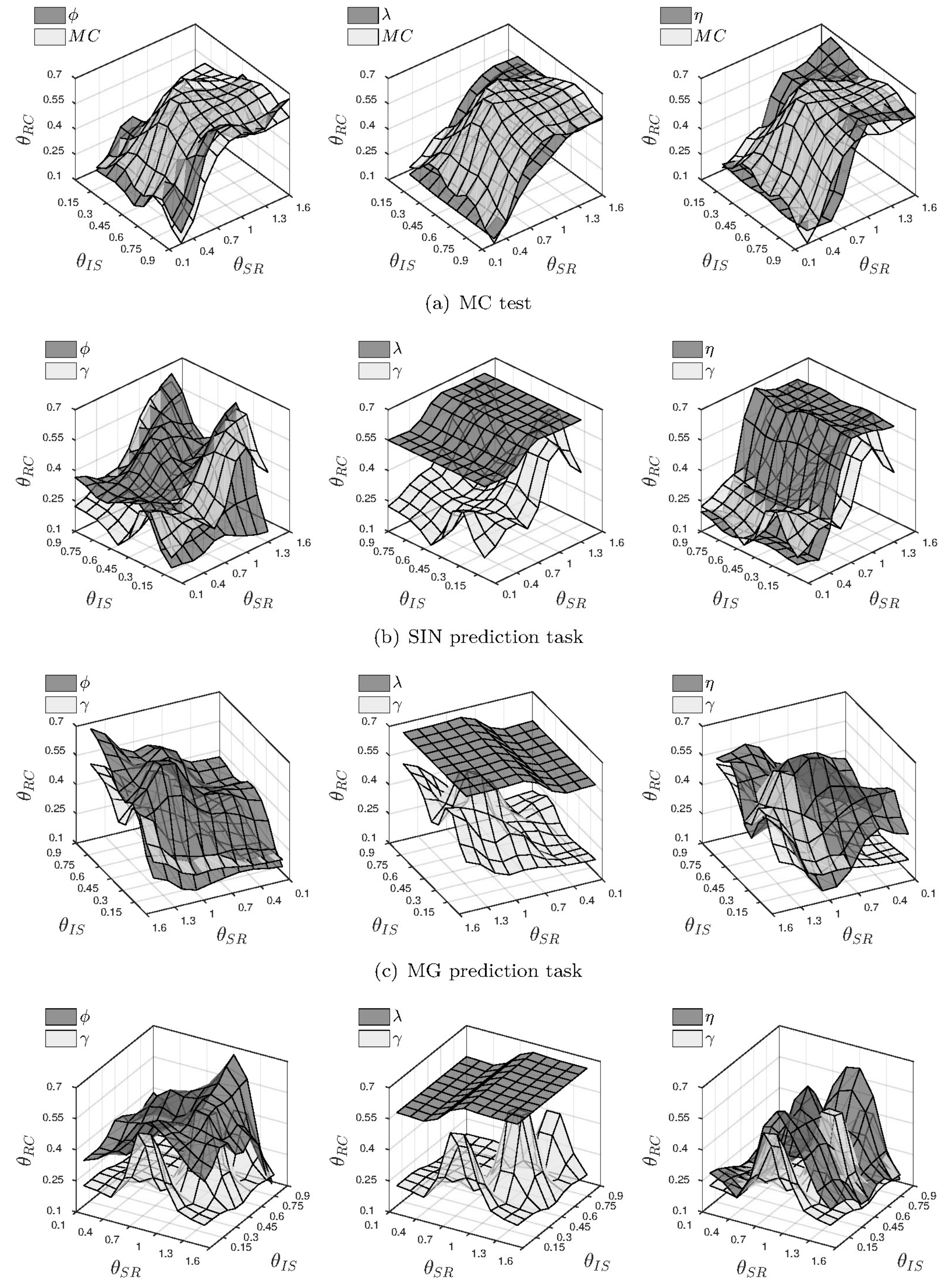
The *NARMA* task, originally proposed in [22], consists in modeling the output of the following order-r system:

$$\begin{aligned} \mathbf{y}[k+1] &= 0.3\mathbf{y}[k] + 0.05\mathbf{y}[k] \cdot \\ &\cdot \left(\sum_{i=0}^{r-1} \mathbf{y}[k-i]\right) + 1.5\mathbf{x}[k-r]\mathbf{x}[k] + 0.1, \end{aligned} \tag{8}$$

being $\mathbf{x}[k]$ an i.i.d. uniform noise in $[0, 1]$. According to the results shown in Fig. 4(d) and Tab. I, in this case ϕ and η achieve a correlation of 52% and 48% respectively. Hence, they perform significantly better than λ for identifying the critical region, which shows a very low correlation of 25% with γ. Even in this case, the best results are achieved by ϕ.

VII. Conclusions

Recurrent neural networks, as well as echo state networks, are driven by inputs and hence their dynamics and related computational capability depend on the type of the input driving signal. With this work, we have established for the first time a connection between the notion of continuous phase transition, ESNs and Fisher information. Based on



(a) MC test

(b) SIN prediction task

(c) MG prediction task

(d) NARMA prediction task

Fig. 4. In each row, we graphically represent the edge of criticality identified by the 3 unsupervised methods for a given task. The light gray manifold represents configurations of spectral radius (θ_{SR}), input scaling (θ_{IS}), and reservoir connectivity (θ_{RC}) that maximize Memory Capacity (MC) or prediction accuracy (γ). The dark gray manifolds represent (from left to right): configurations where the FIM determinant is maximized (ϕ); configurations where MLLE crosses zero (λ); configurations where mSVJ is maximized (η).

this interplay, we have developed a principled approach to configure ESNs on the edge of criticality. The proposed methodology allows the network designer to instantiate a specific architecture based on problem-dependent design choices and to identify optimal hyperparameters in an unsupervised way. Fisher information requires analytic knowledge of the distribution ruling the system. A crucial feature of our approach is that no assumptions regarding the mathematical model of the (input-driven) dynamic system are made, which makes the method independent of

TABLE I
Correlations between the regions where FIM determinant is maximized (ϕ), MLLE crosses zero (λ), mSVJ is maximized (η) and where performances are maximized (γ/MC). Each region is unrolled into a 1-dimensional vector and we compute the Pearson correlation between these vectors. Highest correlations are in bold.

Task	Corr ($\boldsymbol{\phi}$,$\boldsymbol{\gamma}$/MC)	Corr ($\boldsymbol{\lambda}$,$\boldsymbol{\gamma}$/MC)	Corr ($\boldsymbol{\eta}$,$\boldsymbol{\gamma}$/MC)
MC	0.75	**0.81**	0.65
SIN	**0.58**	0.52	0.56
MG	**0.71**	0.66	0.38
NARMA	**0.52**	0.25	0.48

the particular reservoir topology and the specific application under consideration. We have followed an ensemble estimation approach based on a recently proposed non-parametric FIM estimator, which, thanks to a graph-based representation of the data, is also applicable to high-dimensional densities. This last aspect plays a fundamental role in our domain of application, since we analyze the network through a multivariate sequence of reservoir neuron activations; hence the number of dimensions is determined by the number of reservoir neurons.

We evaluated the proposed method on benchmark tasks, designed to assess the computational capability of an ESN. Results are encouraging, since the FIM-based method identifies in every test with high precision the region of the hyperparameters space where prediction accuracy and memory capacity are maximized. We compared our method with established unsupervised criteria based on the sign of the maximum local Lyapunov exponent and the minimum singular value of the Jacobian. Our experiments provided empirical evidence that the proposed indicator describes well the ESN dynamics. In fact, in almost every test, the FIM-based approach outperforms the other unsupervised methods in identifying parameters that yield the highest supervised performance.

We believe that our approach opens new perspectives for analyzing the dynamics of input-driven RNNs. By linking notions taken from statistical mechanics with RNNs, a whole new set of studies and applications might become possible. Future research directions will be focused on transferring knowledge and methodologies between these two areas of research. Further applications of the FIM-based approach include testing the method on real-world applications and evaluating the possibility to reduce the dimensionality of the activations before estimating the FIM. Finally, we stress that, in principle, the proposed method could be used as an unsupervised criterion for training neuron connections in the recurrent layer. This could open interesting perspectives on the characterization of learning procedures in RNNs.

References

[1] M. Lukoševičius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Computer Science Review*, vol. 3, no. 3, pp. 127–149, 2009.

[2] D. Verstraeten, "An experimental comparison of reservoir computing methods," in *Invited talk at NIPS 2007 Workshop on Liquid State Machines and Echo State Networks*, 2006.

[3] F. M. Bianchi, E. De Santis, A. Rizzi, and A. Sadeghian, "Short-term electric load forecasting using echo state networks and PCA decomposition," *IEEE Access*, vol. 3, pp. 1931–1943, 2015.

[4] O. Obst and J. Boedecker, "Guided self-organization of input-driven recurrent neural networks," in *Guided Self-Organization: Inception* (M. Prokopenko, ed.), pp. 319–340, Heidelberg, Germany: Springer Berlin, 2014.

[5] J. Boedecker, O. Obst, J. T. Lizier, N. M. Mayer, and M. Asada, "Information processing in echo state networks at the edge of chaos," *Theory in Biosciences*, vol. 131, no. 3, pp. 205–213, 2012.

[6] F. Bianchi, L. Livi, and C. Alippi, "Investigating echo state networks dynamics by means of recurrence analysis," *arXiv preprint arXiv:1601.07381*, 2016.

[7] B. Schrauwen, L. Büsing, and R. Legenstein, "On computational power and the order-chaos phase transition in reservoir computing," in *Proceedings of the 22nd Annual conference on Neural Information Processing Systems*, vol. 21, pp. 1425–1432, NIPS Foundation, 2009.

[8] M. Massar and S. Massar, "Mean-field theory of echo state networks," *Physical Review E*, vol. 87, no. 4, p. 042809, 2013.

[9] D. Xu, J. Lan, and J. C. Principe, "Direct adaptive control: an echo state network and genetic algorithm approach," in *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, vol. 3, pp. 1483–1486, IEEE, 2005.

[10] M. Prokopenko, J. T. Lizier, O. Obst, and X. R. Wang, "Relating Fisher information to order parameters," *Physical Review E*, vol. 84, no. 4, p. 041116, 2011.

[11] R. Legenstein and W. Maass, "Edge of chaos and prediction of computational performance for neural circuit models," *Neural Networks*, vol. 20, no. 3, pp. 323–334, 2007.

[12] T. Mora and W. Bialek, "Are biological systems poised at criticality?," *Journal of Statistical Physics*, vol. 144, no. 2, pp. 268–302, 2011.

[13] J. Hidalgo, J. Grilli, S. Suweis, A. Maritan, and M. A. Muñoz, "Cooperation, competition and the emergence of criticality in communities of adaptive systems," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2016, no. 3, p. 033203, 2016.

[14] M. Scheffer, S. R. Carpenter, T. M. Lenton, J. Bascompte, W. Brock, V. Dakos, J. van De Koppel, I. A. van De Leemput, S. A. Levin, E. H. van Nes, M. Pascual, and J. Vandermeer, "Anticipating critical transitions," *Science*, vol. 338, no. 6105, pp. 344–348, 2012.

[15] P. Zegers, "Fisher information properties," *Entropy*, vol. 17, no. 7, pp. 4918–4939, 2015.

[16] M. Hermans and B. Schrauwen, "Recurrent kernel machines: Computing with infinite echo state networks," *Neural Computation*, vol. 24, no. 1, pp. 104–133, 2012.

[17] M. C. Ozturk, D. Xu, and J. C. Príncipe, "Analysis and design of echo state networks," *Neural Computation*, vol. 19, no. 1, pp. 111–138, 2007.

[18] G. Manjunath and H. Jaeger, "Echo state property linked to an input: Exploring a fundamental characteristic of recurrent neural networks," *Neural Computation*, vol. 25, no. 3, pp. 671–696, 2013.

[19] D. Verstraeten and B. Schrauwen, "On the quantification of dynamics in reservoir computing," in *Artificial Neural Networks–ICANN 2009*, pp. 985–994, Springer Berlin Heidelberg, 2009.

[20] I. Mastromatteo and M. Marsili, "On the criticality of inferred models," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2011, no. 10, p. P10012, 2011.

[21] V. Berisha and A. O. Hero III, "Empirical non-parametric estimation of the Fisher information," *IEEE Signal Processing Letters*, vol. 22, pp. 988–992, Jul. 2015.

[22] H. Jaeger, "Adaptive nonlinear system identification with echo state networks," in *Advances in Neural Information Processing Systems* (S. Becker, S. Thrun, and K. Obermayer, eds.), pp. 593–600, MIT Press, 2002.