Theoretical Foundation and Design Guideline for Reservoir Computing-Based MIMO-OFDM Symbol Detection

Shashank Jere, Ramin Safavinejad, and Lingjia Liu, *Senior Member, IEEE*

***Abstract*— In this paper, we derive a theoretical upper bound on the generalization error of reservoir computing (RC), a special category of recurrent neural networks (RNNs). The specific RC implementation considered in this paper is the echo state network (ESN), and an upper bound on its generalization error is derived via the empirical Rademacher complexity (ERC) approach. While recent work in deriving risk bounds for RC frameworks makes use of a non-standard ERC measure and a direct application of its definition, our work uses the standard ERC measure and tools allowing fair comparison with conventional RNNs. The derived result shows that the generalization error bound obtained for ESNs is tighter than the existing bound for vanilla RNNs, suggesting easier generalization for ESNs. With the ESN applied to symbol detection in MIMO-OFDM (Multiple Input Multiple Output-Orthogonal Frequency Division Multiplexing) systems, we show how the derived generalization error bound can guide underlying system design. Specifically, the derived bound together with the empirically characterized training loss is utilized to identify the optimum reservoir size in neurons for the ESN-based symbol detector. Finally, we corroborate our theoretical findings with results from simulations that employ 3GPP standards-compliant wireless channels, signifying the practical relevance of our work.**

***Index Terms*— Reservoir computing, echo state network, deep neural network, generalization error, receive processing, MIMO-OFDM, symbol detection.**

I. Introduction

DEEP Neural Networks (DNNs) [2] have delivered remarkable empirical performance on multi-dimensional grid-type datasets. Examples of these include image recognition [3], speech recognition [4] and language translation [5], to name a few. More recently, in the context of wireless networks, artificial intelligence (AI)-enabled cellular networks have been envisioned as the critical path towards realizing Beyond-5G networks [6]. In current 4G/5G systems, symbol detection methods are based on modeling the underlying wireless link and applying model-based signal processing techniques [7]. However, due to the dynamic nature of the underlying wireless channels (e.g. mmWave and Terahertz channels for Beyond-5G), it becomes extremely difficult to analytically model such behavior in a tractable and accurate manner. Furthermore, Beyond-5G (B5G) systems will be required to perform high speed transmit symbol detection at the receiver while supporting user mobility upto 500 km/h. In such scenarios, learning-based approaches to symbol detection, particularly those using neural networks can offer a promising alternative, in contrast with traditional model-based approaches which typically rely on accurate Channel State Information (CSI) which is not possible to obtain in the low Signal to Noise Ratio (SNR) regime. Additionally, end-to-end system non-linearities, e.g. due to Power Amplifiers in the transmitter or due to finite quantization resolution of analog-to-digital converters in the receiver can make traditional signal processing approaches to symbol detection challenging. Offline training driven DNN strategies such as *DetNet* [8], *MMNet* [9] have shown promising symbol detection performance in wireless channels, in some cases outperforming conventional model-based methods [9]. Since temporal correlation is inherent in wireless communications and recurrent neural networks (RNNs) are universal approximators of dynamic systems under fairly mild and general assumptions [10], we focus on the family of RNNs for symbol detection.

In this paper, we theoretically analyze reservoir computing (RC) [11], which is a special paradigm within the RNN family. RC avoids the back-propagation through time (BPTT)-incurred issue of vanishing and exploding gradients [12], which is encountered while training conventional RNNs. Furthermore, the training of RC is only conducted on the output layer of the particular RC network while its input layers and hidden layers are fixed after being initialized from a certain pre-determined distribution. Thus, the amount of training needed can be significantly reduced, leading to improved sample complexity. This makes the RC framework a promising candidate for wireless networks where online training data is extremely limited and physical layer operations are highly latency-sensitive, making conventional RNNs that have a prohibitively high training complexity almost unusable. Furthermore, RC-based receivers have been shown to outperform state-of-the-art model-based strategies as well as other NN-based receive processing in a

Manuscript received 30 August 2022; revised 31 January 2023; accepted 23 March 2023. Date of publication 21 April 2023; date of current version 18 September 2023. This work is supported in part by the U.S. National Science Foundation (NSF) under grant CNS-2003059. An earlier version of this paper was presented in part at the 2022 IEEE International Conference on Communications (ICC) in [DOI: 10.1109/ICC45855.2022.9839095]. The associate editor coordinating the review of this article and approving it for publication was Z. Qin. *(Corresponding author: Lingjia Liu.)*

The authors are with the Wireless@Virginia Tech, Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24060 USA (e-mail: ljliu@ieee.org).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2023.3263874.

Digital Object Identifier 10.1109/TCOMM.2023.3263874

variety of realistic environments, making it a promising NN-based technique for symbol detection in Beyond-5G (B5G) networks.

A. Our Contributions

With the growing prevalence of RC-based wireless symbol detectors, developing a principled understanding of the effectiveness of these approaches is critical in closing the existing knowledge gap in the theoretical insights into neural networks and RC in particular. Gaining a fundamental and clear understanding into the generalization performance of ESN-based detector architectures is key in designing such systems for symbol detection where the training data in the form of pilots is extremely limited. This work derives a generalization error bound for the single-layer ESN using tools from statistical learning theory and adapts it to develop a theory-guided procedure for optimum reservoir design in single-layer ESN-based detectors. This work, along with our most recent work in [13] which assigns model interpretability to single-neuron reservoir ESNs, contributes to the growing body of knowledge that will be crucial in developing "designable" machine learning solutions that go beyond a black-box view and rely on trial and error for model optimization. The main contributions of this work are summarized as follows:

- Using tools from statistical learning theory, we derive an upper bound on the generalization error for a single-layer (single reservoir) echo state network (ESN), which is a specific form of reservoir computing (RC).
- In the MIMO-OFDM symbol detection application, we prove analytically that the ESN training loss is a monotonically decreasing function of the reservoir size, which is a key hyperparameter that is tuned for performance.
- The derived generalization error bound is used in conjunction with the characterized training loss to develop a systematic procedure for the design of an optimum ESN-based symbol detector for MIMO-OFDM systems. This avoids the conventional trial and error process involved in hyper-parameter tuning of similar neural network-based methods.

The remainder of this paper is organized as follows. In Sec. II, we discuss existing work on generalization error bounds, particularly of RNNs and also review state-of-the-art NN-based approaches for wireless symbol detection, with an emphasis on RC-based methods. The problem setup and derivation of the generalization error bound for ESNs is elaborated in Sec. III. The MIMO-OFDM system model, applying ESNs for symbol detection and the optimum ESN-based detector design is introduced in Sec. IV. Sec. V provides numerical evaluation of the introduced procedure and corroboration with simulation results. Finally, Sec. VI concludes the paper.

Notation: We use the following notation throughout this paper: $\mathbf{C}$ is a matrix, $\mathbf{c}$ is a column vector, c is a scalar; $(\cdot)^T$ and $(\cdot)^H$ denote transpose and conjugate transpose respectively; $(\cdot)^\dagger$ denotes the Moore-Penrose matrix pseudoinverse; $\|\mathbf{C}\|_F$ is the Frobenius norm of $\mathbf{C}$, and $\|\mathbf{C}\|_2$ is its spectral norm; $\|\mathbf{c}\|_p$ is the p-norm of $\mathbf{c}$. $[\mathbf{c} \mid \mathbf{d}]$ and $[\mathbf{c} \mid \mathbf{d}]^T$ denote horizontal and vertical concatenation respectively of column vectors; $\mathbf{I}_N$ is the $N \times N$ identity matrix; $\mathbf{0}_{M\times N}$ is the $M \times N$ all-zeros matrix; $\mathcal{T}_n(\mathbf{c})$ denotes a $n \times n$ lower triangular Toeplitz matrix with the first column $\mathbf{c}$. We use short-hands 'TD' for 'time-domain', 'Tx' for 'transmit' and 'Rx' for 'receive'.

II. Related Work

A. Generalization Bounds for Recurrent Neural Networks

Generalization of neural networks, i.e., intuitively the difference in their performance between the training stage and the testing stage, has been a topic of deep investigation, including determining upper bounds on this generalization error for various neural network architectures. A more formal definition of generalization error and the problem formulation of finding its upper bound is provided in Sec. III-B. The generalization error of deep learning frameworks has been studied via many approaches including: 1) Model-based approaches such as: the Vapnik-Chervonenkis (VC) dimension theory [14], the Rademacher complexity approach [15], the Probably Approximately Correct (PAC)-Bayes theory [16]; and 2) Approaches that utilize learning theory-based metrics such as stability [17] and robustness [18]. Alternate approaches for deriving generalization bounds, such as norm-based methods, have been studied in [19]. Using these approaches, there have been a wide array of studies investigating the expressive ability of DNNs [20], the depth efficiency [21] for feedforward neural networks, and the generalization ability of specific neural network types such as convolutional neural networks (CNNs) [22].

Currently, there is limited research in the direction of generalization bounds of RNNs. A generalization bound for vanilla RNNs has been established in [23] using the PAC-Bayes approach. This bound contains the network size parameter J and increases as the square of the input sequence length t. Tighter generalization bounds for vanilla RNNs and its variants, including Minimal Gated United (MGU) and Long Short Term Memory (LSTM), have been derived in [24], where the bounds are tighter by a factor of t^2 compared to [23]. This work also utilizes the PAC-Bayes framework, incorporating the spectral norms of the RNN's weights matrices. [25] investigates RNN generalization bounds using the matrix 1-norm and the Fisher-Rao norm to get a tighter bound. With these techniques, network size parameters do not appear in the bound. However, this bound only applies to vanilla RNNs employing ReLU activations. More recently, there has been work [26] investigating the risk bounds of RC frameworks, including ESNs. In contrast with our work, however, it makes use of a special 'Rademacher-type' complexity measure instead of the standard empirical Rademacher Complexity (ERC) approach, thereby making it difficult to provide a fair comparison with available bounds for vanilla RNNs and making its extension to deeper structures potentially intractable. This work, on the other hand, uses the standard ERC measure, allowing tractable analysis and extension to deep ESN structures and makes way for a fair comparison with vanilla RNNs. In this work, we derive a theoretical upper bound on the generalization error

of the ESN and show that it is tighter than the best known bound for vanilla RNNs, thus implying easier generalization under limited training.

B. Neural Network-Based Wireless Symbol Detection

There has been considerable work done recently in applying deep neural network (DNN)-based strategies for symbol detection in wireless receivers. Multi-layer perceptron (MLP)-based symbol detection strategies have been introduced in prior works such as *DetNet* [8], *MMNet* [9], *OAMPNet* [27], and *HyperMIMO* [28], whereby each work uniquely incorporates trainable parameters from conventional iterative algorithms. While these approaches can achieve promising performance, they typically require large amount of training data, making them hard to utilize in cellular systems, e.g., LTE-Advanced and 5G NR, where training data is extremely limited. Additionally, they also usually need perfect CSI which is difficult, if not impossible, to obtain in practice.

Owing to the lightweight training characteristic of RC-based approaches, they offer a promising alternative to tackle the scarcity of over-the-air training data. ESNs were first applied as a symbol detector in MIMO-OFDM systems in [29]. Subsequent improvements in the ESN architecture such as the ability to handle a 'windowed' input were performed in [30] giving demonstrated performance gains. A novel deep RC structure *RCNet* was introduced in [31], while *RC-Struct* in [32] leverages the time-frequency structure of the OFDM waveform, both showing significant performance improvements over conventional signal processing techniques and other established learning-based approaches. Also, [33] focused on tracking channel change and updating RC weights adaptively within a subframe in high mobility environments with scattered pilots in Wi-Fi systems, while demonstrating superior performance on a real hardware testbed. A key advantage of the aforementioned RC-based methods compared to vanilla NN-based detectors is that the network training in the former is fully online leading to a significantly lower computational complexity. This also allows RC-based methods to be much more robust to changes in dynamic transmission mode and wireless environments as the underlying detector is trained completely online in every new subframe. In the same spirit as prior works, this paper provides a theoretical grounding towards understanding generalization abilities of RC-based symbol detectors and uses that insight in the design of an optimum RC-based symbol detector for MIMO-OFDM systems.

III. Generalization of Echo State Networks

A. Data Space and Network Definition

In this section, we set up the problem of generalization of single reservoir ESNs and define it formally. Our learning problem can be defined by the tuple $(\mathcal{Z}, \mathcal{P}, \mathcal{H}, \ell)$, where:

- $\mathcal{X}$ and $\mathcal{Y}$ are the input and output spaces respectively. In our case, $\mathcal{X} \in \mathbb{R}^{D\times T}$ represents a time sequence of length T. The output space is $\mathcal{Y} \in \mathbb{R}^K$ or $\mathcal{Y} \in \{0,1\}^K$, depending on whether the network is being employed for a regression or a classification task respectively.
- $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ represents the joint input-output space. $\mathcal{P}$ is the space of all probability distributions defined on $\mathcal{Z}$. $\mathcal{H}$ is the space of all predictors $h : \mathcal{X} \to \mathcal{Y}$ where h denotes the network function. The loss function $\ell(\cdot)$ is defined as $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

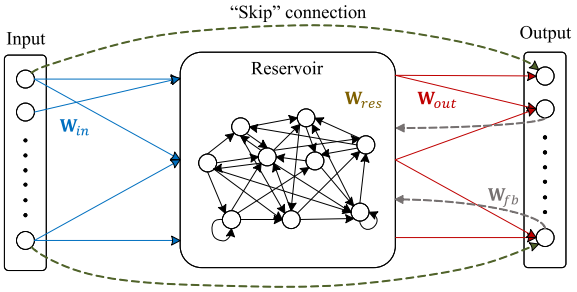

Fig. 1. A single-layer (single reservoir) Echo State Network (ESN). Dashed lines represent optional connections.

Let us define the input and output data along with the details of the ESN structure. Define an input sequence $\mathbf{U} = [\mathbf{u}(1), \mathbf{u}(2), \cdots, \mathbf{u}(T)]$ of length T such that $\mathbf{u}(t) \in \mathbb{R}^D$ and $\mathbf{U} \in \mathbb{R}^{D\times T}$, $t = 1, 2, \cdots, T$. Note that each data sample $\mathbf{u}(t)$ in $\mathbf{U}$ is a (column) vector of dimension D. For every training sequence $\mathbf{U}$, a label (ground truth) sequence $\mathbf{S}$ is available for training the network, where $\mathbf{S} = [\mathbf{s}(1), \mathbf{s}(2), \cdots, \mathbf{s}(T)]$ such that $\mathbf{s}(t) \in \mathbb{R}^K$ for a regression task and $\mathbf{s}(t) \in \{0,1\}^K$ for a classification task. The training set Z^N of size N is then defined as the set of input-label tuples $Z^N := \{(\mathbf{U}_1, \mathbf{S}_1), (\mathbf{U}_2, \mathbf{S}_2), \cdots, (\mathbf{U}_N, \mathbf{S}_N)\}$, where Z^N is generated i.i.d. according to some (unknown) probability distribution $P \in \mathcal{P}$. A single-layer ESN, i.e., with a single reservoir containing M neurons with random and sparse interconnections and a single output (readout) weights matrix is depicted in Fig. 1. The dashed lines from the input to the output denote 'skip' connections, originally introduced in [3] and a concatenated version introduced in [34], which will be elaborated in Sec. IV-B.

In the following, we define the input, output and the model weights for the ESN.

- $\mathbf{x}_{\text{res}}(t) \in \mathbb{R}^M$ is the state vector at the discrete time index t. $\mathbf{X}_{\text{res}} = [\mathbf{x}_{\text{res}}(1), \cdots, \mathbf{x}_{\text{res}}(T)] \in \mathbb{R}^{M\times T}$ is defined as the "reservoir states matrix" of the individual states across time from $t = 1$ to the end of the input sequence $t = T$ stacked sequentially.
- $\mathbf{x}_{\text{in}}(t) \in \mathbb{R}^D$ denotes the ESN input and $\mathbf{y}(t) \in \mathbb{R}^K$ denotes the ESN output.
- $\mathbf{W}_{\text{in}} \in \mathbb{R}^{M\times D}$ is the input weights matrix, $\mathbf{W}_{\text{res}} \in \mathbb{R}^{M\times M}$ is the reservoir weights matrix, $\mathbf{W}_{\text{out}} \in \mathbb{R}^{K\times M}$ is the output weights matrix, and $\mathbf{W}_{\text{fb}} \in \mathbb{R}^{M\times K}$ is the feedback weights matrix when teacher forcing [35] is enabled.

Let $t = \{1, 2, \cdots, T\}$ denote the discrete-time indices in a particular input sequence $\mathbf{U}$. For a single-reservoir ESN structure, its input $\mathbf{x}_{\text{in}}(t)$ is simply $\mathbf{u}(t)$. If $\sigma(\cdot)$ is a pointwise non-linear activation function, the state update equation and

the output equation are respectively:

$$\mathbf{x}_{\text{res}}(t) = \sigma\big(\mathbf{W}_{\text{res}}\mathbf{x}_{\text{res}}(t-1) + \mathbf{W}_{\text{in}}\mathbf{x}_{\text{in}}(t) + \mathbf{W}_{\text{fb}}\mathbf{s}(t-1)\big), \tag{1}$$

$$\mathbf{y}(t) = \mathbf{W}_{\text{out}}\mathbf{x}_{\text{res}}(t). \tag{2}$$

In this setup, $\mathbf{W}_{\text{in}}$, $\mathbf{W}_{\text{res}}$ and $\mathbf{W}_{\text{fb}}$ (if teacher forcing is enabled, otherwise $\mathbf{W}_{\text{fb}} = \mathbf{0}$) are initialized from a certain pre-determined distribution, e.g., the Uniform or Gaussian distributions, and then kept fixed throughout the training and inference (test) stages. For example, each element of $\mathbf{W}_{\text{in}}$, $\mathbf{W}_{\text{res}}$ and $\mathbf{W}_{\text{fb}}$ can be initialized independently from $\mathcal{U}(-1,1)$, $\mathcal{N}(0,1)$ or another distribution of choice. Unlike vanilla RNNs and its variants where all network weights are trained using BPTT, the only trainable network parameter in the ESN is $\mathbf{W}_{\text{out}}$, which is trained using a pseudoinverse-based closed-form linear update rule. This greatly reduces the number of trainable parameters as well as the training computational complexity, lending well to applications with limited training data availability. Additionally, the sparsity of $\mathbf{W}_{\text{res}}$ is controlled via the hyperparameter named 'sparsity' (denoted as κ), which represents the probability of an element of $\mathbf{W}_{\text{res}}$ being zero. The internal reservoir structure depicted in Fig. 1 depicts this random and potentially sparse nature of the interconnections between the neurons.

B. Problem Formulation

Given a tuple of an input sequence and the corresponding ground truth $\big(\mathbf{u}(t), \mathbf{s}(t)\big)_{t=1}^{T}$, we define $\mathbf{U}_t \in \mathbb{R}^{D\times t}$ by concatenating $\{\mathbf{u}(1), \mathbf{u}(1), \cdots, \mathbf{u}(t)\}$ into the columns of $\mathbf{U}_t$. Denote $\mathcal{F}_t = \{f_t : \mathbf{U}_t \to \mathbf{y}(t)\}$ as the class of mappings from the first t inputs to the t-th output $\mathbf{y}(t) = \mathbf{y}_t$, computed by the ESN structure. We also use $\mathbf{s}_t := \mathbf{s}(t)$ interchangeably for brevity of notation.

Unlike using a gradient-based algorithm such as Back-Propagation Through Time (BPTT) for training vanilla RNNs, only the readout (output) layer of the ESN needs to be trained, and this can be done by solving a simple problem such as minimizing an ℓ_2-loss via Least Squares (LS), which has a simple closed-form solution involving the reservoir states matrix's pseudo-inverse. An example training algorithm for the output weights matrix involves the Alternating Least Squares (ALS) method, such as that employed in [31], where the ESN is used for wireless symbol detection. Additionally, for the theoretical analysis in this paper, the loss function $\ell(\cdot)$ is only required to be bounded and Lipschitz continuous in $f_t \in \mathcal{F}_t$, i.e., for $\{(x_1, y_1), (x_2, y_2)\} \in \mathcal{X} \times \mathcal{Y}$,

$$|\ell(f_t(x_1), y_1) - \ell(f_t(x_2), y_2)| \le \rho_\ell |f_t(x_1) - f_t(x_2)|, \tag{3}$$

where ρ_ℓ is the Lipschitz constant for $\ell(\cdot)$. Depending on the type of task considered, $\ell(\cdot)$ can be chosen accordingly. Note that we do not impose the smoothness constraint on $\ell(\cdot)$. In regression tasks, the loss function considered is the ℓ_p-norm ($p \in \mathbb{Z}_+$), where typically $p = 2$. The output for a new 'test' input sequence $\mathbf{U}'$ is then simply $f_t(\mathbf{U}')$.

We now proceed to set up the learning problem. A learning algorithm A_t is defined as a mapping $A_t : Z^N \to \mathcal{F}_t$, i.e., $A(Z^N) = f_t \in \mathcal{F}_t$. Then, the *risk* of f_t is defined as:

$$\mathcal{L}(f_t) := \mathbb{E}_{\mathcal{Z}}\big(\ell\left(f_t(\mathbf{U}_t), \mathbf{S}_t\right)\big), \tag{4}$$

where the expectation is taken over the joint distribution of the input-output space $P \in \mathcal{P}$. The *minimum risk* is $\mathcal{L}_t^* := \inf_{f_t \in \mathcal{F}_t} \mathcal{L}(f_t)$. Since P is unknown, it is not possible to compute the risk $\mathcal{L}(f_t)$. Instead, we calculate the *empirical risk* based on the training dataset as follows:

$$\mathcal{L}_N(f_t) := \frac{1}{N}\sum_{n=1}^{N} \ell\big(f_t(\mathbf{U}_t), \mathbf{S}_t\big). \tag{5}$$

Given (4) and (5), a learning algorithm A_t is said to *generalize*, if for any $\epsilon > 0$, the following holds as $N \to \infty$ [36]:

$$\Pr\left(|\mathcal{L}(f_t) - \mathcal{L}_N(f_t)| \ge \epsilon\right) \to 0, \tag{6}$$

where in (6), the probability is defined over the randomness of the training set Z^N. In this work, our goal is to derive generalization bounds, under any probability distribution $P \in \mathcal{P}$, for an algorithm A_t that learns an ESN. To this end, we adopt the empirical Rademacher complexity (ERC) approach. The ERC is a measure of the "richness" of a function class $\mathcal{H}$, defined as [36]:

$$\mathcal{R}_N(\mathcal{H}) := \mathbb{E}_{\boldsymbol{\epsilon}^N}\left[\frac{1}{N}\sup_{h\in\mathcal{H}}\sum_{n=1}^{N}\epsilon_n \ell\big(h(\mathbf{U}_n), \mathbf{S}_n\big)\right], \tag{7}$$

where $\boldsymbol{\epsilon}^N := [\epsilon_1, \epsilon_2, \cdots, \epsilon_N]$ is a vector of i.i.d. Rademacher random variables, i.e., each $\epsilon_j \in \{1, -1\}$ with probabilities $\{\frac{1}{2}, \frac{1}{2}\}$ respectively for $j = 1, 2, \cdots, N$.

For any learning problem $(\mathcal{Z}, \mathcal{P}, \mathcal{H}, l)$, the ERC $\mathcal{R}_N(\mathcal{H})$ bounds generalization as follows:

Theorem 1 ([36]): For any probability distribution $P \in \mathcal{P}$ and any training set of size N, with probability at least $(1-\delta)$ for $\delta \in (0,1)$,

$$|\mathcal{L}(h) - \mathcal{L}_N(h)| \le 2\mathbb{E}_{Z^N}[\mathcal{R}_N(\mathcal{H})] + \sqrt{\frac{\log(\frac{1}{\delta})}{2N}}. \tag{8}$$

Eq. (8) holds for any $h \in \mathcal{H}$. Based on Eq. (8), our goal boils down to deriving an upper bound to $\mathcal{R}_N(\mathcal{F}_t)$, with $\mathcal{F}_t$ being the class of single-reservoir ESNs.

C. Main Results

In order to establish the main results of this paper, we first make the following mild assumptions for $t = 1, 2, \cdots, T$ that are common in learning theory literature [36].

Assumption 1: The input data is bounded, i.e., $\|\mathbf{u}(t)\|_2 \le B_{X_{in}}$.

Assumption 2: The ground truth and the output data are bounded, i.e. $\|\mathbf{s}(t)\|_2 \le B_S$.

Assumption 3: The spectral norms of the weights matrices are bounded, i.e., $\|\mathbf{W}_{in}\|_2 \le B_{W_{in}}$, $\|\mathbf{W}_{res}\|_2 \le B_{W_{res}}$, $\|\mathbf{W}_{out}\|_2 \le B_{W_{out}}$, $\|\mathbf{W}_{fb}\|_2 \le B_{W_{fb}}$.

Assumption 4: The activation function $\sigma(\cdot)$ is Lipschitz-continuous with Lipschitz constant ρ. Additionally, we assume $\sigma(0) = 0$. This holds for commonly used activation functions such as ReLU and hyperbolic tangent (Tanh) where, $ReLU(\cdot) = \max\{\cdot, 0\}$.

We are now ready to state the main result of this paper in Theorem 2 as follows.

Theorem 2: Let $\mathcal{F}_t = \{f_t : \mathbf{U}_t \to \mathbf{y}(t)\}$ be the class of single-reservoir Echo State Networks (ESNs). Under Assumptions 1-4, for every $f_t \in \mathcal{F}_t$ and $t \leq T$, its ERC is bounded by:

$$\mathcal{R}_N(\mathcal{F}_t) \leq \frac{4}{N} + 24r\sqrt{\frac{MK\log\left(2r\sqrt{JN}\right)}{N}}, \tag{9}$$

where $J = \sqrt{M^2+K^2}$, $r = \rho B B_{W_{out}} a_t$, $B = \left(B_{W_{in}}B_{X_{in}} + B_{W_{fb}}B_S\right)$ *and* $a_t = \frac{(\rho B_{W_{res}})^t - 1}{\rho B_{W_{res}} - 1}$.

D. Proof Strategy

In this section, we sketch the proof of the result obtained in Theorem 2, while stating and proving consequential lemmas along the way that build up to it. Our strategy is as follows:

1) For the single-reservoir ESN, formulate the Lipschitz continuity of its output w.r.t. the model parameters, i.e., reservoir states matrix, input weights matrix, output weights matrix and the feedback weights matrix.
2) Find an upper bound on the covering number of the function class $\mathcal{F}_t$.
3) Using the concept of chaining and Dudley's Entropy Integral, upper bound $\mathcal{R}_N(\mathcal{F}_t)$:
   - Specifically, we consider two different sets of *trainable* network weights matrices: $\mathbf{W}_{\text{out}}$ and $\mathbf{W}'_{\text{out}}$. Unlike conventional RNNs, only the output weights are trained in ESNs, while keeping $\mathbf{W}_{\text{in}}$ and $\mathbf{W}_{\text{res}}$ fixed according to a certain distribution.
   - For the same activation functions as well as the same input data, let the t^{th} output be $\mathbf{y}(t)$ and $\mathbf{y}'(t)$ when the two weights matrices are used respectively.

In what follows, we state lemmas which provide the tools required to prove Theorem 2. First, we characterize the Lipschitz property of $\|\mathbf{y}(t)\|_2$ in Lemma 1.

Lemma 1: Under the Assumptions 1-4 and for a given input, $\|\mathbf{y}(t)\|_2$ is Lipschitz-continuous in $\mathbf{W}_{out}$, i.e.,
$\|\mathbf{y}(t) - \mathbf{y}'(t)\|_2 \leq U_{res,t}\|\mathbf{W}_{out} - \mathbf{W}'_{out}\|_F$,
where $U_{res,t} = \rho(B_{W_{in}}B_{X_{in}} + B_{W_{fb}}B_S)\frac{(\rho B_{W_{res}})^t - 1}{\rho B_{W_{res}} - 1}$.

Proof: In order to prove the Lipschitz-continuity of the output $\mathbf{y}(t)$ of the ESN, note that

$$\begin{aligned}\|\mathbf{y}(t) - \mathbf{y}'(t)\|_2 &= \|\mathbf{W}_{\text{out}}\mathbf{x}_{\text{res}}(t) - \mathbf{W}'_{\text{out}}\mathbf{x}_{\text{res}}(t)\|_2 \\ &= \|(\mathbf{W}_{\text{out}} - \mathbf{W}'_{\text{out}})\,\mathbf{x}_{\text{res}}(t)\|_2 \\ &\overset{(a)}{\leq} \|\mathbf{x}_{\text{res}}(t)\|_2\,\|\mathbf{W}_{\text{out}} - \mathbf{W}'_{\text{out}}\|_2\,,\end{aligned} \tag{10}$$

where (a) holds by the Cauchy-Schwarz inequality. Next, we bound $\|\mathbf{x}_{\text{res}}(t)\|_2$ by establishing a recursive relation between $\|\mathbf{x}_{\text{res}}(t)\|_2$ and $\|\mathbf{x}_{\text{res}}(t-1)\|_2$. Recall that

$$\begin{aligned}&\|\mathbf{x}_{\text{res}}(t)\|_2 \\ &= \left\|\sigma\left(\mathbf{W}_{\text{in}}\mathbf{x}_{\text{in}}(t) + \mathbf{W}_{\text{res}}\mathbf{x}_{\text{res}}(t-1) + \mathbf{W}_{\text{fb}}\mathbf{s}(t-1)\right)\right\|_2 \\ &\overset{(a)}{\leq} \rho\left\|\mathbf{W}_{\text{in}}\mathbf{x}_{\text{in}}(t) + \mathbf{W}_{\text{res}}\mathbf{x}_{\text{res}}(t-1) + \mathbf{W}_{\text{fb}}\mathbf{s}(t-1)\right\|_2 \\ &\overset{(b)}{\leq} \rho\Big(\left\|\mathbf{W}_{\text{in}}\mathbf{x}_{\text{in}}(t)\right\|_2 + \left\|\mathbf{W}_{\text{res}}\mathbf{x}_{\text{res}}(t-1)\right\|_2 \\ &\quad + \left\|\mathbf{W}_{\text{fb}}\mathbf{s}(t-1)\right\|_2\Big) \\ &\overset{(c)}{\leq} \rho\Big(\left\|\mathbf{W}_{\text{in}}\right\|_2\left\|\mathbf{x}_{\text{in}}(t)\right\|_2 + \left\|\mathbf{W}_{\text{res}}\right\|_2\left\|\mathbf{x}_{\text{res}}(t-1)\right\|_2 \\ &\quad + \left\|\mathbf{W}_{\text{fb}}\right\|_2\left\|\mathbf{s}(t-1)\right\|_2\Big) \\ &\overset{(d)}{\leq} \rho\Big(B_{W_{\text{in}}}B_{X_{\text{in}}} + B_{W_{\text{fb}}}B_S + B_{W_{\text{res}}}\left\|\mathbf{x}_{\text{res}}(t-1)\right\|_2\Big).\end{aligned} \tag{11}$$

Here, (a) follows from the fact that $\sigma(\cdot)$ is Lipschitz-continuous with constant ρ, (b) follows from the triangle inequality, (c) follows from the Cauchy-Schwarz inequality and (d) follows from Assumptions 1-3. Applying (11) recursively with the initialization $\mathbf{x}_{\text{res}}(0) = \mathbf{0}_{M\times 1}$, we get

$$\begin{aligned}\|\mathbf{x}_{\text{res}}(t)\|_2 &\leq \rho\left(B_{W_{\text{in}}}B_{X_{\text{in}}} + B_{W_{\text{fb}}}B_S\right)\sum_{j=0}^{t-1}(\rho B_{W_{\text{res}}})^j \\ &= \rho\left(B_{W_{\text{in}}}B_{X_{\text{in}}} + B_{W_{\text{fb}}}B_S\right)\frac{(\rho B_{W_{\text{res}}})^t - 1}{\rho B_{W_{\text{res}}} - 1}.\end{aligned} \tag{12}$$

Therefore,

$$\|\mathbf{x}_{\text{res}}(t)\|_2 \leq U_{\text{res},t}, \tag{13}$$

where $U_{\text{res},t} = \rho\left(B_{W_{\text{in}}}B_{X_{\text{in}}} + B_{W_{\text{fb}}}B_S\right)\frac{(\rho B_{W_{\text{res}}})^t - 1}{\rho B_{W_{\text{res}}} - 1}$. Subsituting this in (10), it follows that

$$\begin{aligned}\|\mathbf{y}(t) - \mathbf{y}'(t)\|_2 &\leq U_{\text{res},t}\,\|\mathbf{W}_{\text{out}} - \mathbf{W}'_{\text{out}}\|_2 \\ &\overset{(a)}{\leq} U_{\text{res},t}\,\|\mathbf{W}_{\text{out}} - \mathbf{W}'_{\text{out}}\|_F\,,\end{aligned} \tag{14}$$

where (a) follows since $\|\mathbf{W}\|_2 \leq \|\mathbf{W}\|_F$ for a matrix $\mathbf{W}$, concluding the proof of Lemma 1. □

Next, we bound the covering number of the class $\mathcal{F}_t$. Let $\mathcal{N}\left(\mathcal{F}_t, \epsilon, \text{dist}(\cdot,\cdot)\right)$ denote its covering number. Then, the following result provides an upper bound.

Lemma 2: For $\epsilon > 0$, under Assumptions 1-4, $\mathcal{N}(\mathcal{F}_t, \epsilon, dist(\cdot,\cdot))$ is bounded by

$$\begin{aligned}&\mathcal{N}(\mathcal{F}_t, \epsilon, dist(\cdot,\cdot)) \\ &\leq \left(1 + 2B_{W_{out}}\frac{\sqrt{J}\rho\left(B_{W_{in}}B_{X_{in}} + B_{W_{fb}}B_S\right)\frac{(\rho B_{W_{res}})^t - 1}{\rho B_{W_{res}} - 1}}{\epsilon}\right)^{MK},\end{aligned} \tag{15}$$

where $J = \sqrt{M^2+K^2}$.

Proof: In order to construct a covering $\mathcal{C}\left(\mathcal{H}, \epsilon, \text{dist}(\cdot,\cdot)\right)$, it is required that for any $h \in \mathcal{H}$, there exists $h' \in \mathcal{H}$ for any input data $\{\mathbf{u}(t)\}_{t=1}^T$ that satisfies

$$\sup\|h(\mathbf{u}(t)) - h'(\mathbf{u}(t))\|_2 \leq \epsilon.$$

This is equivalent to $\sup\|\mathbf{y}(t) - \mathbf{y}'(t)\|_2 \leq \epsilon$. From Lemma 1, we know that

$$\sup\|\mathbf{y}(t) - \mathbf{y}'(t)\|_2 \leq U_{\text{res},t}\,\|\mathbf{W}_{\text{out}} - \mathbf{W}'_{\text{out}}\|_F\,.$$

Therefore, it suffices to construct a matrix covering for $\mathcal{C}\left(\mathbf{W}_{\text{out}}, \frac{\epsilon}{U_{\text{res},t}}, \|\cdot\|_F\right)$. Here, we use the following result

from [24] on the covering number of matrices with a bounded Frobenius norm, stated as Lemma 3 below.

Lemma 3: Let $\mathcal{G} = \{\mathbf{V} \in \mathbb{R}^{d_1 \times d_2} : \|\mathbf{V}\|_2 \leq \lambda\}$ *be the set of matrices with a bounded spectral norm* λ *and* $\epsilon > 0$ *be known. The covering number* $\mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|_F)$ *is upper bounded as*

$$\mathcal{N}(\mathcal{G}, \epsilon, \|\cdot\|_F) \leq \left(1 + 2\frac{\min\{\sqrt{d_1}, \sqrt{d_2}\}\lambda}{\epsilon}\right)^{d_1 d_2}. \quad (16)$$

The proof of this result uses the concept of a packing number of a set and can be found in [24]. Applying the result from Lemma 3 to $\mathbf{W}_{\text{out}}$, we get

$$\begin{aligned}
&\mathcal{N}\big(\mathcal{H}, \epsilon, \text{dist}(\cdot,\cdot)\big) \leq \mathcal{N}\left(\mathbf{W}_{\text{out}}, \frac{\epsilon}{U_{\text{res},t}}, \|\cdot\|_F\right) \\
&\leq \left(1 + 2\frac{\min\{\sqrt{M}, \sqrt{K}\}B_{W_{\text{out}}}U_{\text{res},t}}{\epsilon}\right)^{MK} \\
&\leq \Bigg(1 + 2B_{W_{\text{out}}}\min\{\sqrt{M}, \sqrt{K}\} \\
&\quad \times \frac{\rho\left(B_{W_{\text{in}}}B_{X_{\text{in}}} + B_{W_{\text{fb}}}B_S\right)\frac{(\rho B_{W_{\text{res}}})^t - 1}{\rho B_{W_{\text{res}}} - 1}\}}{\epsilon}\Bigg)^{MK} \\
&\overset{(a)}{\leq} \left(1 + 2B_{W_{\text{out}}}\frac{\sqrt{J}\rho\left(B_{W_{\text{in}}}B_{X_{\text{in}}} + B_{W_{\text{fb}}}B_S\right)\frac{(\rho B_{W_{\text{res}}})^t - 1}{\rho B_{W_{\text{res}}} - 1}}{\epsilon}\right)^{MK},
\end{aligned} \quad (17)$$

where $J = \sqrt{M^2 + K^2}$. This particular formulation of J results in $J > \max\{M, K\}$ and thus (a) holds, concluding the proof of Lemma 2. □

Also, note that for small $\epsilon > 0$ and after taking the logarithm on both sides, we get the following result which will be useful in the proof of Theorem 2.

$$\begin{aligned}
&\log\mathcal{N}\big(\mathcal{H}, \epsilon, \text{dist}(\cdot,\cdot)\big) \leq MK \\
&\times \log\left(2B_{W_{\text{out}}}\frac{\sqrt{J}\rho\left(B_{W_{\text{in}}}B_{X_{\text{in}}} + B_{W_{\text{fb}}}B_S\right)\frac{(\rho B_{W_{\text{res}}})^t - 1}{\rho B_{W_{\text{res}}} - 1}\}}{\epsilon}\right).
\end{aligned} \quad (18)$$

More recently, work in [37] has provided an upper bound on the covering number of a single reservoir ESN with an alternate definition [38] of a cover set. This definition takes into account the number of training sequences N and works with the range of available functions in $\mathcal{F}_t$ instead of the range of those function outputs. The subsequent result is expressed in the following lemma.

Lemma 4: For $\epsilon > 0$, *under Assumptions 1-4,* $\mathcal{N}\big(\mathcal{F}_t, \epsilon, dist(\cdot,\cdot)\big)$ *is bounded by*

$$\log\mathcal{N}\big(\mathcal{F}_t, \epsilon, dist(\cdot,\cdot)\big) \leq \frac{r_F^2 K}{\epsilon^2}\log(2MK), \quad (19)$$

where $r_F = \rho B_{W_{in},F}B_{X_{in},F}B_{W_{out},F}a_t$ *and* $B_{W_{in},F}$, $B_{X_{in},F}$, $B_{W_{out},F}$ *are defined as* $\|\mathbf{W}_{in}\|_F \leq B_{W_{in},F}$, $\|\mathbf{U}\|_F \leq B_{X_{in},F}$, $\|\mathbf{W}_{out}\|_F \leq B_{W_{out},F}$, $\|\mathbf{W}_{res}\|_F \leq B_{W_{res},F}$ *and* a_t *is as defined in (9).*

Finally, employing Dudley's Entropy Integral, we can arrive at an upper bound for the ERC of the single-reservoir ESN. For the purpose of our analysis, we use a slightly modified version of Dudley's Entropy Integral found in [38], which is restated in Lemma 5.

Lemma 5: For a real-valued function class $\mathcal{H}$ *with a bounded output, i.e., assuming values in the range* $[-r, r]$, *the ERC is bounded as*

$$\mathcal{R}_N(\mathcal{H}) \leq \inf_{\alpha > 0}\left(\frac{4\alpha}{\sqrt{N}} + \frac{12}{N}\int_{\alpha}^{2r\sqrt{N}}\sqrt{\log\mathcal{N}(\mathcal{H}, \epsilon, \|\cdot\|)}d\epsilon\right). \quad (20)$$

Using the bound obtained on $\mathcal{N}\big(\mathcal{F}_t, \epsilon, \text{dist}(\cdot,\cdot)\big)$ in Lemma 2 and substituting it in (20), we can obtain an upper bound for $\mathcal{R}_N(\mathcal{F}_t)$ as desired, which is stated in Theorem 2. The complete proof of Theorem 2 can be found in Appendix A. In addition, we also utilize the ESN covering number bound from Lemma 4 which is derived using the alternate cover set definition from [38]. A modified ERC bound similar to Theorem 2 is derived in Appendix C along with its asymptotic generalization gap which is utilized for optimum ESN-based symbol detector design in Sec. IV-E.

E. Comparison With RNN Generalization Bounds

The generalization error upper bound for a single layer ESN (From Theorem 2) as well as the tightest known ERC bound for vanilla RNNs [24] are both given in Table I. Although the bounds for both RNNs and ESNs scale with the network size, i.e., the 'width' parameter J (number of hidden units for RNNs and number of reservoir neurons for ESNs), the bound derived in this work for ESNs is independent of the sequence length t. Note that when $a_t = O(e^t)$, the ESN and RNN bounds will be of the same order, i.e., $O(t)$. However in ESNs, we configure $\mathbf{W}_{\text{res}}$ such that $B_{W_{\text{res}}} < 1$ almost always to satisfy the echo state property [35]. Thus, proper initialization of the reservoir ensures $\rho B_{W_{\text{res}}} < 1$, giving $a_t = O(1)$. This represents an asymptotic improvement (as $N \to \infty$) by a factor of $\sqrt{\log t}$ in the generalization gap of ESNs over vanilla RNNs.

IV. Symbol Detection in Modern Wireless Networks

One of the most promising applications of ESNs can be in standardized wireless communication systems, especially 5G and Beyond-5G (B5G) networks, where online training resources are extremely limited. Specifically, symbol detection is a critical classification task in any wireless receiver. The goal of the ESN-based symbol detector is to recover the frequency-domain QAM/PSK symbols from the corresponding time-domain observation. To accomplish this, an online learning method can be utilized in a supervised learning framework, where the term "online" emphasizes the fact that the training data for each subframe is present within the same subframe, without the need for prior offline training of the ESN. It is extremely important to be able to conduct symbol detection completely online since the operation modes of 5G and B5G can change on a sub-millisecond level leading to completely different environments from subframe (typically one millisecond) to subframe. For example, in the LTE/LTE-Advanced (4G) as well as 5G NR cellular standards, the first few

TABLE I
COMPARISON OF ASYMPTOTIC GENERALIZATION ERROR BOUNDS BETWEEN ESN AND VANILLA RNN

$\mathcal{R}_N\left(\mathcal{F}_t^{(\text{ESN})}\right) = O\left(J\sqrt{\frac{\log(\sqrt{JN})}{N}}\right)$	$\mathcal{R}_N\left(\mathcal{F}_t^{(\text{RNN})}\right) = O\left(J\sqrt{\frac{\log(t\sqrt{JN})}{N}}\right)$

(N_p) OFDM symbols within each subframe are pilot symbols known to the receiver that are used for synchronization and channel estimation in a conventional receiver. These pilot symbols are particularly limited and so the key challenge becomes: *How to achieve good generalization performance with the extremely limited training dataset for symbol detection?* With ESN-based symbol detection, the pilot symbols can be used as labels or ground truth to train the underlying ESN. Thus, no additional offline "training overhead" is required, unlike most existing neural network-based detection schemes. The system model for a MIMO-OFDM wireless communications system is described next.

A. MIMO-OFDM System Model

The TD Rx signal in a MIMO-OFDM system is $\mathbf{x}_i^{(m)} = \sum_{n=1}^{N_t} \widetilde{\mathbf{H}}_i^{(mn)} \widetilde{\mathbf{x}}_i^{(n)} + \mathbf{n}_i^{(m)}$, where:

- $\mathbf{x}_i^{(m)} \in \mathbb{C}^{(N_{\text{cp}}+N_{\text{sc}})}$: i^{th} TD OFDM symbol on the m^{th} Rx antenna out of N_r total Rx antennas.
- $\widetilde{\mathbf{x}}_i^{(n)} \in \mathbb{C}^{(N_{\text{cp}}+N_{\text{sc}})}$: i^{th} TD OFDM symbol from the n^{th} Tx antenna out of N_t total Tx antennas.
- $\widetilde{\mathbf{H}}_i^{(mn)} \in \mathbb{C}^{(N_{\text{cp}}+N_{\text{sc}})\times(N_{\text{cp}}+N_{\text{sc}})}$: Matrix of channel impulse response (CIR) coefficients between the n^{th} Tx antenna and the m^{th} Rx antenna for the i^{th} OFDM symbol. N_{sc} is the number of OFDM subcarriers and N_{cp} is the length of the cyclic prefix (CP). $L_t(< N_{\text{cp}})$ is the number of delay taps in the CIR vector $\mathbf{h}_i^{(mn)} = \left[h_{0,i}^{(mn)}\, h_{1,i}^{(mn)} \cdots h_{L_t-1,i}^{(mn)}\right]^T \in \mathbb{C}^{L_t\times 1}$. Note that $\widetilde{\mathbf{H}}_i^{(mn)} = \mathcal{T}_{(N_{\text{cp}}+N_{\text{sc}})}\left([\mathbf{h}_i^{(mn)} \mid \mathbf{0}_{(N_{\text{cp}}+N_{\text{sc}}-L_t)\times 1}]^T\right)$ is a lower triangular Toeplitz matrix.
- $\mathbf{n}_i^{(m)} \in \mathbb{C}^{(N_{\text{cp}}+N_{\text{sc}})\times 1}$: White Gaussian Noise (WGN) added to the i^{th} OFDM symbol added at the m^{th} Rx antenna, such that $\mathbf{n}_i^{(m)} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2\mathbf{I})$, where σ_n^2 is the noise variance.

B. ESN Dynamics for MIMO-OFDM Symbol Detection

Adapting the ESN dynamics to the symbol detection problem, the state update and output equations can be written as:

$$\mathbf{x}_{\text{res}}(t) = \sigma\big(\mathbf{W}_{\text{res}}\mathbf{x}_{\text{res}}(t-1) + \mathbf{W}_{\text{in}}\mathbf{x}(t)\big), \tag{21}$$

$$\mathbf{y}(t) = \mathbf{W}_{\text{out}}\mathbf{z}(t), \tag{22}$$

where $\mathbf{z}(t) = [\mathbf{x}_{\text{res}}(t)^T, \mathbf{x}(t)^T]^T \in \mathbb{C}^{(M+N_r)\times 1}$ is the concatenated vector of the ESN internal state and the input at time t. Note here that we concatenate the input $\mathbf{x}(t)$ to the "hidden" layer's output, in this case $\mathbf{x}_{\text{res}}(t)$ to compute the output, in accordance with the concatenation-style 'skip' connection introduced in [34]. This is unlike the dynamics of Eq. (2), where the input was not directly concatenated with the state in the output computation step. Teacher forcing is disabled, i.e., $\mathbf{W}_{\text{fb}} = \mathbf{0}$. The ESN parameters in the context of MIMO-OFDM symbol detection can be defined as follows:

- $\mathbf{x}(t) \in \mathbb{C}^{N_r\times 1}$ is the input to the ESN at time t on all N_r antennas; $\mathbf{y}(t) \in \mathbb{C}^{N_t\times 1}$ is the output of the ESN at time t on all N_t antennas.
- $\mathbf{W}_{\text{in}} \in \mathbb{C}^{M\times N_r}$ is the input weights matrix, $\mathbf{W}_{\text{res}} \in \mathbb{C}^{M\times M}$ is the reservoir (recurrent) weights matrix, and $\mathbf{W}_{\text{out}} \in \mathbb{C}^{N_t\times(M+N_r)}$ is the output weights matrix.

Due to the inherent transmit power constraint in practical wireless communication systems and the ESN parameter initialization process which satisfies the echo state property (ESP), Assumptions 1-3 from Sec. III-B still hold for the ESN input and output and the parameter weights matrices. The training phase of the ESN comprises finding the optimal output weights matrix $\widehat{\mathbf{W}}_{\text{out}}$ via linear regression, such that the mean squared error between the ESN output and the actual transmitted signal (ground truth) is minimized. This can be described as:

$$\begin{aligned}\widehat{\mathbf{W}}_{\text{out}} &= \arg\min_{\mathbf{W}_{\text{out}}} \sum_{i=0}^{N_p-1}\sum_{t=0}^{N_{cp}+N_{sc}-1} \|\mathbf{y}_i(t) - \widetilde{\mathbf{x}}_i(t)\|_2^2 \\ &= \arg\min_{\mathbf{W}_{\text{out}}} \sum_{i=0}^{N_p-1} \left\|\mathbf{Y}_i - \widetilde{\mathbf{X}}_i\right\|_F^2 = \arg\min_{\mathbf{W}_{\text{out}}} \left\|\mathbf{Y} - \widetilde{\mathbf{X}}\right\|_F^2,\end{aligned} \tag{23}$$

where $\mathbf{Y}_i = [\mathbf{y}_i(1), \mathbf{y}_i(2), \cdots, \mathbf{y}_i(N_{cp}+N_{sc})] \in \mathbb{C}^{N_t\times(N_{cp}+N_{sc})}$ is the matrix form of the ESN output, covering all $(N_{\text{cp}}+N_{\text{sc}})$ samples of the i^{th} OFDM symbol. Similarly, the matrix form of the transmitted signal (target/label) is $\widetilde{\mathbf{X}}_i = [\tilde{\mathbf{x}}_i(1), \tilde{\mathbf{x}}_i(2), \cdots, \tilde{\mathbf{x}}_i(N_{cp}+N_{sc})] \in \mathbb{C}^{N_t\times(N_{cp}+N_{sc})}$. Finally, collecting all the N_p pilot OFDM symbols in the training set, $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_{N_p}] \in \mathbb{C}^{N_t\times N_p(N_{cp}+N_{sc})}$ and $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2, \cdots, \widetilde{\mathbf{X}}_{N_p}] \in \mathbb{C}^{N_t\times N_p(N_{cp}+N_{sc})}$. Also during the training phase, the concatenated reservoir state vector $\mathbf{z}(t)$ is recorded at each time step t and stacked to form the "reservoir state matrix" according to $\mathbf{Z}_i = [\mathbf{z}_i(1), \mathbf{z}_i(2), \cdots, \mathbf{z}_i(N_{\text{cp}}+N_{\text{sc}})] \in \mathbb{C}^{(M+N_r)\times(N_{\text{cp}}+N_{\text{sc}})}$. Then, the reservoir states are collected across the N_p pilot/training OFDM symbols in the matrix $\mathbf{Z}$ as $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_{N_p}] \in \mathbb{C}^{(M+N_r)\times N_p(N_{\text{cp}}+N_{\text{sc}})}$. Finally, the optimal $\widehat{\mathbf{W}}_{\text{out}}$ is found using the solution of the least squares regression problem [35] of Eq. (23), i.e.,

$$\widehat{\mathbf{W}}_{\text{out}} = \widetilde{\mathbf{X}}\mathbf{Z}^{\dagger}, \tag{24}$$

where $\mathbf{Z}^{\dagger}$ is the Moore-Penrose pseudoinverse of $\mathbf{Z}$, defined as $\mathbf{Z}^{\dagger} = \mathbf{Z}^H\left(\mathbf{Z}\mathbf{Z}^H\right)^{-1}$.

With regards to the input of the ESN for TD regression, it is empirically shown in our recent work [39] that the temporal correlation caused by the channel is embedded contiguously in both the CP and the non-CP (payload) part of the Rx TD OFDM symbols. Thus, removing the CP before being

input to the ESN creates a discontinuity in the original temporal correlation, resulting in the information learnt by the ESN being incomplete and the test BER ($P_{b,\text{test}}$) being high. Therefore, we retain the CP in the input to the ESN in the training procedure setup of Eq. (23).

C. Analyzing Trends in the ESN Training Loss

The input-output relation of the ESN can be written in matrix form as $\mathbf{Y} = \mathbf{W}_{\text{out}}\sigma\big(\mathbf{W}_{\text{res}}\mathbf{Z} + \mathbf{W}_{\text{in}}\mathbf{X}\big)$, where $\mathbf{X} \in \mathbb{C}^{N_r \times N_p(N_{\text{cp}}+N_{\text{sc}})}$ is the matrix of received time-domain OFDM symbols over the training set. After the ESN is trained, i.e., $\widehat{\mathbf{W}}_{\text{out}}$ is determined using Eq. (24), the above matrix equation becomes $\mathbf{Y} = \widetilde{\mathbf{X}}\mathbf{Z}^{\dagger}\sigma\big(\mathbf{W}_{\text{res}}\mathbf{Z} + \mathbf{W}_{\text{in}}\mathbf{X}\big)$. Then, the training loss $\mathcal{L}_{\text{train}}$ is

$$\begin{aligned}\mathcal{L}_{\text{train}} &= \frac{1}{N_p}\|\mathbf{Y} - \widetilde{\mathbf{X}}\|_F^2 \qquad\qquad = \frac{1}{N_p}\|\mathbf{Y}^T - \widetilde{\mathbf{X}}^T\|_F^2 \\ &= \frac{1}{N_p}\|\mathbf{V}\widetilde{\mathbf{X}}^T - \widetilde{\mathbf{X}}^T\|_F^2. \end{aligned} \tag{25}$$

The square matrix $\mathbf{V} \in \mathcal{C}^{N_p(N_{\text{cp}}+N_{\text{sc}}) \times N_p(N_{\text{cp}}+N_{\text{sc}})}$ is given by

$$\begin{aligned}\mathbf{V} &= \big(\mathbf{Z}^{\dagger}\sigma\big(\mathbf{W}_{\text{res}}\mathbf{Z} + \mathbf{W}_{\text{in}}\mathbf{X}\big)\big)^T \\ &= \big(\sigma\left(\mathbf{W}_{\text{res}}\mathbf{Z} + \mathbf{W}_{\text{in}}\mathbf{X}\right)\big)^T\big(\mathbf{Z}^{\dagger}\big)^T = \mathbf{A}\mathbf{B}, \end{aligned} \tag{26}$$

where $\mathbf{A} = \big(\sigma\left(\mathbf{W}_{\text{res}}\mathbf{Z} + \mathbf{W}_{\text{in}}\mathbf{X}\right)\big)^T \in \mathbb{C}^{N_p(N_{\text{cp}}+N_{\text{sc}}) \times (M+N_r)}$ and $\mathbf{B} = \big(\mathbf{Z}^{\dagger}\big)^T \in \mathbb{C}^{(M+N_r) \times N_p(N_{\text{cp}}+N_{\text{sc}})}$. We know that a matrix $\mathbf{V}_{m\times m}$ has rank ν if it can be factorized as $\mathbf{V}_{m\times n} = \mathbf{A}_{m\times\nu}\mathbf{B}_{\nu\times m}$. Since one of the dimensions is much larger than the other in both $\mathbf{A}$ and $\mathbf{B}$, i.e., $N_p(N_{\text{cp}} + N_{\text{sc}}) \gg (M + N_r)$, by definition, rank$(\mathbf{V}) = (M + N_r)$. Going back to Eq. (25), we can see that $\mathcal{L}_{\text{train}} \to 0$ when $\mathbf{V} \to \mathbf{I}_{m\times m}$. One of the necessary conditions to satisfy this is that $\mathbf{V}$ must be full rank, i.e., $(M + N_r)$ must approach $N_p(N_{\text{cp}} + N_{\text{sc}})$. Since N_r is a system parameter and thus fixed, we can conclude that the training loss $\mathcal{L}_{\text{train}}$ decreases monotonically with the reservoir size M, i.e., number of neurons in the reservoir. We will show this experimentally in Sec. V.

D. Simpler ESN Dynamics for MIMO-OFDM Symbol Detection

A simpler ESN dynamics model in the symbol detection application can be written without the state vector being concatenated to the input, i.e., without the skip connection. following the dynamics of Eq. (1) and Eq. (2) with $\mathbf{W}_{\text{fb}} = \mathbf{0}$. The previous derivation in Sec. IV-C for the trend in $\mathcal{L}_{\text{train}}$ is still valid with $\mathbf{Z}$ being replaced by $\mathbf{X}_{\text{res}}$ and $\mathbf{W}_{\text{out}} \in \mathbb{C}^{N_t \times M}$. Note that this is the same ESN model for which an upper bound on the generalization error was derived in Theorem 2. The main motivation for using the concatenated ESN model from Sec. IV-B especially for wireless symbol detection is two-fold:

1) Significantly lower $P_{b,\text{train}}$ and $P_{b,\text{test}}$, i.e., superior performance for given channel statistics.
2) Greater robustness of $\mathcal{L}_{\text{test}}$ and thereby $P_{b,\text{test}}$ to change in the reservoir size M.

We show the above outcomes with and without input concatenation to the state experimentally in Sec. V. The generalization bound derived in Theorem 2 and that derived in Appendix C using the alternate cover set definition, both change slightly when using the concatenated dynamics model of Eq. (22). This minor change in the ERC bound in either case under the concatenated model is outlined in Appendix B. Its effect on the asymptotic gap is provided in Sec. IV-E.

E. System Design Guidance

The objective of the ESN-based symbol detector design process is to optimize the reservoir size M such that the test BER ($P_{b,\text{test}}$) is minimized. This is because $P_{b,\text{test}}$ is one of the main key performance indicators (KPI) quantifying the reliability of a symbol detector deployed in practice. However, directly minimizing $P_{b,\text{test}}$, which can be characterized as an ℓ_0-norm, may not necessarily result in a convex optimization problem. Therefore, we focus on the test "loss" $\mathcal{L}_{\text{test}}$ instead, which makes the optimization problem convex since $\mathcal{L}_{\text{test}}$ is defined as a function of an ℓ_2-norm. Similar to $\mathcal{L}_{\text{train}}$, $\mathcal{L}_{\text{test}}$ is defined as $\mathcal{L}_{\text{test}} = \frac{1}{N_d}\|\mathbf{Y} - \widetilde{\mathbf{X}}\|_F^2$, where $\mathbf{Y}$ and $\widetilde{\mathbf{X}}$ are respectively the ESN output and the transmitted OFDM symbols in the testing set (payload) consisting of N_d OFDM symbols (typically the last N_d OFDM symbols in a subframe).

In the optimum design process, we assume that the statistics of the wireless channel are known to us during the design procedure. This includes knowledge of the empirical distributions of relevant channel parameters including but not limited to: i) Number of (dominant) multipath components, ii) Angle of Arrival (AoA) and Angle of Departure (AoD) angular spreads, iii) Number of clusters, and iv) Path loss for each path. Since the received time-domain symbols $\mathbf{x}_i^{(m)}\ \forall\ m = (1, 2, \cdots, N_r)$ are a function of the wireless channel $\widetilde{\mathbf{H}}_i^{(mn)}$, the optimum ESN design procedure must take into account channel statistics. This can be done by evaluating the expected value of the training loss $\mathbb{E}_H[\mathcal{L}_{\text{train}}]$, where the expectation is taken w.r.t. the realization of the channel $\widetilde{\mathbf{H}}_i$. Typically, arriving at an analytical expression for $\mathbb{E}_H[\mathcal{L}_{\text{train}}]$ can be challenging, especially with the non-linear activation $\sigma(\cdot)$ in the ESN dynamics equation. To overcome this issue, we can get an empirical approximation for $\mathbb{E}_H[\mathcal{L}_{\text{train}}]$, for which a large number of channel realizations e.g., $\sim 10^3$ or higher, from a known channel statistical distribution can be used. The empirical distributions of important channel parameters can be typically measured in the field, or are available in standards documents such as those from 3GPP, e.g., the Extended Pedestrian-A (EPA) model, Clustered Delay Line (CDL) models, etc. Alternatively, simulators such as QuaDRiGa [40] enable performance evaluation with 3GPP-compliant channel models and scenarios adhering to realistic electromagnetic environments, thereby allowing the computation of $\mathbb{E}_H[\mathcal{L}_{\text{train}}]$.

The second part required for optimum ESN design is a numerical approximation of the generalization error bound, i.e., the difference between $\mathcal{L}_{\text{train}}$ and $\mathcal{L}_{\text{test}}$. We approximate $\mathcal{L}_{\text{gap}}$ asymptotically ($N \to \infty$) for $\mathcal{L}_{\text{gap}}$, where only the dominant term is retained in Eq. (9), i.e.,

$$\mathcal{L}_{\text{gap}} = \beta O\left(\sqrt{\frac{MN_t \log\sqrt{MN_p}}{N_p}}\right). \tag{27}$$

Here, the value of β depends on whether the state-input concatenated or non-concatenated model is used. For the non-concatenated model, β is given by $\beta = 24B_{W_{\text{in}}}B_{X_{\text{in}}}B_{W_{\text{out}}}a_0$ while for the concatenated model $\beta = 24B_{W_{\text{out}}}\sqrt{(B_{W_{\text{in}}}B_{X_{\text{in}}}a_0)^2 + B_{X_{in}}^2}$ (see Appendix B), and $a_0 = \frac{(\rho B_{W_{\text{res}}})^{N_{\text{cp}}+N_{\text{sc}}}-1}{\rho B_{W_{\text{res}}}-1}$. Using the alternate definition of the cover set from [38], the alternate asymptotic approximation for the generalization gap $\mathcal{L}'_{\text{gap}}$ is given by (full derivation in Appendix C):

$$\mathcal{L}'_{\text{gap}} = \beta' O\left(\frac{\log(\beta N_p)}{2N_p}\sqrt{N_t \log(2MN_t)}\right), \tag{28}$$

where $\beta' = 24\rho B_{W_{\text{in},F}}B_{X_{\text{in},F}}B_{W_{\text{out},F}}a_0 \leq \beta\sqrt{N_r N_p M}$. Finally, the overall optimum ESN-based detector design process, i.e., finding the best M_{opt} that minimizes $P_{b,\text{test}}$, can be summarized as: 1) For each potential value of M, numerically evaluate $\mathbb{E}_H[\mathcal{L}_{\text{train}}]$, 2) Add the theoretical derived generalization gap upper bound $\mathcal{L}_{\text{gap}}$ or $\mathcal{L}'_{\text{gap}}$ to the numerically evaluated $\mathbb{E}_H[\mathcal{L}_{\text{train}}]$ to get an upper bound on $\mathcal{L}_{\text{test}}$, and 3) Find the value of M that minimizes $\mathcal{L}_{\text{test}}$.

F. Complexity Analysis

We also perform complexity analysis for both the concatenated and the non-concatenated ESN detector models and especially their comparison with a conventional method, e.g., LMMSE channel estimation with LMMSE symbol detection. For this analysis, assume $N_{\text{ant}} = N_t = N_r$ is the number of antennas each at the transmitter and the receiver. In our previous work [30], we have shown that the training complexity in terms of FLOPS for the ESN is $\mathcal{O}(MN_{\text{sc}}N_{\text{ant}}^2 + M^2N_{\text{sc}}N_{\text{ant}} + M^3)$, while its testing complexity is only $\mathcal{O}(MN_{\text{sc}}N_{\text{ant}})$. On the other hand, the overall complexity for LMMSE channel estimation with symbol detection is $\mathcal{O}(N_{\text{sc}}^2N_{\text{ant}}^2 + N_{\text{sc}}N_{\text{ant}}^3)$. This suggests that when M is small and N_{sc} is large, the ESN symbol detector has a significantly lower overall computational cost than the conventional LMMSE symbol detector. The implications of this for both the concatenated and the non-concatenated ESN models are discussed in Sec. V.

V. Simulation Results

In this section, we simulate the performance of the ESN-based symbol detector in a MIMO-OFDM system. The primary objective here is to first, empirically validate the derived generalization error bound, especially in comparison with vanilla RNN variants when both are used for MIMO-OFDM symbol detection. Second, we validate the theory-guided system design procedure with the empirically observed optimum reservoir size for both the concatenated and the non-concatenated ESN models. The optimum design guidance is evaluated under two realistic wireless channel conditions namely, i) 3GPP-specified EPA (Extended Pedestrian-A) channel [41] with a maximum Doppler shift of 20 Hz, and ii) '3GPP_3D_UMa_NLOS' (Urban Macro) and 'BERLIN_UMa_NLOS' scenarios, both generated with the 3GPP-compliant QuaDRiGa simulator [40]. The latter

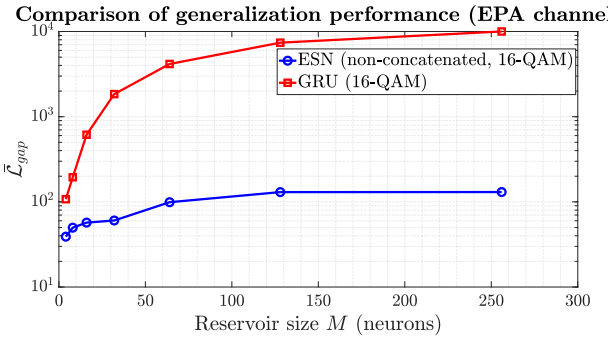

Fig. 2. Comparison of empirical generalization gap between ESN and Gated Recurrent Unit (GRU) for $N = 4$ training "sequences" (pilot OFDM symbols) in MIMO-OFDM symbol detection.

QuaDRiGa scenario is based on terrestrial macrocell measurements performed in Berlin, Germany. The MIMO-OFDM system settings, QuaDRiGa channel scenario settings and the ESN hyperparameter values are outlined in Table II.

A. Generalization Bound Comparison

In this section, we compare the empirical generalization gap ($\bar{\mathcal{L}}_{\text{gap}} = \mathcal{L}_{\text{test}} - \mathcal{L}_{\text{train}}$) between the single-layer ESN symbol detector (non-concatenated version) and a conventional RNN variant namely, the Gated Recurrent Unit (GRU). From Fig. 2, we can see the ESN generalizes much faster with $\bar{\mathcal{L}}_{\text{gap}}$ being at least one order of magnitude greater for the GRU for small M (e.g., $M < 50$), while being almost two orders of magnitude greater for higher values of M (> 100). This empirical result confirms the insight drawn from our theoretical generalization error gap for ESNs and RNNs, and its direct implication for MIMO-OFDM symbol detection, namely that vanilla RNN variants exhibit severe underfitting due to the scant training data in the form of pilots and display a large generalization error, leading to poor test performance. Exhaustive comparisons with multiple RNN variants have been detailed in our prior work [31].

B. Validating Theoretical Design Guidance

In this section, we corroborate the optimum ESN design suggested by theory, i.e., $\widehat{\mathcal{L}}_{\text{test}} = \mathbb{E}_H[\mathcal{L}_{\text{train}}] + \widehat{\mathcal{L}}_{\text{gap}}$, where $\widehat{\mathcal{L}}_{\text{gap}}$ is either $\mathcal{L}_{\text{gap}}$ or $\mathcal{L}'_{\text{gap}}$, with the actual performance metric during test, i.e., $P_{b,\text{test}}$. Next, we compute the test loss for design guidance $\widehat{\mathcal{L}}_{\text{test}}$, where $\mathbb{E}_H[\mathcal{L}_{\text{train}}]$ is evaluated numerically for a given channel distribution, and $\mathcal{L}_{\text{gap}}$ is calculated using the asymptotic expressions of Eq. (27) and Eq. (28). To approximate $\mathbb{E}_H[\mathcal{L}_{\text{train}}]$, we use 5000 different channel realizations, i.e., $\mathbb{E}_H[\mathcal{L}_{\text{train}}] \approx \frac{1}{5000}\sum_{n_c=1}^{5000}\mathcal{L}_{\text{train},n_c}$, where $\mathcal{L}_{\text{train},n_c}$ is the training loss evaluated for the n_c^{th} realization from the channel scenario under consideration. $P_{b,\text{test}}$ is simulated with 1000 Monte-Carlo simulation runs. The range of M is $[4, 1024]$ in Fig. 3 and $[512, 2048]$ in Fig. 4.

First, note from both Fig. 3 and Fig. 4 that for both channel models, $\mathbb{E}_H[\mathcal{L}_{\text{train}}]$ decreases monotonically with M, which is completely aligned with our theoretical justification in Sec. IV-C for this trend. Secondly, from Fig. 3 for the concatenated model (with skip connection), the theoretical guidance

TABLE II
SIMULATION PARAMETER SETTINGS

MIMO-OFDM Settings		QuaDRiGA Channel Parameters		ESN Parameters	
N_t	4	Scenarios	'3GPP_3D_UMa_NLOS' 'BERLIN_UMa_NLOS'	ρ	1
N_r	4	UE Speed	5 km/hr	κ	0.4
N_{sc}	1024	UE antenna height	1.5 m	$B_{W_{res}}$	0.2
N_{cp}	160	BS to UE distance	$\mathcal{U}(10, 500)$ m	$B_{W_{out}}$	0.665
N_p	4	BS antenna height	25 m	β (concatenated)	~ 18
N_d	9	BS antenna type	2×2 UPA	β (non-concatenated)	~ 5
Modulations	QPSK, 16-QAM	E_b/N_0	15 dB		

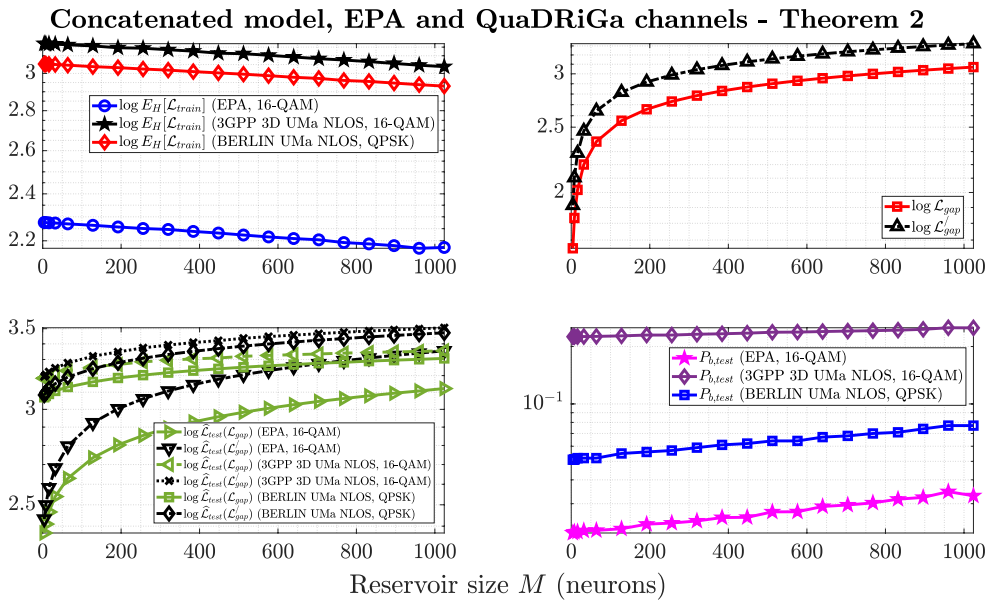

Fig. 3. State-input concatenated model (with skip connection) ESN symbol detector simulated in EPA and QuaDRiGa channels.

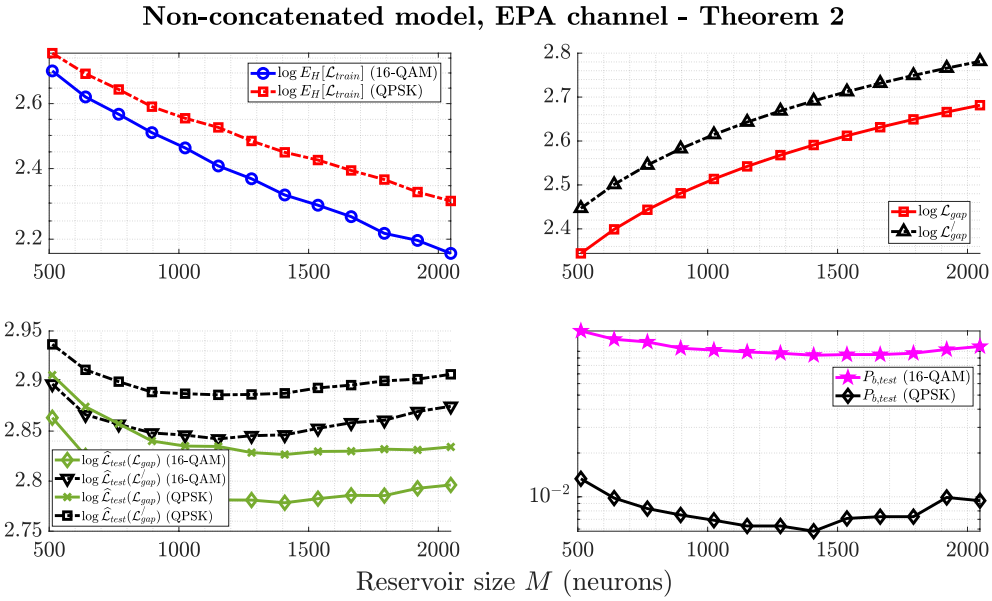

Fig. 4. State-input non-concatenated model (without skip connection) ESN symbol detector simulated in EPA channel.

matches well with the actual simulated test BER, i.e., $P_{b,\text{test}}$ is minimized at $M_{\text{opt}} = 4$ in all cases using the theoretical design guidance provided by either $\mathcal{L}_{\text{gap}}$ or $\mathcal{L}'_{\text{gap}}$. The higher $P_{b,\text{test}} \sim 0.2$ for the QuaDRiGa channel scenarios compared to the EPA channel is also seen in our prior work [32], where other methods such as *MMNet* [9] perform much worse in the

challenging (ill-conditioned due to correlation) QuaDRiGa-generated channel scenarios. Additionally in Fig. 4, we also present the results under the EPA channel utilizing the simpler ESN dynamics model as outlined in Eq. (2). The non-concatenated ESN model (without skip connection) results in a prohibitively high $P_{b,\text{test}} \sim 0.4$ for $M \in [4, 2048]$ under the QuaDRiGa-generated channel scenarios, so that such a detector cannot be used reliably in practice. We can observe the following from Fig. 4:

- Not concatenating the state vector with the input in the ESN output equation, i.e., no 'skip' connections [34], results in significantly higher $\mathbb{E}_H[\mathcal{L}_{\text{train}}]$, $\widehat{\mathcal{L}}_{\text{test}}$ and thus $P_{b,\text{test}}$.
- The optimum reservoir size M_{opt} that minimizes $P_{b,\text{test}}$ with the non-concatenated model is much higher compared to the concatenated model (with skip connection). The optimum M_{opt} suggested by $\mathcal{L}_{\text{gap}}$ from Theorem 2 matches with the value obtained experimentally from $P_{b,\text{test}}$, both occurring at $M_{\text{opt}} = 1408$ for 16-QAM as well as QPSK. The optimum M_{opt} suggested by $\mathcal{L}'_{\text{gap}}$ derived using the alternate cover set definition occurs at $M'_{\text{opt}} = 1152$, deviating only 5.3% from the actual lowest $P_{b,\text{test}}$ which occurs at $M = 1408$.

In practical standardized communication systems, link and rank adaptation mechanisms are employed based on CQI (Channel Quality Indicator) feedback from the receiver to the transmitter to maintain the reliability of the wireless link. Fig. 3 shows that $P_{b,\text{test}} < 10^{-1}$ for the concatenated model with the transmitter using QPSK modulation for the 'BERLIN_UMa_NLOS' channel scenario. Similarly, Fig. 4 shows that $P_{b,\text{test}} < 10^{-2}$ is achievable using QPSK modulation with the non-concatenated ESN model under the EPA channel scenario. These link and rank adaptation mechanisms, coupled with the use of strong channel coding schemes such as LDPC codes, ensure that the achieved link BER is well below 10% to comply with 3GPP standards[1] [42], [43] even with higher-order modulation schemes as shown in our previous work [32].

In terms of computational complexity, Fig. 3 shows that for the concatenated model, $M_{\text{opt}} \ll N_{\text{sc}}$, leading to its overall complexity being much lower than a conventional LMMSE detector based on our analysis from Sec. IV-F. On the contrary, Fig. 4 shows that for the non-concatenated ESN model where M_{opt} is on the order of N_{sc}, its overall complexity becomes $\mathcal{O}(N_{\text{sc}}^3)$ while that for the LMMSE detector is still $\mathcal{O}(N_{\text{sc}}^2)$. Note however, that this is still significantly lower than the computational complexity of other offline learning-based approaches such as *MMNet* [9]. Furthermore, even though RC-based architectures in our previous works [30], [31], [32] use state-input concatenation, i.e., 'skip' connections [34], our theoretical system design guidance agrees well with simulation for both models, with and without concatenation. This shows the practical impact of our theory-guided optimal design on general RC-based MIMO-OFDM symbol detectors.

[1]For example, the UE CQI (Channel Quality Indicator) calculation is based on a target coded BLER (Block Error Rate) of 10% [42], while the radio link monitoring out-of-sync BLER is also set to 10% [43].

VI. Conclusion

In this paper, we have derived a theoretical upper bound on the generalization error of Echo State Networks using tools from statistical learning theory. The derived bound is adapted suitably in terms of practical system parameters when the ESN is utilized as a symbol detector in a MIMO-OFDM wireless transceiver system. The monotonically decreasing trend of the training loss as a function of the reservoir size in neurons is theoretically justified. Combining the derived generalization error bound with the empirically characterized training loss, a systematic procedure is developed to design the optimum ESN-based symbol detector under given channel statistics, thereby avoiding the traditional practice in machine learning of hyper-parameter tuning via trial and error or grid search methods. We corroborate this procedure with experimental results obtained via simulations that employ realistic and standards-compliant wireless channel models. This provides valuable system design guidance and highlights the practical impact of our results.

Appendix A
Proof of Theorem 2

First, we verify the fact that $\mathcal{F}_t$ does indeed take values [-r, r]. To show this, consider

$$\|\mathbf{y}(t)\|_2 = \|\mathbf{W}_{\text{out}}\mathbf{x}_{\text{res}}(t)\|_2 \leq \|\mathbf{W}_{\text{out}}\|_2 \left\|\mathbf{x}_{\text{res}(t)}\right\|_2. \tag{29}$$

From Lemma 1, we know that $\|\mathbf{x}_{\text{res}}(t)\|_2 \leq \rho\left(B_{W_{\text{in}}}B_{X_{\text{in}}} + B_{W_{\text{fb}}}B_S\right)\frac{(\rho B_{W_{\text{res}}})^t - 1}{\rho B_{W_{\text{res}}} - 1}$. Therefore,

$$\|\mathbf{y}(t)\|_2 \leq \rho B_{W_{\text{out}}}\left(B_{W_{\text{in}}}B_{X_{\text{in}}} + B_{W_{\text{fb}}}B_S\right)\frac{(\rho B_{W_{\text{res}}})^t - 1}{\rho B_{W_{\text{res}}} - 1}, \tag{30}$$

i.e., $\|\mathbf{y}(t)\|_2 \leq \rho B B_{W_{\text{out}}} a_t = r$, where $B = \left(B_{W_{\text{in}}}B_{X_{\text{in}}} + B_{W_{\text{fb}}}B_S\right)$ and $a_t = \frac{(\rho B_{W_{\text{res}}})^t - 1}{\rho B_{W_{\text{res}}} - 1}$. To prove the main result, we use the upper bound derived on covering number for ESNs in Lemma 2. We first evaluate the integral in (20), i.e.,

$$\begin{aligned}
&\int_{\alpha}^{2r\sqrt{N}} \sqrt{\log \mathcal{N}(\mathcal{H}, \epsilon, \|\cdot\|)}d\epsilon \\
&\leq \int_{\alpha}^{2r\sqrt{N}} \sqrt{MK} \\
&\quad \times \sqrt{\log\left(\frac{2B_{W_{\text{out}}}\sqrt{J}\rho\left(B_{W_{\text{in}}}B_{X_{\text{in}}} + B_{W_{\text{fb}}}B_S\right)\frac{(\rho B_{W_{\text{res}}})^t - 1}{\rho B_{W_{\text{res}}} - 1}}{\epsilon}\right)}d\epsilon, \\
&\leq 2r\sqrt{N} \\
&\quad \times \sqrt{MK\log\left(2B_{W_{\text{out}}}\frac{\sqrt{J}\rho\left(B_{W_{\text{in}}}B_{X_{\text{in}}} + B_{W_{\text{fb}}}B_S\right)\frac{(\rho B_{W_{\text{res}}})^t - 1}{\rho B_{W_{\text{res}}} - 1}}{\alpha}\right)}.
\end{aligned} \tag{31}$$

Selecting $\alpha = \frac{1}{\sqrt{N}}$ and using the above result in (20) with the substitution for r, we arrive at (32), as shown at the top of the next page.

$$\mathcal{R}_N(\mathcal{H}) \leq \frac{4}{N} + \frac{24r}{\sqrt{N}}\sqrt{MK} \times \sqrt{\log\left(2B_{W_{\text{out}}}\sqrt{JN}\rho\left(B_{W_{\text{in}}}B_{X_{\text{in}}} + B_{W_{\text{fb}}}B_S\right)\frac{(\rho B_{W_{\text{res}}})^t - 1}{\rho B_{W_{\text{res}}} - 1}\right)}. \tag{32}$$

Substituting back with r, we arrive at the final form of Theorem 2, concluding its proof.

$$\mathcal{R}_N(\mathcal{F}_t) \leq \frac{4}{N} + \frac{24r}{\sqrt{N}}\sqrt{MK\log\left(2r\sqrt{JN}\right)}.\square$$

Appendix B
ERC Upper Bound for the Concatenated ESN Model

Here, we derive an upper bound on the ERC for the state-input concatenated ESN model (with skip connection). The difference between this model and the simpler non-concatenated ESN model is that the output is given by (22) in the former compared to (2). Therefore from (13) and since $\mathbf{z}(t) = [\mathbf{x}_{\text{res}}(t)^T, \mathbf{x}(t)^T]^T$, we have $\|\mathbf{z}(t)\|_2 = \sqrt{\|\mathbf{x}_{\text{res}}(t)\|_2^2 + \|\mathbf{x}(t)\|_2^2} \leq \sqrt{U_{res,t}^2 + B_{X_{in}}^2}$. Thus, we can rewrite the ERC bound in (32) as

$$\mathcal{R}_N(\mathcal{H}) \leq \frac{4}{N} + \frac{24r}{\sqrt{N}}\sqrt{MK} \times \sqrt{\log\left(2B_{W_{\text{out}}}\sqrt{JN}\sqrt{U_{res,t}^2 + B_{X_{in}}^2}\right)}. \tag{33}$$

The asymptotic approximation $\mathcal{L}'_{\text{gap}}$ or $\mathcal{L}_{\text{gap}}$ would therefore, be the same as (27) or (28) respectively, except that we would have $\beta = 24B_{W_{\text{out}}}\sqrt{\left(B_{W_{\text{in}}}B_{X_{\text{in}}}a_0\right)^2 + B_{X_{in}}^2}$.$\square$

Appendix C
ERC Upper Bound Using Alternate Cover Set Definition

From Lemma 4, we have $\log\mathcal{N}\big(\mathcal{F}_t, \epsilon, \text{dist}(\cdot,\cdot)\big) \leq \frac{r_F^2 K}{\epsilon^2}\log(2MK)$. As before, we first evaluate the integral in (20) by substituting the above, i.e.,

$$\begin{aligned}
&\int_{\alpha}^{2r\sqrt{N}} \sqrt{\log\mathcal{N}(\mathcal{F}_t, \epsilon, \text{dist}(\cdot,\cdot))}d\epsilon \\
&\quad\leq \int_{\alpha}^{2r\sqrt{N}} \sqrt{\frac{r_F^2 K}{\epsilon^2}\log(2MK)}d\epsilon \\
&\quad= r_F\sqrt{K\log(2MK)}\int_{\alpha}^{2r\sqrt{N}}\frac{1}{\epsilon}d\epsilon, \\
&\quad= r_F\sqrt{K\log(2MK)}\log\left(\frac{2r\sqrt{N}}{\alpha}\right).
\end{aligned}$$

Setting $\alpha = \frac{1}{\sqrt{N}}$, this becomes

$$\begin{aligned}
&\int_{\alpha}^{2r\sqrt{N}} \sqrt{\log\mathcal{N}(\mathcal{F}_t, \epsilon, \text{dist}(\cdot,\cdot))}d\epsilon \\
&\quad\leq r\sqrt{K\log(2MK)}\log(2rN).
\end{aligned} \tag{34}$$

We arrive at the alternate version of the ERC upper bound by substituting (34) in (20), i.e.,

$$\mathcal{R}'_N(\mathcal{F}_t) \leq \frac{4}{N} + \frac{12r_F\log(2rN)}{N}\sqrt{K\log(2MK)}. \tag{35}$$

Finally, the asymptotic generalization gap (as $N \to \infty$) based on the alternate cover set definition from [38] which can be used for optimum detector design is

$$\mathcal{L}'_{\text{gap}} = \beta' O\left(\frac{\log(2rN)}{2N}\sqrt{N_t\log(2MN_t)}\right), \tag{36}$$

where $\beta' = 24\rho B_{W_{in,F}}B_{X_{in},F}B_{W_{out,F}}a_0$. We know that $B_{W_{in,F}} \leq \sqrt{N_r}B_{W_{in}}$, $B_{X_{in,F}} \leq \sqrt{N}B_{X_{in}}$ and $B_{out,F} \leq \sqrt{M}$. Thus, for $N = N_p$ training "sequences" and since $r = O(\beta)$,

$$\mathcal{L}'_{\text{gap}} = \sqrt{N_r N_p M}\beta O\left(\frac{\log(\beta N_p)}{2N_p}\sqrt{N_t\log(2MN_t)}\right). \tag{37}$$

$\square$

References

[1] S. Jere, H. M. Saad, and L. Liu, "Error bound characterization for reservoir computing-based OFDM symbol detection," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2022, pp. 1349–1354.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[4] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 6645–6649.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)* 2015, San Diego, CA, USA, May 2015, pp. 1–15.

[6] R. Shafin, L. Liu, V. Chandrasekhar, H. Chen, J. Reed, and J. Zhang, "Artificial intelligence-enabled cellular networks: A critical path to beyond-5G and 6G," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 212–217, Apr. 2019.

[7] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1941–1988, 4th Quart., 2015.

[8] N. Samuel, T. Diskin, and A. Wiesel, "Learning to detect," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2554–2564, May 2019.

[9] M. Khani, M. Alizadeh, J. Hoydis, and P. Fleming, "Adaptive neural signal detection for massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5635–5648, Aug. 2020.

[10] K.-I. Funahashi and Y. Nakamura, "Approximation of dynamical systems by continuous time recurrent neural networks," *Neural Netw.*, vol. 6, no. 6, pp. 801–806, 1993.

[11] M. Lukoševičius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Comput. Sci. Rev.*, vol. 3, no. 3, pp. 127–149, 2009.

[12] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.

[13] S. Jere, R. Safavinejad, L. Zheng, and L. Liu, "Channel equalization through reservoir computing: A theoretical perspective," *IEEE Wireless Commun. Lett.*, early access, Jan. 6, 2023, doi: 10.1109/LWC.2023.3234239.

[14] V. Vapnik and A. J. Chervonenkis, "The necessary and sufficient conditions for consistency in the empirical risk minimization method," *Patt. Recog. Imag. Ana.*, vol. 1, no. 3, pp. 283–305, 1991.
[15] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," in *Proc. 14th Annu. Conf. Comput. Learn. Theory 5th Eur. Conf. Comput. Learn. Theory*, 2001, pp. 224–240.
[16] D. A. McAllester, "PAC-Bayesian model averaging," in *Proc. 12th Annu. Conf. Comput. Learn. Theory*, Jul. 1999, pp. 164–170.
[17] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Mach. Learn. Res.*, vol. 2, pp. 499–526, Mar. 2002.
[18] H. Xu and S. Mannor, "Robustness and generalization," *Mach. Learn.*, vol. 86, no. 3, pp. 391–423, 2012.
[19] B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro, "Exploring generalization in deep learning," in *Proc. Adv. Neural Inf. Proc. Sys.*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–10.
[20] H. Mhaskar, Q. Liao, and T. Poggio, "When and why are deep networks better than shallow ones?" in *Proc. 31st AAAI Conf. Art. Intell.*, 2017, pp. 2343–2349.
[21] M. Telgarsky, "Benefits of depth in neural networks," in *Proc. 29th Annu. Conf. Learn. Theory*, vol. 49, Jun. 2016, pp. 1517–1539.
[22] N. Cohen and A. Shashua, "Convolutional rectifier networks as generalized tensor decompositions," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 48, 2016, pp. 955–963.
[23] J. Zhang, Q. Lei, and I. Dhillon, "Stabilizing gradients for deep neural networks via efficient SVD parameterization," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5806–5814.
[24] M. Chen, X. Li, and T. Zhao, "On generalization bounds of a family of recurrent neural networks," in *Proc. 23rd Int. Conf. Artif. Intell. Statist.*, vol. 108, Aug. 2020, pp. 1233–1243.
[25] Z. Tu, F. He, and D. Tao, "Understanding generalization in recurrent neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–17.
[26] L. Gonon, L. Grigoryeva, and J.-P. Ortega, "Risk bounds for reservoir computing," *J. Mach. Learn. Res.*, vol. 21, no. 240, pp. 1–61, 2020.
[27] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "A model-driven deep learning network for MIMO detection," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2018, pp. 584–588.
[28] M. Goutay, F. Ait Aoudia, and J. Hoydis, "Deep hypernetwork-based MIMO detection," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, May 2020, pp. 1–5.
[29] S. S. Mosleh, L. Liu, C. Sahin, Y. R. Zheng, and Y. Yi, "Brain-inspired wireless communications: Where reservoir computing meets MIMO-OFDM," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4694–4708, Oct. 2018.
[30] Z. Zhou, L. Liu, and H.-H. Chang, "Learning for detection: MIMO-OFDM symbol detection through downlink pilots," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3712–3726, Jun. 2020.
[31] Z. Zhou, L. Liu, S. Jere, J. Zhang, and Y. Yi, "RCNet: Incorporating structural information into deep RNN for online MIMO-OFDM symbol detection with limited training," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3524–3537, Jun. 2021.
[32] J. Xu, Z. Zhou, L. Li, L. Zheng, and L. Liu, "RC-Struct: A structure-based neural network approach for MIMO-OFDM detection," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7181–7193, Sep. 2022.
[33] L. Li, L. Liu, J. Zhang, J. D. Ashdown, and Y. Yi, "Reservoir computing meets Wi-Fi in software radios: Neural network-based symbol detection using training sequences and pilots," in *Proc. 29th Wireless Opt. Commun. Conf. (WOCC)*, May 2020, pp. 1–6.
[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[35] M. Lukoševičius, *A Practical Guide to Applying Echo State Networks*. Berlin, Germany: Springer, 2012, pp. 659–686.
[36] B. Hajek and M. Raginsky, "Statistical learning theory," Univ. Illinois Urbana-Champaign, Tech. Rep., 2021. [Online]. Available: http://maxim.ece.illinois.edu/teaching/SLT/SLT.pdf.
[37] R. Safavinejad, H.-H. Chang, and L. Liu, "DRL meets DSA networks: Convergence analysis and its application to system design," *IEEE Trans. Wireless Commun.* [Online]. Available: https://arxiv.org/abs/2305.11237
[38] P. L. Bartlett, D. J. Foster, and M. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Proc. 31st Int. Conf. Neural Inf. Syst.*, 2017, pp. 6241–6250.
[39] L. Li, L. Liu, Z. Zhou, and Y. Yi, "Reservoir computing meets extreme learning machine in real-time MIMO-OFDM receive processing," *IEEE Trans. Commun.*, vol. 70, no. 5, pp. 3126–3140, May 2022.
[40] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antennas Propag.*, vol. 62, no. 6, pp. 3242–3256, Jun. 2014.
[41] *Evolved Universal Terrestrial Radio Access (E-UTRA) User Equipment (UE) Radio Transmission and Reception*, document TS 36.101, version 16.9.0, 3GPP, 2021.
[42] *5G; NR; Physical Layer Procedures for Data*, document TS 38.214, version 16.6.0, 3GPP, 2021.
[43] *5G; NR; Requirements for Support of Radio Resource Management*, document TS 38.133, version 17.2.0, 3GPP, 2021.

Shashank Jere received the B.S. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2014, and the M.S. degree in electrical engineering from the University of California at Los Angeles in 2016. He is currently pursuing the Ph.D. degree with the Wireless@VT, Bradley Department of Electrical and Computer Engineering, Virginia Tech. From 2016 to 2019, he worked as a Platform and Product Development Engineer with Qualcomm Technologies Inc., San Diego, CA, USA. His research interests include wireless communications, optimization, deep learning, and statistical learning theory.

Ramin Safavinejad received the B.S. degree in electrical engineering from the Iran University of Science and Technology in 2018 and the M.S. degree in electrical engineering from the Communication Systems Branch, Sharif University of Technology, Iran, in 2021. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the Wireless@VT, Virginia Tech. His research interests include physical and upper layers of wireless communications, deep learning, and theoretical analysis of algorithms in those areas.

Lingjia Liu (Senior Member, IEEE) received the B.S. degree in electronic engineering from Shanghai Jiao Tong University and the Ph.D. degree in electrical and computer engineering from Texas A&M University. He spent more than four years working with the Mitsubishi Electric Research Laboratory (MERL) and the Standards and Mobility Innovation Laboratory, Samsung Research America (SRA). He was leading Samsung's efforts on multiuser MIMO, CoMP, and HetNets in 3GPP LTE/LTE-advanced standards. He is currently a Professor and Bradley Senior Faculty Fellow with the ECE Department, Virginia Tech (VT). He is also working as the Director of the Wireless@VT. His general research interests include enabling technologies for 5G-advanced/6G networks, including machine learning for wireless networks, massive MIMO, massive MTC communications, and mmWave communications. He received the Global Samsung Best Paper Award from SRA in 2008 and 2010.