Universality and Approximation Bounds for Echo State Networks With Random Weights

Zhen Li and Yunfei Yang

ABSTRACT—We study the uniform approximation of echo state networks (ESNs) with randomly generated internal weights. These models, in which only the readout weights are optimized during training, have made empirical success in learning dynamical systems. Recent results showed that ESNs with ReLU activation are universal. In this article, we give an alternative construction and prove that the universality holds for general activation functions. Specifically, our main result shows that, under certain condition on the activation function, there exists a sampling procedure for the internal weights so that the ESN can approximate any continuous casual time-invariant operators with high probability. In particular, for ReLU activation, we give explicit construction for these sampling procedures. We also quantify the approximation error of the constructed ReLU ESNs for sufficiently regular operators.

INDEX TERMS—Approximation error, echo state networks (ESNs), reservoir computing, universality.

I. INTRODUCTION

RESERVOIR computing [1], such as echo state networks (ESNs) [2] and liquid state machines [3], is a paradigm for supervised learning of dynamical systems, which transforms input data into a high-dimensional space by a state-space nonlinear system called reservoir, and performs the learning task only on the readout. Due to the simplicity of this computational framework, it has been applied to many fields and made remarkable success in many tasks such as temporal pattern prediction, classification, and generation [4]. Motivated by these empirical successes, researchers have devoted a lot of effort in the theoretical understanding of the properties and performances of reservoir computing models (see, for instance, [1], [5], [6], [7], [8], [9], [10] and references therein).

In particular, recent studies [11], [12], [13], [14] showed that ESNs are universal in the sense that they can approximate sufficiently regular input-output systems (i.e., operators) in various settings. However, many of these universality results do not guarantee the important property of ESNs (and reservoir computing) that the state-space system is randomly generated, which is the major difference between reservoir computing and recurrent neural networks. To be concrete, consider the ESN

\[
\begin{align*}
    s_t &= \phi(W_1 s_{t-1} + W_2 u_t + b) \\
    y_t &= a \cdot s_t
\end{align*}
\]

where \(u_t, y_t, \text{and } s_t\) are the input-output and hidden state at time step \(t\), and \(\phi\) is a prescribed activation function. In general, the weight matrices \(W_1, W_2, \text{and } b\) are randomly generated and only the readout vector \(a\) is trained in a supervised learning manner. But the universality results in many papers, such as [11] and [14], require that all the weights depend on the target system that we want to approximate. Hence, these results cannot completely explain the approximation capacity of ESNs. To overcome this drawback of current theories, the recent work [9] studied the approximation of random neural networks in a Hilbert space setting and proposed a sampling procedure for the internal weights \(W_1, W_2, \text{and } b\) so that ESN (1) with ReLU activation is universal in \(L^2\) sense.

In this article, we generalize these results and study the universality of ESNs with randomly generated internal weights. Our main results show that, under weak assumption on the activation function \(\phi\), there exists a sampling procedure for \(W_1, W_2, \text{and } b\) so that the system (1) can approximate any continuous time-invariant operator with prescribed \(L^\infty\) error in high probability. In particular, for ReLU activation, we show that the ESN (1) is universal when \(W_1 = c_\mu W P W^T\) and \(W_2 = WQ\), where \((W, b)\) is a random matrix whose entries are sampled independently of a general symmetric distribution \(\mu\), \(c_\mu\) is a constant depending on \(\mu\) and \(P, Q\) are fixed matrices defined by (10).

Let us compare our results with the most related work [9]. Given a series of random samples from a uniform distribution, Gonon et al. [9] provided a procedure to use these samples to construct the internal weights so that the ReLU ESN is universal in high probability. We prove that such kind of sampling procedure also exists for general activation functions (it is sufficient that the activation does not grow too fast, see Corollary 10 and Assumption 12 for details). Different from the construction in [9], which takes advantage of the piecewise linearity of the ReLU function, our construction makes use of the concentration of probability measures. Hence, it does not rely on the piecewise linearity and can be applied to general activation functions. But our sampling procedure is given implicitly and it is difficult to compute it explicitly for general activation functions. We can only provide explicit examples for ReLU networks. One of the shortcoming of our
construction is that the constructed neural networks may not have the echo state property. But we show that these networks satisfy a property slightly weaker than the echo state property (see Remark 14), which, we think, is good enough for many applications.

A. Notations

We denote \( \mathbb{N} := \{1, 2, \ldots\} \) as the set of positive integers. Let \( \mathbb{Z} (\mathbb{Z}_+ \text{ and } \mathbb{Z}_-, \text{ respectively}) \) be the set of integers (nonnegative and nonpositive integers, respectively). The ReLU function is denoted by \( \sigma(x) := \max\{x, 0\} \). Throughout the article, \( \mathbb{R}^d \) is equipped with the sup-norm \( \| \cdot \|_\infty \), unless it is explicitly stated. For any \( U \subseteq \mathbb{R}^d \), we use \( U^Z \) to denote the set of sequences of the form \( u = (\ldots, u_{-1}, u_0, u_1, \ldots) \), \( u_i \in U, i \in \mathbb{Z} \). The set \( U^{Z+} \text{ and } U^{Z-} \) consist of right and left infinite sequences of the form \( (u_0, u_1, \ldots) \) and \( (\ldots, u_{-1}, u_0) \), respectively. For any \( i \leq j \) and \( u \in U^Z \), we denote \( u_{i:j} := (u_i, \ldots, u_j) \in U^{j-i+1} \) and use the convention that, when \( i = -\infty \) or \( j = \infty \), it denotes the left or right infinite sequence. And we will often regard \( u_{-\infty:j} \) as an element of \( U^{Z-} \). The supremum norm of the sequence is denoted by \( \| u \|_\infty := \sup_{t \in \mathbb{Z}} |u_t| \). A weighting sequence \( \eta : \mathbb{Z}_+ \rightarrow (0, 1] \) is a decreasing sequence with zero limit. The weight norm on \( (\mathbb{R}^d)^{Z-} \) associated with \( \eta \) is denoted by \( \| u \|_\eta := \sup_{t \in \mathbb{Z}} \eta_t |u_t| \|_\infty \).

II. CONTINUOUS CAUSAL TIME-INVARIENT OPERATORS

We study the uniform approximation problem for input-output systems or operators of signals in discrete time setting. We mainly consider the operators \( F : ([1, 1)^d]^Z \rightarrow \mathbb{R}^Z \) that satisfy the following three properties.

1) Causality: For any \( u, v \in ([1, 1)^d]^Z \) and \( t \in \mathbb{Z} \), \( u_{-\infty} = v_{-\infty} \) implies \( F(u)_t = F(v)_t \).
2) Time-Invariance: \( F(T_t(u)) = F(u)_{-t} \), for any \( \tau \in \mathbb{Z} \), where \( T_t : (\mathbb{R}^d)^Z \rightarrow (\mathbb{R}^d)^Z \) is time delay operator defined by \( T_t(u)_i = u_{i+t} \).
3) Continuity (Fading Memory Property): \( F : ([1, 1)^d]^Z \rightarrow [-B, B]^Z \) is uniformly bounded for some \( B > 0 \) and is continuous with respect to the product topology.

Grigoryeva and Ortega [11] gave a comprehensive study of these operators. We recall that the causal time-invariant operator \( F : ([1, 1)^d]^Z \rightarrow \mathbb{R}^Z \) is one-to-one correspondence with the functional \( F^* : ([1, 1)^d]^Z \rightarrow \mathbb{R} \) defined by \( F^*(u) := F(u)_0 \) where \( u \in ([1, 1)^d]^Z \) is any extension of \( u_- \in ([1, 1)^d]^Z \) such that \( u_{-\infty} = u_- \). \( F^* \) is well-defined by causality. And we can reconstruct \( F \) from \( F^* \) by \( F(u)_t = F^*(P_{Z-} \circ T_{-t}(u)) \), where \( P_{Z-} : (\mathbb{R}^d)^Z \rightarrow (\mathbb{R}^d)^Z \) is the natural projection. In particular, for any causal time-invariant operators \( F, G : ([1, 1)^d]^Z \rightarrow \mathbb{R}^Z \),

\[
\sup_{u \in ([1, 1)^d]^Z} \| F(u) - G(u) \|_\infty = \sup_{u_- \in ([1, 1)^d)^Z} |F^*(u_-) - G^*(u_-)|.
\]

Hence, the approximation problem of causal time-invariant operators \( F \) can be reduced to the approximation of functionals \( F^* \) on \( ([1, 1)^d]^Z \).

We remark that the product topology on \( (\mathbb{R}^d)^{Z-} \) is different from the uniform topology induced by the sup-norm. However, for the set \( ([1, 1)^d)^{Z-} \) of uniformly bounded sequences, the product topology coincides with the topology induced by the weight norm \( \| u \|_\eta \) of any weighting sequence \( \eta \). It is shown by [11, Sec. 2.3] that the causal time-invariant operator \( F^* \) is continuous (i.e., has the fading memory property) if and only if the corresponding functional \( F^* \) is continuous with respect to the product topology, which is equivalent to \( F^* \) has fading memory property with respect to some (and hence any) weighting sequence \( \eta \), i.e., \( F^* \) is a continuous function on the compact metric space \( (([1, 1)^d)]^{Z-}, \| \cdot \|_\eta \) \). In other words, for any \( \epsilon > 0 \), there exists a \( \delta(\epsilon) > 0 \) such that for any \( u, v \in ([1, 1)^d]^Z \),

\[
\| u - v \|_\eta = \sup_{t \in \mathbb{Z}} \eta_t |u_t - v_t| < \delta(\epsilon) \implies |F^*(u) - F^*(v)| < \epsilon.
\]

Next, we introduce several notations to quantify the regularity of causal time-invariant operators. For any \( m \in \mathbb{N} \), we can associate a function \( F^*_m : ([1, 1)^d]^m \rightarrow \mathbb{R} \) to any causal time-invariant operator \( F : ([1, 1)^d]^Z \rightarrow \mathbb{R}^Z \) by

\[
F^*_m(u_{m+1}, \ldots, u_0) : = F^*(\ldots, 0, u_{m+1}, \ldots, u_0).
\]

If the functional \( F^* \) can be approximated by \( F^*_m \) arbitrarily well, we say \( F \) has approximately finite memory. The following definition quantifies this approximation.

Definition 1 (Approximately Finite Memory): For any causal time-invariant operator \( F : ([1, 1)^d]^Z \rightarrow \mathbb{R}^Z \), let \( \epsilon \geq 0 \) and \( m \in \mathbb{N} \), we denote

\[
E_F(m) := \sup_{u \in ([1, 1)^d]^Z} |F^*(u) - F^*_m(u_{m+1:0})|,
\]

\[
m_F(\epsilon) := \min_{m \in \mathbb{N}} : E_F(m) \leq \epsilon.
\]

If \( m_F(\epsilon) < \infty \) for all \( \epsilon > 0 \), we say \( F \) has approximately finite memory. If \( m_F(0) < \infty \), we say \( F \) has finite memory.

Note that \( m_F(\epsilon) \) is a nonincreasing function of \( \epsilon \). By the time-invariance of \( F \), for any \( t \in \mathbb{Z} \)

\[
\sup_{u \in ([1, 1)^d]^Z} |F(u)_t - F^*_m(u_{t-m+1:1})| = \sup_{u \in ([1, 1)^d]^Z} |F^*(u_{-\infty}) - F^*_m(u_{t-m+1:1})| = E_F(m)
\]

Hence, \( E_F(m) \) quantifies how well \( F \) can be approximated by functionals with finite memory.

Definition 2 (Modulus of Continuity): Suppose the functional \( F^* : ([1, 1)^d)^{Z-} \rightarrow \mathbb{R} \) has fading memory property with respect to some weighting sequence \( \eta \). We denote the modulus of continuity with respect to \( \| \cdot \|_\eta \) by

\[
\omega_F(\delta; \eta) := \sup_{u, u' \in ([1, 1)^d)^{Z-}} |F^*(u) - F^*(u')| : u, u' \in ([1, 1)^d)^{Z-}, \quad \| u - u' \|_\eta \leq \delta
\]

and the inverse modulus of continuity

\[
\omega_F^{-1}(\epsilon; \eta) := \sup \{ \delta > 0 : \omega_F(\delta; \eta) \leq \epsilon \}.
\]
Similarly, the modulus of continuity (with respect to \( \| \cdot \|_\infty \)) of a continuous function \( f : [-1, 1]^m \to \mathbb{R} \) and the inverse modulus of continuity are defined by

\[
\omega_f(\delta) := \sup \{ |f(x) - f(x')| : x, x' \in [-1, 1]^m, \|x - x'\|_\infty \leq \delta \}
\]

\[
\omega_f^{-1}(\epsilon) := \sup \{ \delta > 0 : \omega_f(\delta) \leq \epsilon \}.
\]

The next proposition quantifies the continuity of causal time-invariant operators by the approximately finite memory and modulus of continuity. This proposition is a modification of similar result in [15] to our setting. It shows that a causal time-invariant operator \( F \) is continuous if and only if it has approximately finite memory and each \( F_m^* \) is continuous.

**Proposition 3:** Let \( F : ([{-1, 1}]^d)^Z \to \mathbb{R}^Z \) be a causal time-invariant operator.

1) If \( F \) has approximately finite memory and \( \omega_{F_m}(\delta) < \infty \) for any \( m \in \mathbb{N} \) and \( \delta > 0 \), then \( F^* \) has fading memory property with respect to any weighting sequence \( \eta \); for any \( \epsilon > 0 \), \( u, v \in ([{-1, 1}]^d)^Z \) and \( m = m_F(\epsilon/3) \)

\[
\|u - v\|_\infty \leq \eta_m - \omega_{F_m}(\epsilon/3) \quad \Longrightarrow \quad \|F^*(u) - F^*(v)\|_\infty \leq \epsilon.
\]

In other words, \( \omega_{F_m}(\eta_m - \omega_{F_m}(\epsilon/3); \eta) \leq \epsilon \) for \( m = m_F(\epsilon/3) \).

2) If \( F^* \) has fading memory property with respect to some weighting sequence \( \eta \), then \( F \) has approximately finite memory with \( \bar{E}_F(m) \leq \omega_F(\eta_m, \eta) \)

\[
m_F(\epsilon) \leq \min \{ m \in \mathbb{N} : \eta_m - \omega_{F_m}(\epsilon; \eta) \}
\]

and \( \omega_{F_m}(\delta) \leq \omega_F(\delta; \eta) \) for any \( m \in \mathbb{N} \).

In order to approximate the continuous causal time-invariant operator \( F \), we only need to approximate the functional \( F^* \), which can be approximated by the continuous function \( F_m^* \) if \( m \) is chosen sufficiently large. Hence, any approximation theory of continuous functions can be translated to an approximation result for continuous causal time-invariant operators.

For instance, if we approximate \( F_m^* \) by some function \( h \), then we can approximate \( F \) by

\[
y_t = h(u_{t-m+1}, \ldots, u_t), \quad t \in \mathbb{Z}.
\]

The function \( h \) uniquely determine a causal time-invariant operator \( H \) such that \( H(u)_t = h(u_{t-m+1}) \). Since \( H \) is time-invariant, \( H \) is continuous if and only if \( H \) is continuous by Proposition 3. When we approximate \( F_m^* \) by polynomials, then \( H \) is the Volterra series [16]. When \( h \) is a neural network, then \( H \) is a temporal convolution neural network, studied by [15].

In this article, we focus on the approximation by ESN, which are special state space models of the form

\[
\begin{align*}
s_t &= \phi(W_1 s_{t-1} + W_2 u_t + b) \\
y_t &= a \cdot s_t
\end{align*}
\]

where \( s_t, a, b \in \mathbb{R}^n \), \( W_1 \in \mathbb{R}^{n \times n} \), \( W_2 \in \mathbb{R}^{n \times d} \) and the activation function \( \phi : \mathbb{R} \to \mathbb{R} \) is applied element-wise.

**Definition 4 (Existence of Solutions and Echo State Property):** We say the system (4) has existence of solutions property if for any \( u \in ([{-1, 1}]^d)^Z \), there exist \( s \in ([{-1, 1}]^d)^Z \) such that \( s_t = \phi(W_1 s_{t-1} + W_2 u_t + b) \) holds for each \( t \in \mathbb{Z} \). If the solution \( s \) is unique, we say the system has echo state property.

Grigoryeva and Ortega [11, Th. 3.1] gave sufficient conditions for the system (4) to have existence of solutions property and echo state property. In particular, they showed that, if \( \phi \) is a bounded continuous function, then the existence of solutions property holds. And if \( \phi \) is a bounded Lipschitz continuous function with Lipschitz constant \( L_\phi < \infty \) and \( \|W_1\|L_\phi < 1 \), then the system (4) has echo state property, where \( \|W_1\| \) is the operator norm of the matrix \( W_1 \). As a sufficient condition to ensure the echo state property, the hypothesis \( \|W_1\|L_\phi < 1 \) has been extensively studied in the ESN literature [2], [5], [6], [7], [17].

If the system (4) has existence of solutions property, the axiom of choice allows us to assign \( s(u) \in ([{-1, 1}]^d)^Z \) to each \( u \in ([{-1, 1}]^d)^Z \), and hence define a functional \( h : ([{-1, 1}]^d)^Z \to \mathbb{R}^Z \) by \( h(u) := a \cdot s(u_0) = y_0 \). Thus, we can assign a causal time-invariant operator \( H : ([{-1, 1}]^d)^Z \to \mathbb{R}^Z \) to the system such that \( H^* = h \). When the echo state property holds, this operator is unique. The operator \( H \) is continuous if and only if the mapping \( u \mapsto s(u_0) \) is a continuous function. In Section III, we study the universality of these operators.

**III. Universal Approximation**

As mentioned in the introduction, the recent works of [11] and [14] showed that the ESNs (4) are universal: Assume \( \phi \) is a bounded Lipschitz continuous function. Let \( F : ([{-1, 1}]^d)^Z \to \mathbb{R}^Z \) be a causal continuous time-invariant operator, then for any \( \epsilon > 0 \), for sufficiently large \( n \), there exists an ESN (4) such that the corresponding causal time-invariant operator \( F_{ESN} \) satisfies

\[
\sup_{u \in ([{-1, 1}]^d)^Z} \|F(u) - F_{ESN}(u)\|_\infty \leq \epsilon.
\]

In this universal approximation theorem, the weights \( W_1, W_2, a, b \) in the network (4) depend on the target operator \( F \). However, in practice, the parameters \( W_1, W_2, b \) are drawn at random from certain given distribution and only the readout vector \( a \) are trained by linear regression using observed data related to the target operator \( F \). Hence, this universal approximation theorem cannot completely explain the empirical performance of ESNs.

In this section, our goal is to show that, with randomly generated weights, ESNs are universal: For any \( \epsilon > 0 \), for sufficiently large \( n \), with high probability on \( W_1, W_2, and b \), which are drawn from certain distribution, there exists \( a \) such that we can associate a causal time-invariant operator \( F_{ESN} \) to the ESN (4) and the approximation bound (5) holds. In the context of standard feed-forward neural networks, one can show that similar universal approximation theorem holds for random neural networks. This will be the building block of our main theorem of ESNs.

**A. Universality of Random Neural Networks**

It is well-known that feed-forward neural networks with one hidden layer are universal [18], [19], [20], [21]. We recall the
universal approximation theorem proved by [20], which has minimal requirements on the activation function.

**Theorem 5:** If $\phi : \mathbb{R} \to \mathbb{R}$ is continuous and is not a polynomial, then for any compact set $\mathcal{X} \subseteq \mathbb{R}^d$, any function $f \in C(\mathcal{X})$ and $\epsilon > 0$, there exists $a_i, b_i \in \mathbb{R}$ and $w_i \in \mathbb{R}^d$, $i = 1, \ldots, n$, such that

$$\sup_{x \in \mathcal{X}} \left| f(x) - \sum_{i=1}^n a_i \phi(w_i \cdot x + b_i) \right| \leq \epsilon. \quad (6)$$

In Theorem 5, the parameters $a_i, b_i$, and $w_i$ depend on the target function $f$. In order to take into account the fact that for ESN the inner weights $w_i, b_i$ are randomly chosen and only $a_i$ are trained from data, we consider the random neural networks whose weights $(w_i, b_i)$ are drawn from some probability distribution $\mu$. This motivates the following definition.

**Definition 6:** Suppose $(w_i^T, b_i)_{i \in \mathbb{N}}$ is a sequence of i.i.d. random vectors drawn from probability distribution $\mu$ defined on $\mathbb{R}^{d+1}$. If for any compact set $\mathcal{X} \subseteq \mathbb{R}^d$, any function $f \in C(\mathcal{X})$ and $\epsilon, \delta \in (0, 1)$, there exists $n \in \mathbb{N}$ such that, with probability at least $1 - \delta$, the inequality (6) holds for some $(a_i)_{1 \leq i \leq n}$, then we say the pair $(\phi, \mu)$ is universal.

**Remark 7:** We note that the readout weights $(a_i)_{1 \leq i \leq n}$ depend on the randomly generated weights $(w_i, b_i)_{1 \leq i \leq n}$. In the definition, we only require the existence of neural networks that converge to the target function in probability. This requirement is actually equivalent to almost sure convergence. To see this, convergence in probability implies that there exists a sequence $n_k$ such that there exists $f_{n_k}$, which is a linear combination of $\{\phi(w_i \cdot x + b_i)\}_{i=1}^{n_k}$ converging to the target function $f$ almost surely as $k \to \infty$. Notice that, for any $n \geq n_k$, $f_{n_k}$ is also a linear combination of $\{\phi(w_i \cdot x + b_i)\}_{i=1}^{n_k}$. Hence, there exists $f_n$, which is a linear combination of $\{\phi(w_i \cdot x + b_i)\}_{i=1}^n$, such that $f_n$ converges to $f$ almost surely.

The universality of random neural networks was widely studied in the context of extreme learning machine [22], [23], [24]. In particular, Huang et al. [22] used an incremental construction to establish the random universal approximation theorem in $L^2$-norm for bounded nonconstant piecewise continuous activation function. Gonon et al. [9] studied the approximation of random neural networks in a Hilbert space setting and established the universal approximation of ReLU neural networks. The recent work of [25] considered the approximation in $C^1$-norm and assumed that the activation function $\phi \in C^1(\mathbb{R})$ satisfying $0 < \int_{\mathbb{R}} |\phi'(x)| dx < \infty$. They argued that, since there exists a neural network $g$ that approximates the target function $f$ by universality of neural networks, there will eventually be some randomly generated samples $(w_i^T, b_i)$ that are close to the weights of $g$ and we can discard other samples by setting the corresponding $a_i = 0$, hence the random universal approximation holds. It is possible to generalize their argument to the approximation of continuous functions. Nevertheless, we will give an alternative approach based on law of large numbers and show that $(\phi, \mu)$ is universal under very weak conditions. Our analysis will need the following uniform law of large numbers.

**Lemma 8** [26, Th. 2]: Let $w_1, \ldots, w_n$ be i.i.d. samples from some probability distribution $\mu$ on $\Omega$. Suppose

1) $\mathcal{X}$ is compact;
2) $f(x, w)$ is continuous on $x$ for each $w \in \Omega$, and measurable on $w$ for each $x \in \mathcal{X}$;
3) there exists $g(w)$ with $|f(x, w)| \leq g(w)$ for all $x \in \mathcal{X}$ and $E_\mu[g(w)] < \infty$.

Then $E_\mu[f(x, w)]$ is continuous on $x$ and

$$\sup_{x \in \mathcal{X}} \left| \frac{1}{n} \sum_{i=1}^n f(x, w_i) - E_\mu[f(x, w)] \right| \to 0$$

$\mu$-almost surely as $n \to \infty$.

Now, we give a sufficient condition for the pair $(\phi, \mu)$ to be universal. Our proof is a combination of the uniform law of large number (Lemma 8) and the universality of neural networks (Theorem 5).

**Theorem 9:** Suppose the continuous function $\phi : \mathbb{R} \to \mathbb{R}$ is not a polynomial and $|\phi(x)| \leq C_1|x|^a + C_2$ for some $a, C_1, C_2 \geq 0$. If $\mu$ is a probability distribution with full support on $\mathbb{R}^{d+1}$ such that $\int \|\phi^T w\|_\infty d\mu(w) < \infty$, then $(\phi, \mu)$ is universal.

**Proof:** By Hahn–Banach theorem, for any compact set $\mathcal{X} \subseteq \mathbb{R}^d$, the linear span of a function class $\mathcal{H} \subseteq C(\mathcal{X})$ is dense in $C(\mathcal{X})$ if and only if

$$\left\{ v \in M(\mathcal{X}) : \int_{\mathcal{X}} h(x) d\nu(x) = 0, \forall h \in \mathcal{H} \right\} = \{0\}$$

where $M(\mathcal{X})$ is the dual space of $C(\mathcal{X})$, that is, the space of all signed Radon measures with finite total variation [27, Ch. 7.3].

We consider the linear space

$$\mathcal{H}_\mu := \left\{ h(x) = \int_{\mathbb{R}^{d+1}} g(w, b) \phi(w \cdot x + b) d\mu(w^T, b) : \right. \left. g \in L^\infty(\mu) \right\}. \quad (7)$$

Observe that $|g(w, b)\phi(w \cdot x + b)| \leq \|g\|_\infty (C_1 \|w \cdot x + b\|^a + C_2) \leq C_4 \|g\|_\infty (\|w^T\|_\infty + C_2) \|x\|^a$, where $C_4$ is a constant depending on the compact set $\mathcal{X}$. By assumption and Lemma 8, any $h \in \mathcal{H}_\mu$ is continuous and hence $\mathcal{H}_\mu \subseteq C(\mathcal{X})$. Suppose $v \in M(\mathcal{X})$ satisfies $\int_{\mathcal{X}} h(x) d\nu(x) = 0$ for all $h \in \mathcal{H}_\mu$. Then, by Fubini’s theorem

$$0 = \int_{\mathcal{X}} \int_{\mathbb{R}^{d+1}} g(w, b) \phi(w \cdot x + b) d\mu(w^T, b) d\nu(x) = \int_{\mathbb{R}^{d+1}} g(w, b) \int_{\mathcal{X}} \phi(w \cdot x + b) d\nu(x) d\mu(w^T, b)$$

for all function $g \in L^\infty(\mu)$. Therefore,

$$\int_{\mathcal{X}} \phi(w \cdot x + b) d\nu(x) = 0, \quad \mu\text{-almost surely.}$$

Since this function is continuous and $\mu$ has full support, the equality holds for all $(w^T, b) \in \mathbb{R}^{d+1}$ by Theorem 5, the linear span of $\{\phi(w \cdot x + b) : (w^T, b) \in \mathbb{R}^{d+1}\}$ is dense in $C(\mathcal{X})$, which implies $\nu = 0$. We conclude that $\mathcal{H}_\mu$ is dense in $C(\mathcal{X})$. In other words, for any $f \in C(\mathcal{X})$ and $\epsilon > 0$, there exists $h \in \mathcal{H}_\mu$ such that $|f(x) - h(x)| \leq \epsilon/2$ for all $x \in \mathcal{X}$.
By Lemma 8, for any $\delta > 0$, there exists $n \in \mathbb{N}$ such that, with probability at least $1 - \delta$ on the samples $(w_i, b_i)$ from $\mu$
\[
\sup_{x \in X} \left| \frac{1}{n} \sum_{i=1}^{n} g(w_i, b_i) \phi(w_i \cdot x + b_i) - h(x) \right| \leq \frac{\epsilon}{2}.
\]
By triangle inequality, we conclude that (6) holds with $a_i = (1/n)g(w_i, b_i)$ and the pair $(\phi, \mu)$ is universal.

In practice, it is more convenient to sample each weight independently of certain distribution $\mu$ on $\mathbb{R}$, so that $(w^T, b)$ is sample from $\mu^{d+1}$, the $d+1$ products of $\mu$. The next corollary is a direct application of Lemma 9 to this situation.

**Corollary 10**: Suppose the continuous function $\phi : \mathbb{R} \to \mathbb{R}$ is not a polynomial and $|\phi(x)| \leq C_1|x|^\alpha + C_2$ for some $\alpha, C_1, C_2 \geq 0$. If $\mu$ is a probability distribution with full support on $\mathbb{R}$ such that $\int |x|^\alpha d\mu(x) < \infty$, then for any $d \in \mathbb{N}$, the pair $(\phi, \mu^{d+1})$ is universal.

**Proof**: When $0 \leq \alpha \leq 1$
\[
\int \|(w^T, b)\|_2^\alpha d\mu^{d+1}(w^T, b)
\leq \int \|(w^T, b)\|_1^\alpha d\mu^{d+1}(w^T, b)
\leq \int (|b|^\alpha + \sum_{i=1}^{d} |w_i|^\alpha) d\mu^{d+1}(w^T, b) < \infty.
\]
When $\alpha > 1$, since all norms of $\mathbb{R}^{d+1}$ are equivalent
\[
\int \|(w^T, b)\|_\infty^\alpha d\mu^{d+1}(w^T, b)
\leq C \int \|(w^T, b)\|_1^\alpha d\mu^{d+1}(w^T, b)
= C \int (|b|^\alpha + \sum_{i=1}^{d} |w_i|^\alpha) d\mu^{d+1}(w^T, b) < \infty.
\]
In any cases, the pair $(\phi, \mu^{d+1})$ satisfies the condition in Lemma 9, hence it is universal.

So far, we have assumed that $\mu$ has full support. When the activation $\phi(x) = \sigma(x) = \max(x, 0)$ is the ReLU function, this assumption can be weaken due to the absolute homogeneity of ReLU.

**Corollary 11**: For the ReLU function $\sigma$, if $\mu$ is a probability distribution on $\mathbb{R}$ whose support contains the interval $[-r, r]$ for some $r > 0$, then $(\sigma, \mu^{d+1})$ is universal for any $d \in \mathbb{N}$.

**Proof**: We consider the continuous mapping $T : \mathbb{R}^{d+1} \to [-r, r]^{d+1}$ defined by
\[
T(w^T, b) = \begin{cases} (w^T, b) & \|(w^T, b)\|_\infty \leq r \\ r \frac{(w^T, b)}{\|(w^T, b)\|_\infty} & \|(w^T, b)\|_\infty > r. \end{cases}
\]
Let $\gamma = T_\#\mu^{d+1}$ be the push-forward measure of $\mu^{d+1}$ under $T$ defined by $\gamma(S) = \mu^{d+1}(T^{-1}(S))$ for any measurable set $S \subseteq [-r, r]^{d+1}$. Then, the support of $\gamma$ is $[-r, r]^{d+1}$ by assumption.

We first show that $(\sigma, \gamma)$ is universal. As in the proof of Theorem 9, if $v \in M(X)$ satisfies $\int_X h(x)dv(x) = 0$ for all $h \in H$, then by Fubini’s theorem
\[
\int_X \sigma(\tilde{\omega} \cdot x + \tilde{b})dv(x) = 0
\]
holds for all $(\tilde{\omega}, \tilde{b}) \in [-r, r]^{d+1}$. By the absolute homogeneity of $\sigma$, this equation actually holds for all $(\tilde{\omega}, \tilde{b}) \in \mathbb{R}^{d+1}$.

The argument in the proof of Theorem 9 implies that $(\sigma, \gamma)$ is universal.

Observe that any sample $(w^T, b)$ from $\mu^{d+1}$ corresponding to a sample $(\tilde{w}^T, \tilde{b}) = T(w^T, b)$ from $\gamma = T_\#\mu^{d+1}$, and
\[
\sum_{i=1}^{n} a_i\phi(w_i \cdot x + b_i) = \sum_{i=1}^{n} \frac{a_i}{c_i} \phi(\tilde{w}_i \cdot x + \tilde{b}_i)
\]
where $c_i = 1$ if $\|(w^T, b)\|_\infty \leq r$ and $c_i = \|(w^T, b)\|_\infty / r$ otherwise. We conclude that $(\sigma, \mu^{d+1})$ is universal.

**B. Universality of ESNs**

In this section, we will state and prove the random universal approximation theorem for ESNs. Our analysis is based on the uniform law of large numbers and the universality of random feed-forward neural networks. For simplicity, we will assume that all internal weights in the network have the same distribution $\mu$ from now on. We make the following assumption on the activation function $\phi : \mathbb{R} \to \mathbb{R}$ and the distribution $\mu$, where we also assume that the input $x \in [-2, 2]^m$ (any compact set slightly larger than $[-1, 1]^m$ is enough for our purpose).

**Assumption 12**: For any $m \in \mathbb{N}$, the pair $(\phi, \mu^{m+1})$ is universal and, for any $\epsilon_0 > 0$, there exists a measurable mapping $\varphi_m : \mathbb{R}^{m+1} \to \mathbb{R}^m$ such that
\[
\sup_{x \in [-2,2]^m} \|x - \mathbb{E}((w^T, b) \sim \mu^{m+1}) \varphi_m(w, b) \phi(w \cdot x + b)\|_\infty \leq \epsilon_0
\]
and $\|\varphi_m(w, b) \phi(w \cdot x + b)\|_\infty \leq g_m(w, b)$ for some $g_m$ satisfying $\mathbb{E}[g_m(w, b)] < \infty$.

Corollary 10 gives sufficient condition for the pair $(\phi, \mu^{m+1})$ to be universal. Besides, in the proof of Theorem 9, we have shown that the function class $H_\mu$ defined by (7) is dense in $C(X)$ for any compact set $X$. Therefore, if $\phi$ and $\mu$ satisfy Corollary 10 (so the pair $(\phi, \mu^{m+1})$ satisfies Theorem 9), then there exists a bounded mapping $\varphi_m$ that satisfies (8) by the denseness of $H_\mu$. In summary, any activation $\phi$ and distribution $\mu$ that satisfy the Conditions in corollary 10 also satisfy Assumption 12. However, the function $\varphi_m$ in (8) may be difficult to compute for general activation function.

We will give explicit construction for the ReLU function in Corollary 16.

By Lemma 8, Assumption 12 ensures that if we have $n$ i.i.d. samples $(w^T_1, b_1), \ldots, (w^T_n, b_n)$ from $\mu^{m+1}$, we can approximately reconstruct the input by
\[
x \approx \frac{1}{n} \sum_{i=1}^{n} \varphi_m(w_i, b_i) \phi(w_i \cdot x + b_i).
\]
In other words, the features $\phi(w_i \cdot x + b_i)$ contain enough information about the history of the input, and hence we can view it as a “reservoir” and approximate the desired function at any time steps. We make this idea precise in the next theorem.
Theorem 13: Suppose the activation \( \phi : \mathbb{R} \to \mathbb{R} \) and the distribution \( \mu \) satisfy Assumption 12. For \( n, m, d \in \mathbb{N} \), let \((w^1_1, b_1), \ldots, (w^d_n, b_n)\) be \( n \) i.i.d. samples from \( \mu^{md+1} \), and define the ESN

\[
\begin{aligned}
    s_t &= \phi(W \left( \frac{1}{n} \sum_{i=1}^{n} \varphi_{md}(w_i, b_i) s_{t-1} + Q u_t \right) + b) \\
    y_t &= a \cdot s_t
\end{aligned}
\]

where \( u_t \in \mathbb{R}^d, a, s_t \in \mathbb{R}^d, W = (w_1, \ldots, w_n)^T \in \mathbb{R}^{md} \), \( b = (b_1, \ldots, b_n)^T \in \mathbb{R}^d \) and we denote \( \varphi_{md}(W, b) := (\varphi_{md}(w_1, b_1), \ldots, \varphi_{md}(w_n, b_n)) \in \mathbb{R}^{md \times n} \) for measurable function \( \varphi_{md} : \mathbb{R}^{md+1} \to \mathbb{R}^{md} \) and

\[
P = \begin{pmatrix}
    0_d & I_d \\
    \vdots & \vdots \\
    0_d & I_d
\end{pmatrix} \in \mathbb{R}^{md \times md}, \quad Q = \begin{pmatrix}
    0_d \\
    \vdots \\
    0_d
\end{pmatrix} \in \mathbb{R}^{md \times d}
\]

where \( 0_d \) and \( I_d \) are the \( d \times d \) zero matrix and identity matrix, respectively. Then, for any continuous causal time-invariant operator \( F : ([-1, 1]^d)^Z \to \mathbb{R}^Z \), any \( \epsilon, \delta \in (0, 1) \), there exist \( n, m \in \mathbb{N} \) and measurable function \( \varphi_{md} \) such that, with probability at least \( 1 - \delta \), the system (9) has existence of solutions property and there exists \( a \in \mathbb{R}^d \) so that the corresponding operator \( F_{ESN} \) satisfies

\[
    \sup_{u \in ([-1, 1]^d)^Z} \| F(u) - F_{ESN}(u) \|_\infty \leq \epsilon. \tag{11}
\]

Proof: The proof is divided into three steps. Step 1: Choose parameters \((W, b)\).

We choose \( m = m_F(\epsilon/3) \), then by the definition of \( m_F(\epsilon/3) \) and \( F_m^* \), for any \( u \in \mathbb{R} \), \( t \in \mathbb{Z} \)

\[
    \| F(u) - F_m^*(u_{t-1}) \|_\infty \leq \epsilon/3.
\]

Since \( F_m^* : [-1, 1]^md \to \mathbb{R} \) is continuous by Proposition 3, we can continuously extend its domain to \([-2, 2]^md \) by setting \( F(x_1, \ldots, x_{md}) = F(\tilde{x}_1, \ldots, \tilde{x}_{md}) \) where \( \tilde{x}_i = \max[\min[x_i, 1], -1] \). By Assumption 12, the pair \((\varphi, \mu^{md+1})\) is universal. Hence, there exists \( N_1 \in \mathbb{N} \) such that for any \( n \geq N_1 \), with probability at least \( 1 - \frac{\delta}{2} \)

\[
    \sup_{x \in [-2, 2]^md} \left| F_m^*(x) - \sum_{i=1}^n a_i \varphi(w_i \cdot x + b_i) \right| \leq \epsilon/3 \tag{12}
\]

for some \( a = (a_1, \ldots, a_n)^T \in \mathbb{R}^n \).

By Assumption 12, there exists \( \varphi_{md} \) such that

\[
    \sup_{x \in [-2, 2]^md} \| x - E[\varphi_{md}(w, b) \varphi(w \cdot x + b)] \|_\infty \leq \frac{1}{2m} \omega_F^{-1}(\epsilon/3).
\]

Lemma 8 implies that

\[
    \sup_{x \in [-2, 2]^md} \left| E[\varphi_{md}(w, b) \varphi(w \cdot x + b)] - \frac{1}{n} \sum_{i=1}^n \varphi_{md}(w_i, b_i) \varphi(w_i \cdot x + b_i) \right| \to 0
\]

almost surely as \( n \to \infty \). Thus, there exists \( N_2 \in \mathbb{N} \) such that for any \( n \geq N_2 \), with probability at least \( 1 - \delta/2 \)

\[
    \sup_{x \in [-2, 2]^md} \left| x - \frac{1}{n} \sum_{i=1}^n \varphi_{md}(w_i, b_i) \varphi(w_i \cdot x + b_i) \right|_\infty \leq \frac{1}{m} \omega_F^{-1}(\epsilon/3)
\]

which can be written as

\[
    \sup_{x \in [-2, 2]^md} \left| x - \frac{1}{n} \sum_{i=1}^n \varphi_{md}(W, b) \varphi(wx + b) \right|_\infty \leq \frac{1}{m} \omega_F^{-1}(\epsilon/3). \tag{13}
\]

Without loss of generality, we can assume that \( \omega_F^{-1}(\epsilon/3) \leq 1/2 \) by setting \( \epsilon \) sufficiently small. Now, we choose \( n = \max[N_1, N_2] \), then with probability at least \( 1 - \delta \), (12) and (13) hold. We will fix such parameters \((W, b)\) that satisfy (12) and (13).

Step 2: Establish the existence of solutions property.

For any fixed \( u \in ((-1, 1]^d)^Z \), we consider the operator

\[
    \psi : (\mathbb{R}^md)^Z \to (\mathbb{R}^md)^Z \text{ defined by}
\]

\[
    \psi(r)_i := \frac{1}{n} P \varphi_{md}(W, b) \varphi(Wr_{t-1} + b) + Qu_i \in \mathbb{R}^md
\]

For convenience, we will regard any vector in \( \mathbb{R}^md \) as \( m \) vectors in \( \mathbb{R}^d \). To do this, for any \( i = 1, \ldots, m \), we denote the \((i - 1)d + 1 \) to \( id \) entries of \( \psi(r) \), by \( \psi(r)_i \in \mathbb{R}^d \). Then, by the definition of matrices \( P \) and \( Q \), \( \psi(r)_i[m] = u_i \) and \( \psi(r)_i[i] = ((1/n)\varphi_{md}(W, b) \varphi(Wr_{t-1} + b))[i + 1] \) for \( 1 \leq i \leq m - 1 \).

For \( j = 1, \ldots, m \), we define the \( j \) composition of \( \psi \) inductively by \( \psi^{(j)} = \psi \circ \psi^{(j-1)} \), where \( \psi^{(0)} \) is the identity map. We are going to show that \( \psi^{(m)} \) maps the compact set \( \{(-3/2, 3/2)^{md}\}^{Z} \) into itself and then use fixed point theorem to establish the existence of solutions property. To do this, we assert that, if \( r \in \{(-3/2, 3/2)^{md}\}^{Z} \), then, for \( 1 \leq i \leq m - j \)

\[
    \| \psi^{(j)}(r)[i] \|_\infty \leq \frac{3}{2} + \frac{1}{m} \omega_F^{-1}(\epsilon/3)
\]

and for \( m - j < k \leq m \)

\[
    \| \psi^{(j)}(r)[k] - u_{t-1+k} \|_\infty \leq \frac{m - k}{m} \omega_F^{-1}(\epsilon/3).
\]

We prove this by induction. When \( j = 1 \), \( \psi(r)_i[m] = u_i \in [-1, 1]^d \) and, by inequality (13), for \( 1 \leq i \leq m - 1 \)

\[
    \| \psi(r)_i[i] \|_\infty \leq \| r_{t-1}[i + 1] \|_\infty + \frac{1}{m} \omega_F^{-1}(\epsilon/3) \leq \frac{3}{2} + \frac{1}{m} \omega_F^{-1}(\epsilon/3).
\]

Hence, the assertion is true for \( j = 1 \). Assume that the assertion is true for some \( 1 \leq j \leq m - 1 \), we are going to show that it is true for \( j + 1 \). Since \( \psi^{(j)}(r)_i[-2, 2]^{md} \), by inequality (13), for \( 1 \leq i \leq m - 1 \)

\[
    \| \psi^{(j+1)}(r)_i[i] - \psi^{(j)}(r)_{i-1}[i + 1] \|_\infty \leq \frac{1}{m} \omega_F^{-1}(\epsilon/3)
\]
and $\psi^{(j+1)}(r_i)[m] = u_i$. Hence, by induction assumption, for $1 \leq i \leq m - j - 1$

$$\|\psi^{(j+1)}(r_i)[i]\|_{\infty} \leq \|\psi^{(j)}(r_i)[i+1]\|_{\infty} + \frac{1}{m} \omega_{F_n}^{-1}(\epsilon/3) \leq \frac{3}{2} + \frac{j + 1}{m} \omega_{F_n}^{-1}(\epsilon/3)$$

and for $m - j - 1 < k \leq m - 1$

$$\|\psi^{(j+1)}(r_i)[k] - u_{i-m+k}\|_{\infty} \leq \|\psi^{(j)}(r_i)-[k+1] - u_{i-1-m+k+1}\|_{\infty} + \frac{1}{m} \omega_{F_n}^{-1}(\epsilon/3) \leq \frac{m}{m} \omega_{F_n}^{-1}(\epsilon/3) + \frac{1}{m} \omega_{F_n}^{-1}(\epsilon/3) = \frac{m - k}{m} \omega_{F_n}^{-1}(\epsilon/3).$$

Hence, the assertion is true and we conclude that

$$\|\psi^{(m)}(r_i) - u_{i-m+1}\|_{\infty} \leq \omega_{F_n}^{-1}(\epsilon/3) \leq \frac{1}{2}$$

holds for $r \in ((-3/2), (3/2)^{md})^Z$, where we regard $u_{i-m+1} \in ((-1, 1)^m)^Z$ as the vector $(u_1^t, \ldots, u_m^t) \in \mathbb{R}^{md}$ to simplify the notation. This inequality implies that $\psi^{(m)}$ maps the compact convex set $((-3/2), (3/2)^{md})^Z$ into itself. By the continuity of $\psi^{(m)}$ and Schauder’s Fixed Point Theorem (see [28, Th. 7.1]), $\psi^{(m)}$ has at least a fixed point, that is, a point $r^* \in ((-3/2), (3/2)^{md})^Z$ such that $\psi^{(m)}(r^*) = r^*$. Note that the fixed point $r^*$ depends on $u$ because the operator $\psi$ is defined through $u$.

Now, we are ready to establish the existence of solution property for the system (9). For any $u \in ((-1, 1)^m)^Z$, we define

$$r = \psi^{(j)}(r^*)$$

where $j = t - [(t/m)] \in [0, 1, \ldots, m - 1]$. If $j = 0$, then $r = r^*$, $r_{t-1} = \psi^{(m)}(r^*)$ and

$$\psi(r) = \frac{1}{n} P_{\varphi_{md}}(W, b)\phi(Wr_{t-1} + b) + Qu_t$$

$$= \frac{1}{n} P_{\varphi_{md}}(W, b)\phi(W \psi^{(m-1)}(r^*), t-1 + b) + Qu_t$$

$$= \psi^{(m)}(r^*)$$

If $j = 1$, then $r = \psi^{(j)}(r^*)$, $r_{t-1} = \psi^{(j-1)}(r^*)$ and

$$\psi(r) = \frac{1}{n} P_{\varphi_{md}}(W, b)\phi(Wr_{t-1} + b) + Qu_t$$

$$= \frac{1}{n} P_{\varphi_{md}}(W, b)\phi(W \psi^{(j-1)}(r^*), t-1 + b) + Qu_t$$

$$= \psi^{(j-1)}(r^*)$$

In any cases, we have $\psi(r) = r$, or equivalently

$$r = \frac{1}{n} P_{\varphi_{md}}(W, b)\phi(Wr_{t-1} + b) + Qu_t.$$  

Since $r$ is a fixed point of $\psi$, it is also a fixed point of $\psi^{(m)}$ and hence

$$\|r - u_{t-m+1}\|_{\infty} = \|\psi^{(m)}(r_i) - u_{i-m+1}\|_{\infty} \leq \omega_{F_n}^{-1}(\epsilon/3)$$

where we regard $u_{i-m+1}$ as a vector of $\mathbb{R}^{md}$ as before.

Let $s_t := \phi(Wr_t + b)$, then $s$ satisfies the equation

$s_t = \phi \left( W \left( \frac{1}{n} P_{\varphi_{md}}(W, b)s_{t-1} + Qu_t + b \right) \right)$.

Hence, the system (9) has the existence of solution property with probability at least $1 - \delta$.

Step 3: Bound the approximation error.

For the chosen parameters $(W, b)$, we let $a = (a_1, \ldots, a_m)^T \in \mathbb{R}^n$ be the vector that satisfies inequality (12), then

$$\sup_{x \in [(-2, 2)^m]} |F_m^a(x) - a \cdot \phi(Wx + b)| \leq \epsilon/3.$$

Since the system (9) has the existence of solution property, we can associate a causal time-invariant operator $F_{ESN}$ to the system (9) such that

$$F_{ESN}(u_t) = a \cdot s_t = a \cdot \phi(Wr_t + b), \quad u \in ((-1, 1)^m)^Z.$$ 

Therefore, for any $u \in ((-1, 1)^m)^Z$, we have $|F_{ESN}(u_t) - F(u_t)| \leq \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon$

which completes the proof.

Theorem 13 establishes the random universal approximation for a special form of ESNs. But we are only able to prove the existence of solutions property for the constructed network. We note that the echo state property is established for certain ReLU neural networks constructed in [9] and [14]. It would be interesting to see whether this property also holds for the construction in Theorem 13 (it seems that some regularity assumption on the activation function is needed). In the following remark, we show that the networks constructed in the proof of Theorem 13 satisfy a property slightly weaker than the echo state property.

Remark 14: Recall that the echo state property can be formulated in a forward version [7, Th. 2.1]: there exists a sequence $\delta_{j, 0, \infty}$ such that $\delta_j \to 0$ as $t \to \infty$ and, for any state sequences $s_{0, \infty}$ and $s_{0, \infty}'$, generated by the system (9) with the same input $u_{1, \infty}$ and different initial states $s_0$ and $s_0'$, respectively, it holds $\|s_t - s_t'\|_{\infty} \leq \delta_t$. Thus, the effect of the initial state vanishes as $t \to \infty$. In the proof of Theorem 13, we have shown that $s_t = \phi(Wr_t + b)$ with $|r_{t-1} - u_{i-m+1}\|_{\infty} \leq \omega_{F_n}^{-1}(\epsilon/3)$, which can be smaller than any prescribed accuracy $\epsilon_0$ by setting $\epsilon$ small enough. Hence, for any fixed sample $(W, b)$, the difference of the states $|s_t - s_t'|_{\infty}$ is bounded for large $t$, by the continuity of the activation $\phi$. For example, if $\phi$ is Lipschitz continuity with Lipschitz constant $L$, then $\|s_t - s_t'\|_{\infty} \leq \|\phi(Wr_t + b) - \phi(Wr_t + b)\|_{\infty} \leq 2L\|W\|\epsilon_0$ by the triangle inequality, where $\|W\|$ denotes the operator norm induced by $\|\cdot\|_{\infty}$. Furthermore, the output sequences $y_t = a \cdot s_t$ and $y_t' = a' \cdot s_t'$ computed by the corresponding operators constructed in Theorem 13 satisfy the same approximation accuracy (11) for large $t$, which means the outputs are closed.
to each other. So, in our construction, the effect of the initial state is small.

Let us take a closer look at the result of Theorem 13. For a given pair of activation and distribution \((\phi, \mu)\), it is shown that if the internal weights are sampled as in (9), then the ESN is a universal approximator with high probability. Notice that the sampling procedure in (9) depends on the function \(\varphi_{md}\), which is implicitly given in Assumption 12. As mentioned above, if the pair \((\phi, \mu)\) satisfies Corollary 10, we are guaranteed the existence of \(\varphi_{md}\). However, it may be difficult to compute it explicitly. In the following, we show how to compute \(\varphi_{md}\) for the widely used ReLU activation \(\phi = \sigma\), when the distribution \(\mu\) is symmetric.

**Proposition 15:** Let \(\mu\) be a symmetric distribution on \(\mathbb{R}\) with variance \(\int x^2 d\mu(x) = M_2 < \infty\). Then, for the ReLU function \(\sigma\) and any \(m \in \mathbb{N}\)

\[
\mathbb{E}_{(w^1, b^1) \sim \mu^m}[w \sigma(w \cdot x + b)] = \frac{M_2}{2} x.
\]

**Proof:** Let us denote \(w = (w_1, \ldots, w_m)^T\) and \(x = (x_1, \ldots, x_m)^T\). We need to show that \(2\mathbb{E}[w_j \sigma(w \cdot x + b)] = x_j M_2\) for any \(j = 1, \ldots, m\). By symmetry

\[
\mathbb{E}[w_j \sigma(w \cdot x + b)] = -\mathbb{E}[w_j \sigma(-w \cdot x - b)].
\]

Using the equality \(\sigma(x) - \sigma(-x) = x\) for the ReLU function, we have

\[
2\mathbb{E}[w_j \sigma(w \cdot x + b)] = \mathbb{E}[w_j \sigma(w \cdot x + b)] - \mathbb{E}[w_j \sigma(-w \cdot x - b)] = \mathbb{E}[w_j w_j \sigma(x_1 + \ldots + w_m x_m + b)] = x_j \mathbb{E}[w_j^2] = x_j M_2
\]

where we use the fact that \(\mathbb{E}[w_j x_i] = \mathbb{E}[w_j w_i x_i] = 0\) for \(i \neq j\), because the random variables \(w_i\) and \(w_j\) are independent and symmetric.

By Proposition 15, for the ReLU function \(\sigma\) and symmetric distribution \(\mu\), we can choose \(\varphi_{md}(w, b) = (2/M_2)w\) in Assumption 12 with error \(e_0 = 0\). As a corollary, we can obtain the random universal approximation theorem for ReLU ESNs under very weak assumption on the distribution \(\mu\).

**Corollary 16:** Let \(\mu\) be a symmetric probability distribution on \(\mathbb{R}\) whose support contains the interval \([-r, r]\) for some \(r > 0\), and variance \(\int x^2 d\mu(x) = M_2 < \infty\). For \(n, m, d \in \mathbb{N}\), let \((w_1^m, b_1), \ldots, (w_n^m, b_n)\) be \(n\) i.i.d. samples from \(\mu^{md+1}\), and define the ReLU ESN

\[
\begin{align*}
\sigma = & \sigma \left( \left( \frac{2}{n M_2} P W \sigma_{s_{t-1}} + Q u_t \right) + b \right) \\
\gamma_i = & a \cdot s_i \\
\end{align*}
\]

where \(u_t \in \mathbb{R}^d\), \(a, s_i \in \mathbb{R}^n\), \(W = (w_1, \ldots, w_n)^T \in \mathbb{R}^{n \times md}\), \(b = (b_1, \ldots, b_n)^T \in \mathbb{R}^n\) and \(P, Q\) are defined by (10).

Then, for any continuous causal time-invariant operator \(F: ([1, 1]^d)^Z \to \mathbb{R}^Z\), any \(\epsilon, \delta \in (0, 1)\), there exist \(m, n \in \mathbb{N}\) such that, with probability at least \(1 - \delta\), the system (14) has existence of solutions property and there exists \(\alpha \in \mathbb{R}\) so that the corresponding operator \(F_{ESN}\) satisfies

\[
\sup_{u \in ([1, 1]^d)^Z} \| F(u) - F_{ESN}(u) \| \leq \epsilon.
\]

**Proof:** For any \(m \in \mathbb{N}\), the pair \((\sigma, \mu^{md+1})\) is universal by Corollary 11 and satisfies Assumption 12 with \(\varphi_{md}(w, b) = (2/M_2)w\) by Proposition 15. The result then follows from Theorem 13 by observing that \(\varphi_{md}(W, b) = ((2/M_2)w_1, \ldots, (2/M_2)w_n) = (2/M_2)WT\).

IV. APPROXIMATION BOUNDS FOR RELU NETWORKS

We have shown that the ReLU ESN (14) with randomly generated weights can approximate any continuous causal time-invariant operator with high probability. In this section, we derive approximation bounds for sufficiently regular functions. To simplify the analysis, we will assume that each random weight in (14) is sampled from the uniform distribution on \([-1/2, 1/2]\). Similar results can be derived for distributions \(\mu\) with bounded support. To do that, one can modify Assumption 17 by replacing the expectation (15) over uniform distribution by the expectation over \(\mu\). Then, the proofs of Lemma 19 and Theorem 20 can be applied.

The key ingredients of our approximation results are certain integral representation of sufficiently regular functions and the random neural networks approximation of these functions. We make the following regularity assumption on the target continuous causal time-invariant operator \(F\).

**Assumption 17:** The function \(F_m: [-1, 1]^md \to \mathbb{R}\) defined by (2) can be represented as

\[
F_m(x) = \int_{-1}^{1} \int_{[-1, 1]^{md}} g_m(w, b) \sigma(w \cdot x + b) dwdb
\]

for some function \(g_m\) with uniform bound \(|g_m(w, b)| \leq B_m\).

Neural network approximation of functions with integral representations was initially studied by the classical work of Barron [29]. This result has been generalized and widely discussed in recent articles [9], [30], [31], [32], [33]. In particular, Gonon et al. [9] and Klusowski and Barron [30] showed that sufficient regular functions (whose Fourier transforms decay sufficiently fast) can be represented by integral forms similar to (15). The following proposition is a modification of these results. The proof is deferred to the Appendix.

**Proposition 18:** Suppose \(f: \mathbb{R}^d \to \mathbb{R}\) admits a Fourier representation

\[
f(x) = \int_{\mathbb{R}^d} \hat{f}(w) e^{i w x} dw, \quad x \in [-1, 1]^d
\]

and \(\|f\|_{L^1} \leq B\). Then, there exists \(g: [-1/2, 1/2]^{d+1} \to \mathbb{R}\) with \(|g(w, b)| \leq 40B\) such that

\[
f(x) = \int_{-1/2}^{1/2} \int_{[-1, 1]^{d}} g(w, b) \sigma(w \cdot x + b) dwdb,
\]

\(x \in [-1, 1]^d\).

The next lemma gives error bound for the random approximation of the integral form (15). This result can be seen as a quantitative version of Lemma 8. Intuitively, the integral can be regarded as an expectation over uniform distribution. Hence, the law of large numbers implies that the empirical average converges to the expectation. We can bind this approximation error by using statistical learning theory [34] and concentration inequalities [35]. Specifically, we view the
uniform approximation error as the supremum of difference between the expectation \( f(x) \) and its empirical average. Then, we can apply concentration inequality to bound the deviation of this supremum. The expectation of the supremum is controlled by the Rademacher complexity [36]. The proof is given in the Appendix.

**Lemma 19:** Let
\[
f(x) = \int_{[-1,1]^d} g(w)\sigma(w \cdot x) dw
\]
with \( |g(w)| \leq B \). Then, with probability at least \( 1 - \delta \)
\[
\sup_{x \in [-1,1]^d} \left| f(x) - \frac{1}{n} \sum_{i=1}^{n} g(w_i)\sigma(w_i \cdot x) \right| 
\leq Br d \left( 2 \sqrt{\frac{2d \log(n + 1)}{n}} + \frac{\sqrt{\log(2/\delta)}}{2n} \right)
\]
where \( w_{1:n} := (w_i)_{1 \leq i \leq n} \) are i.i.d. samples from the uniform distribution on \([-1/2, 1/2]^d\).

Now, we are ready to prove our approximation bound for ReLU ESNs. The proof is similar to the proof of Theorem 13, in which we approximate the activation function \( F^* \) and construct an ESN to approximate \( F^* \). Here, we quantify this approximation error by Lemma 19.

**Theorem 20:** For \( n, m, d \in \mathbb{N} \), let \((w^T_1, b_1), \ldots, (w^T_n, b_n)\) be \( n \) i.i.d. samples from the uniform distribution on \([-1/2, 1/2]^{md+1}\), and define the ReLU ESN

\[
\begin{align*}
  s_t &= \sigma \left( W \left( \frac{24}{n} PW^T s_{t-1} + Qu_t \right) + b \right) \\
  y_t &= a \cdot s_t
\end{align*}
\]

where \( u_t \in \mathbb{R}^d \), \( a, s_t \in \mathbb{R}^n \), \( W = (w_1, \ldots, w_n)^T \in \mathbb{R}^{n \times md} \), \( b = (b_1, \ldots, b_n)^T \in \mathbb{R}^n \) and \( P, Q \) are defined by (10). Then, for sufficiently large \( n/\log(n + 1) \geq c(m, \delta, d) \) and any continuous causal time-invariant operator \( F : (\mathbb{R}^d)^\mathbb{T} \rightarrow \mathbb{R}^d \) satisfying Assumption 17, with probability at least \( 1 - \delta \), the system has existence of solutions property and there exists \( a \in \mathbb{R}^d \) so that the corresponding operator \( F_{\text{ESN}} \) satisfies

\[
\sup_{u \in ([-1,1]^d)^\mathbb{T}} \| F(u) - F_{\text{ESN}}(u) \|_{\infty} 
\leq C(m, \delta, d) \sqrt{\frac{\log(n + 1)}{n}} + E_F(m)
\]

where we can choose \( C(m, \delta, d) = (2)^{1/2} B_m (md + 1)^{(3m^2d + 1)(\log(4md/\delta))^{1/2}} \) and \( c(m, \delta, d) = (2)^{1/2} B_m (md + 1)^{(3m^2d + 1)(\log(4md/\delta))^{1/2}} \).

**Proof:** For any \( t \in \mathbb{Z} \), by (3)

\[
\sup_{u \in ([-1,1]^d)^\mathbb{T}} \left| F(u)_t - F^*_{m}(u_{t-m+1:1}) \right| \leq E_F(m).
\]

By Assumption 17, we can extend the domain of \( F^*_m \) to \([-2, 2]^{md}\) by the integral representation (15). Using Lemma 19, it holds with probability at least \( 1 - \delta/2 \) that

\[
\sup_{x \in [-2,2]^{md}} \left| F^*_m(x) - \frac{1}{n} \sum_{i=1}^{n} g(w_i, b_i)\sigma(w_i \cdot x + b_i) \right| 
\leq \sqrt{2B_m(md + 1)} \left( 4 \sqrt{\frac{(md + 1) \log(n + 1)}{n}} + \frac{\sqrt{\log(4/\delta)}}{n} \right) 
=: \mathcal{E}_1
\]

On the other hand, Proposition 15 shows that \( x = 24Ew\sigma(w \cdot x + b) \). Hence, applying Lemma 19 to each coordinate of \( x \), it holds with probability at least \( 1 - \delta/2 \) that

\[
\sup_{x \in [-2,2]^{md}} \left\| x - \frac{24}{n} \sum_{i=1}^{n} w_i\sigma(w_i \cdot x + b_i) \right\|_{\infty} 
\leq 12\sqrt{2(md + 1)} \times \left( 4 \sqrt{\frac{(md + 1) \log(n + 1)}{n}} + \frac{\sqrt{\log(4md/\delta)}}{n} \right) 
=: \mathcal{E}_2
\]

where the last inequality holds if \( n/\log(n + 1) \geq 1152m^2(md + 1)^2(4(md + 1)^{1/2} + (\log(4md/\delta)))^{1/2} \). We note that, in matrix form

\[
x - \frac{24}{n} \sum_{i=1}^{n} w_i\sigma(w_i \cdot x + b_i) = x - \frac{24}{n} W^T \sigma(Wx + b)
\]

Therefore, with probability at least \( 1 - \delta \) over parameters \((W, b)\), inequalities (16) and (17) hold. For such parameters \((W, b)\), Step 2 of the proof of Theorem 13 shows that, for any \( u \in ([-1,1]^d)^\mathbb{T} \), there exists \( \tilde{\mathbf{F}} \in ([-2, 2]^{md})^\mathbb{T} \) such that

\[
\tilde{\mathbf{F}}_t = \frac{24}{n} PW^T \sigma(W\tilde{F}_{t-1} + b) + Qu_t
\]

and

\[
\| \tilde{\mathbf{F}}_t - u_{t-m+1:1} \|_{\infty} \leq m \mathcal{E}_2
\]

where we regard \( u_{t-m+1:1} \) as a vector of \( \mathbb{R}^{md} \) in the natural way. By the integral representation (15), we have

\[
F^*_m(\tilde{\mathbf{F}}_t) - F^*_m(u_{t-m+1:1}) 
\leq B_m \int_{[-1/2, 1/2]^d} \| w \|_1 \| \tilde{\mathbf{F}}_t - u_{t-m+1:1} \|_d dw \leq \frac{B_m m^2 d}{4} \mathcal{E}_2.
\]

Let \( s_t := \sigma(W \tilde{F}_t + b) \), then \( s_t \) satisfies the equation

\[
s_t = \sigma \left( W \left( \frac{24}{n} PW^T s_{t-1} + Qu_t \right) + b \right)
\]

Hence, the system has the existence of solution property with probability at least \( 1 - \delta \) and we can associate a causal time invariant operator \( F_{\text{ESN}} \) to the system such that

\[
F_{\text{ESN}}(u)_t = a \cdot s_t = a \cdot \sigma(W \tilde{F}_t + b), \quad u \in ([-1,1]^d)^\mathbb{T}.
\]

Let \( a = (1/n)(g(w_1, b_1), \ldots, g(w_n, b_n))^T \), then by inequality (16)

\[
|F_{\text{ESN}}(u)_t - F(u)_t| = |a \cdot \sigma(W \tilde{F}_t + b) - F(u)_t| 
\leq |a \cdot \sigma(W \tilde{F}_t + b) - F^*_m(\tilde{\mathbf{F}}_t)| + |F^*_m(\tilde{\mathbf{F}}_t) - F^*_m(u_{t-m+1:1})| 
+ |F^*_m(u_{t-m+1:1}) - F(u)_t| 
\leq \mathcal{E}_1 + \frac{B_m m^2 d}{4} \mathcal{E}_2 + E_F(m).
\]
\[
\leq \sqrt{2}B_m(md + 1) \left( 12m^2d + 4 \right) \sqrt{\frac{m(d + 1)\log(n + 1)}{n}}
+ \sqrt{\frac{\log(4/\delta)}{n}}
+ \frac{3m^2d}{n} \sqrt{\frac{\log(4md/\delta)}{n}} + \mathcal{E}_F(m)
\leq C(m, \delta, d) \sqrt{\frac{\log(n + 1)}{n}} + \mathcal{E}_F(m)
\]

where \( C(m, \delta, d) = (2)^{1/2}B_m(md + 1)(3m^2d + 1)(4(md + 1)^{1/2} + (\log(4md/\delta))^{1/2}) \).

**Remark 21:** Similar approximation bound has been obtained in [9]. They made assumptions that the input sequences are sampled from some probability distribution and the target operator is Lipschitz continuous, and proved a bound for the expected \( L^2 \) approximation error. In contrast, we generalize the Lipschitz continuity to general modulus of continuity, and obtain a high probability bound for the universal approximation error in Theorem 20. Note that, if \( B_m \) is at most polynomial on \( m \) and \( \mathcal{E}_F(m) \) decays exponentially, say \( \mathcal{E}_F(m) \leq C_0 e^{-c_1m} \) for some \( C_0, c_1 > 0 \), then we can choose \( m \approx (2c_1)^{-1} \log n \) and the approximation error is \( \tilde{O}(\log(1/\delta)/n)^{1/2} \), where the notation \( \tilde{O} \) hides polylogarithmic factors of \( n \). In this case, the approximation rate is the same as [9, Th. 2].

**V. Conclusion**

We have analyzed the approximation capacity of ESNs with randomly generated internal weights. Under weak assumptions on the activation function, we constructed a sampling procedure for the internal weights so that the ESNs can approximate any continuous time-invariant operators with prescribed accuracy. For ReLU activation, we explicitly computed the random weights in our construction and derived the approximation bounds for sufficiently regular operators.

Our results also raise some questions for future study. For example, our construction of the sampling procedure relies on the function \( \varphi_m \) in Assumption 12. In other words, we only establish the universality of ESNs when the parameters are sampled in a special way. In practical applications, the parameters in ESNs are directly sampled from some simple distributions, and they do not have the particular form constructed in this article. It would be interesting to extend the universality result to these “practical” sampling procedures. Another interesting question is the echo state property of the constructed networks. In Theorem 13 and Remark 14, we only prove that the constructed networks have existence of solution property and satisfy a property weaker than the echo state property. In order to establish the echo state property, we think more regularity condition on the activation function is needed.

**APPENDIX**

**A. Proof of Proposition 3**

For the first part, by the definition of \( m = m_F(\epsilon/3) \)

\[
\sup_{u \in \{-1, 1\}^{12}} |F^*(u) - F^*_m(u_{m+1:0})| \leq \epsilon/3.
\]

Since the weighting sequence is decreasing

\[
\|u_{m+1:0} - v_{m+1:0}\|_\infty \leq \eta_{m+1} \|u - v\| \leq \omega_{F^*_m}(\epsilon/3)
\]

which implies \( |F^*_m(u_{m+1:0}) - F^*_m(v_{m+1:0})| \leq \epsilon/3. \)

Therefore,

\[
|F^*(u) - F^*(v)|
\leq |F^*_m(u_{m+1:0}) - F^*_m(v_{m+1:0})| + |F^*_m(u_{m+1:0}) - F^*_m(v_{m+1:0})|
\leq \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon.
\]

For the second part, we observe that, for any \( m \in \mathbb{N} \) and \( u \in \{-1, 1\}^{12} \)

\[
\|u - (\ldots, 0, u_{m+1:0})\|_\infty = \sup_{i \leq m} \eta_i \|u_i\|_\infty \leq \eta_m.
\]

Then, by the definition of \( \omega_{F^*}(\delta; \eta) \)

\[
\mathcal{E}_F(m) = \sup_{u \in \{-1, 1\}^{12}} |F^*(u) - F^*(\ldots, 0, u_{m+1:0})|
\leq \omega_{F^*}(\eta_m, \eta) \leq \epsilon.
\]

Finally, by the definition of \( \omega_{F^*}(\delta; \eta) \),

\[
\mathcal{E}_F(m) \leq \|u_{m+1:0} - u'_{m+1:0}\|_\infty \leq \omega_{F^*}(\delta; \eta).
\]

**B. Proof of Proposition 18**

The proof is a modification of the argument in [9] and [30]. First, observe that, for any \( t \in \mathbb{R} \)

\[
-\int_0^\infty \sigma(t - b)e^{ib} + \sigma(-t - b)e^{-ib} db = e^{it} - it - 1.
\]

For any \( x \in [-1, 1]^d \), letting \( t = \mathbf{w} \cdot x \), we have

\[
f(x) = \nabla f(0) \cdot x - f(0)
= \int_{\mathbb{R}^d} (e^{i\mathbf{w} \cdot x} - i\mathbf{w} \cdot x - 1) \hat{f}(\mathbf{w}) d\mathbf{w}
= -\int_{\mathbb{R}^d} \int_0^\infty (\sigma(\mathbf{w} \cdot x - b)e^{ib} + \sigma(-\mathbf{w} \cdot x - b)e^{-ib})
\times \hat{f}(\mathbf{w}) db d\mathbf{w}
= -\int_{\mathbb{R}^d} \int_0^\infty \sigma(\mathbf{w} \cdot x + b)e^{-ib} \hat{f}(\mathbf{w}) db d\mathbf{w}
- \int_{\mathbb{R}^d} \int_{-\infty}^0 \sigma(\mathbf{w} \cdot x + b)e^{ib} \hat{f}(-\mathbf{w}) db d\mathbf{w}
= \int_{\mathbb{R}^d} \int_{|\mathbf{w}|}^0 \text{Re} [e^{-ib} \hat{f}(\mathbf{w}) - e^{ib} \hat{f}(-\mathbf{w})]
\times \sigma(\mathbf{w} \cdot x + b) db d\mathbf{w}
\]

where, in the last equality, we use the fact that \( \sigma(\mathbf{w} \cdot x + b) = 0 \) for \( b < -\|\mathbf{w}\|_1 \). Since \( f(0), \nabla f(0) \cdot x \in \mathbb{R} \), we have
\[
\int_{\mathbb{R}^d} \text{Im}[\hat{f}(w)]dw = 0 \quad \text{and} \quad \int_{\mathbb{R}^d} (w \cdot x) \text{Re}[\hat{f}(w)]dw = 0,
\]
which implies
\[
\nabla f(0) \cdot x + f(0) = \int_{\mathbb{R}^d} -(w \cdot x) \text{Im}[\hat{f}(w)] + \text{Re}[\hat{f}(w)]dw
\]
\[
= \int_{\mathbb{R}^d} \int_0^{1/2} (w \cdot x + b)(8 \text{Re}[\hat{f}(w)] - \text{Im}[\hat{f}(w)])dbdw
\]
\[
= \int_{\mathbb{R}^d} \int_0^{1/2} (\sigma(w \cdot x + b) - \sigma(-w \cdot x - b)) \times \text{Re}(8 + i)\hat{f}(w)dbdw
\]
\[
= \int_{\mathbb{R}^d} \int_0^{1/2} \text{Re}[(8 + i)\hat{f}(w)]\sigma(w \cdot x + b)dbdw
\]
\[
- \int_{\mathbb{R}^d} \int_{-1/2}^0 \text{Re}[(8 + i)\hat{f}(w)]\sigma(w \cdot x + b)dbdw.
\]
By Fubini's theorem, we conclude that
\[
f(x) = \int_{\mathbb{R}^d} h(w, b)\sigma(w \cdot x + b)dbdw
\]
with
\[
h(w, b) := 1_{|w|_1} \text{Re} [-e^{-ib}\hat{f}(w) - e^{ib}\hat{f}(-w)] + 1_{[0,1/2]}(b) \text{Re}(8 + i)\hat{f}(w)
\]
\[
- 1_{[-1/2,0]}(b) \text{Re}(8 + i)\hat{f}(-w)
\]
where \(1_S\) denote the indicator function of the set \(S\). Note that, if \(|b| > \|w\|_1\) and \(|b| > 1/2\), then \(h(w, b) = 0\).

Next, we denote \(A = \{w \in \mathbb{R}^d : \|w\|_1 > 1/2\}\), \(B = \{w \in \mathbb{R}^d : 0 < \|w\|_1 < 1/2\}\) and consider the mapping
\[
T : B \rightarrow A, \quad T(w) = \frac{w}{4\|w\|_1^2}.
\]
Then, the Jacobian of \(T\) is
\[
\nabla T(w) = \frac{1}{4\|w\|_1^2} \left( I_d - 2 \frac{w}{\|w\|_1} \text{sgn}(w)^T \right)
\]
with determinant
\[
|\text{det(\nabla T(w))}| = \frac{1}{(2\|w\|_1)^{2d}} \left| 1 - 2 \text{sgn}(w)^T \frac{w}{\|w\|_1} \right|
\]
where we use Sylvester’s determinant theorem. The change of variables formula implies
\[
\int_{\mathbb{R}^d} \int_A h(w, b)\sigma(w \cdot x + b)dbdw
\]
\[
= \int_{\mathbb{R}^d} \int_B h\left(\frac{w}{4\|w\|_1^2}, b\right)\sigma\left(\frac{w \cdot x}{4\|w\|_1} + b\right) \frac{1}{(2\|w\|_1)^{2d}} dbdw
\]
\[
= \int_{\mathbb{R}^d} \int_B h\left(\frac{w}{4\|w\|_1^2}, \frac{b}{4\|w\|_1^2}\right)\sigma(w \cdot x + b) \times \frac{1}{(2\|w\|_1)^{2d+4}} dbdw
\]
where we use the homogeneity of ReLU in the last equality. Hence,
\[
f(x) = \int_{\mathbb{R}^d} \int_B h(w, b)\sigma(w \cdot x + b)dbdw
\]
\[
+ \int_{\mathbb{R}^d} \int_A h(w, b)\sigma(w \cdot x + b)dbdw
\]
\[
= \int_{\mathbb{R}^d} \int_B \left( h(w, b) + \frac{1}{(2\|w\|_1)^{2d+4}} h\left(\frac{w}{4\|w\|_1^2}, \frac{b}{4\|w\|_1^2}\right) \right) \sigma(w \cdot x + b) dbdw.
\]
We define
\[
g(w, b) := 1_{g(w)}(h(w, b) + \frac{1}{(2\|w\|_1)^{2d+4}} h\left(\frac{w}{4\|w\|_1^2}, \frac{b}{4\|w\|_1^2}\right)).
\]
Then, \(g(w, b) = 0\) for \(\|w\|_1 > 1/2\). If \(\|w\|_1 \leq 1/2\) and \(|b| > 1/2\), then \(|b| > \|w\|_1\) and \(|b|/(4\|w\|_1^2) > 1/2\), and hence \(h(w, b) = h((w/4\|w\|_1^2), (b/4\|w\|_1^2)) = 0\), which implies \(g(w, b) = 0\). This shows that
\[
f(x) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{[-1/2,1/2]^d} g(w, b)\sigma(w \cdot x + b)dbdw, \quad x \in [-1, 1]^d.
\]
Finally, for \(0 < \|w\|_1 \leq 1/2\), by the definition of \(h(w, b)\) and the assumption that
\[
\sup_{r > 0} \sup_{\|w\|_1 = 1/2} \max(1, r^{2d+4}) |\hat{f}(rw)| \leq B
\]
we have \(|h(w, b)| \leq 10(1|\hat{f}(w)| + |\hat{f}(-w)|) \leq 20B\) and
\[
\frac{1}{(2\|w\|_1)^{2d+4}} h\left(\frac{w}{4\|w\|_1^2}, \frac{b}{4\|w\|_1^2}\right) \leq 10r^{2d+4}(1|\hat{f}(rw)| + |\hat{f}(-rw)|) \leq 20B
\]
where \(\hat{w} = w/(2\|w\|_1)\) and \(r = 1/(2\|w\|_1)\). Therefore, \(|g(w, b)| \leq 40B\).

\section{Proof of Lemma 19}
We denote the function \(h(x, w) := g(w)\sigma(w \cdot x)\) and the random variables
\[
Z := \sup_{x \in [-r, r]^d} f(x) - \frac{1}{n} \sum_{i=1}^n h(x, w_i)
\]
\[
Z_+ := \sup_{x \in [-r, r]^d} f(x) - \frac{1}{n} \sum_{i=1}^n h(x, w_i)
\]
\[
Z_- := \sup_{x \in [-r, r]^d} -\frac{1}{n} \sum_{i=1}^n h(x, w_i) - f(x).
\]
We are going to bound the expectation of these random variables by the Rademacher complexity [36]. Let \(w_{1,\sigma} = (w_{1,\sigma})_1 \leq s\) be i.i.d. samples from the uniform distribution on \([-1/2, 1/2]^d\), independent of \(w_{1,\sigma}\). Since \(E_{w_{1,\sigma}} h(x, w_{1,\sigma}) = f(x)\)
\[
E_{w_{1,\sigma}} Z_+ = \sum_{x \in [-r, r]^d} \frac{1}{n} \sum_{i=1}^n h(x, w_i) - \frac{1}{n} \sum_{i=1}^n h(x, w_i)
\]
\[
\leq E_{w_{1,\sigma}, w_{1,\sigma}} \left[ \sup_{x \in [-r, r]^d} \frac{1}{n} \sum_{i=1}^n h(x, w_i) - h(x, w_i) \right].
\]
Let \( \xi_{1:n} = (\xi_j)_{j \leq 2n} \) be a sequence of i.i.d Rademacher random variables, independent of \( w_{1:n} \) and \( w'_{1:n} \). Then, by symmetrization argument, we have
\[
\mathbb{E}_{w_{1:n}, Z_+} \leq \mathbb{E}_{w_{1:n}, w'_{1:n}, \xi_{1:n}} \left[ \sup_{x \in [-r, r]^d} \frac{1}{n} \sum_{i=1}^{n} h(x, w'_i) - h(x, w_i) \right] \\
= \mathbb{E}_{w_{1:n}, w'_{1:n}, \xi_{1:n}} \left[ \sup_{x \in [-r, r]^d} \frac{1}{n} \sum_{i=1}^{n} \xi_i (h(x, w'_i) - h(x, w_i)) \right] \\
\leq \mathbb{E}_{w_{1:n}, w'_{1:n}, \xi_{1:n}} \left[ \sup_{x \in [-r, r]^d} \frac{1}{n} \sum_{i=1}^{n} \xi_i h(x, w'_i) \right] \\
+ \sup_{x \in [-r, r]^d} \frac{1}{n} \sum_{i=1}^{n} \xi_i h(x, w_i) \\
= 2 \mathbb{E}_{w_{1:n}, \xi_{1:n}} \left[ \sup_{x \in [-r, r]^d} \frac{1}{n} \sum_{i=1}^{n} \xi_i h(x, w'_i) \right]
\]
where the last equality is because \( w_i \) and \( w'_i \) have the same distribution and \( \xi_i \) and \( -\xi_i \) have the same distribution.

Observe that, for any \( w \in [-1/2, 1/2]^d \) and \( x, y \in [-r, r]^d \), we have \( |h(x, w)| \leq B \|x - y\| \leq B r \sqrt{d} \) and
\[
|h(x, w) - h(y, w)| = |g(w) \sigma (w \cdot x) - g(w) \sigma (w \cdot y)| \\
\leq B \|w \cdot (x - y)\| \leq \frac{1}{2} B d \sqrt{\|x - y\|_{\infty}}.
\]

For any \( \epsilon > 0 \), there exists a set \( S_{\epsilon} = \{x_j\}_{j=1}^{N} \subseteq [-r, r]^d \) of \( N \leq (1 + r/\epsilon)^d \) points such that \( S_{\epsilon} \) is an \( \epsilon \)-covering of \( [-r, r]^d \); for any \( x \), there exists \( x_j \in S_{\epsilon} \) such that \( \|x - x_j\|_{\infty} \leq \epsilon \). Hence,
\[
\mathbb{E}_{w_{1:n}, Z_+} \leq 2 \mathbb{E}_{w_{1:n}, \xi_{1:n}} \left[ \sup_{x \in S_{\epsilon}} \frac{1}{n} \sum_{i=1}^{n} \xi_i h(x_j, w_i) \right] \\
+ \sup_{\|x - x_j\|_{\infty} \leq \epsilon} \frac{1}{n} \sum_{i=1}^{n} \xi_i (h(x, w_i) - h(x_j, w_i)) \\
\leq 2 \mathbb{E}_{w_{1:n}, \xi_{1:n}} \left[ \sup_{x \in S_{\epsilon}} \frac{1}{n} \sum_{i=1}^{n} \xi_i h(x_j, w'_i) \right] \leq B d \epsilon.
\]

By Massart’s lemma [34, Lemma 26.8]
\[
\mathbb{E}_{\xi_{1:n}} \left[ \sup_{x \in S_{\epsilon}} \frac{1}{n} \sum_{i=1}^{n} \xi_i h(x_j, w'_i) \right] \\
\leq \text{Brd} \sqrt{\frac{2 \log N}{2 \sqrt{n}}} \leq \text{Brd} \sqrt{\frac{2 \log (1 + r/\epsilon)}{n}}.
\]

If we choose \( \epsilon = r/\sqrt{n} \), then
\[
\mathbb{E}_{w_{1:n}, Z_+} \leq \text{Brd} \sqrt{\frac{2 \log (1 + \sqrt{n})}{n}} + \text{Brd} \sqrt{\frac{1}{n}} \\
\leq 2 \text{Brd} \sqrt{\frac{2 \log (n + 1)}{n}}.
\]

Next, we apply McDiarmid’s inequality [35, Th. 6.2] to the random variable \( Z_+ \). Observe that, if one of the \( w_j \) is replaced by \( w'_j \), then the difference is
\[
\sup_{x \in [-r, r]^d} \left[ f(x) - \frac{1}{n} \sum_{i=1}^{n} h(x, w'_i) \right] - \sup_{x \in [-r, r]^d} \left[ f(x) - \frac{1}{n} \sum_{i=1}^{n} h(x, w_i) + \frac{1}{n} h(x, w'_i) - \frac{1}{n} h(x, w_i) \right] \\
\leq \frac{1}{n} \sup_{x \in [-r, r]^d} \left| h(x, w_i) - h(x, w'_i) \right| \leq \text{Brd} \frac{1}{n}
\]

because the inequalities hold for each fixed \( x \in [-r, r]^d \) and taking supremum cannot increase the difference. Hence, McDiarmid’s inequality implies
\[
\mathbb{P}(Z_+ - \text{E}Z_+ > t) \leq \exp\left(-\frac{2n^2 t^2}{B^2 \epsilon^2 d^2}\right).
\]

For any \( \delta \in (0, 1) \), if we choose \( t = \text{Brd} (\log (2/\delta)/(2n))^{1/2} \), then with probability at least \( 1 - \delta/2 \)
\[
Z_+ \leq \text{E}Z_+ + \frac{t}{2} \leq \text{Brd} \left( 2 \sqrt{\frac{2 \log (n + 1)}{n}} + \sqrt{\frac{\log (2/\delta)}{2n}} \right) =: \mathcal{E}.
\]

Applying similar argument to \( Z_- \), shows that \( Z_- \leq \mathcal{E} \) with probability at least \( 1 - \delta/2 \). Hence,
\[
\mathbb{P}(Z_+ > \mathcal{E}) \leq \mathbb{P}(Z_+ > \mathcal{E}) + \mathbb{P}(Z_- > \mathcal{E}) \leq \delta.
\]

REFERENCES

[1] M. LUKOŠEVIČIUS and H. Jaeger, “Reservoir computing approaches to recurrent neural network training,” Comput. Sci. Rev., vol. 3, no. 3, pp. 127–149, Aug. 2009.

[2] H. Jaeger and H. Haas, “Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication,” Science, vol. 304, no. 5667, pp. 78–80, Apr. 2004.

[3] W. Maass, T. Natschlager, and H. Markram, “Real-time computing without stable states: A new framework for neural computation based on perturbations,” Neural Comput., vol. 14, no. 11, pp. 2531–2560, Nov. 2002.

[4] G. Tanaka et al., “Recent advances in physical reservoir computing: A review,” Neural Netw., vol. 115, pp. 100–123, Jul. 2019.

[5] H. Jaeger, “The ‘echo state’ approach to analysing and training recurrent neural networks—with an erratum note,” German Nat. Res. Center Inf. Technol., Bonn, Germany, GMD Tech. Rep. 148, 2001.

[6] M. Buehner and P. Young, “A tighter bound for the echo state property,” IEEE Trans. Neural Netw., vol. 17, no. 3, pp. 820–824, May 2006.

[7] I. B. Yildiz, H. Jaeger, and S. J. Kielb, “Re-visiting the echo state property,” Neural Netw., vol. 35, pp. 1–9, Nov. 2012.

[8] L. Grigoryeva and J.-P. Ortega, “Differentiable reservoir computing,” J. Mach. Learn. Res., vol. 20, no. 179, pp. 1–62, 2019.

[9] L. Gonon, L. Grigoryeva, and J.-P. Ortega, “Approximation bounds for random neural networks and reservoir systems,” Ann. Appl. Probab., vol. 33, no. 1, pp. 28–69, Feb. 2023.

[10] 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.

[11] L. Grigoryeva and J.-P. Ortega, “Echo state networks are universal,” Neural Netw., vol. 108, pp. 495–508, Dec. 2018.

[12] L. Grigoryeva and J.-P. Ortega, “Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneous state-affine systems,” J. Mach. Learn. Res., vol. 19, no. 24, pp. 1–40, 2018.

[13] L. Gonon and J.-P. Ortega, “Reservoir computing universality with stochastic inputs,” IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 1, pp. 100–112, Jan. 2020.

[14] L. Gonon and J.-P. Ortega, “Fading memory echo state networks are universal,” Neural Netw., vol. 138, pp. 10–13, Jun. 2021.
[15] J. Hanson and M. Raginsky, “Universal approximation of input–output maps by temporal convolutional nets,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32. Red Hook, NY, USA: Curran Associates, 2019, pp. 14071–14081.

[16] S. Boyd and L. Chua, “Fading memory and the problem of approximating nonlinear operators with Volterra series,” IEEE Trans. Circuits Syst., vol. CS-32, no. 11, pp. 1150–1161, Nov. 1985.

[17] G. Manjunath and H. Jaeger, “Echo state property linked to an input: Exploring a fundamental characteristic of recurrent neural networks,” Neural Comput., vol. 25, no. 3, pp. 671–696, Mar. 2013.

[18] G. Cybenko, “Approximation by superpositions of a sigmoidal function,” Math. Control, Signals, Syst., vol. 2, no. 4, pp. 303–314, 1989.

[19] K. Hornik, “Approximation capabilities of multilayer feedforward networks,” Neural Netw., vol. 4, no. 2, pp. 251–257, 1991.

[20] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, “Multilayer feedforward networks with a nonpolynomial activation function can approximate any function,” Neural Netw., vol. 6, no. 6, pp. 861–867, Jan. 1993.

[21] A. Pinkus, “Approximation theory of the MLP model in neural networks,” Acta Numerica, vol. 8, pp. 143–195, Jan. 1999.

[22] G.-B. Huang, L. Chen, and C.-K. Siew, “Universal approximation using incremental constructive feedforward networks with random hidden nodes,” IEEE Trans. Neural Netw., vol. 17, no. 4, pp. 879–892, Jul. 2006.

[23] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: Theory and applications,” Neurocomputing, vol. 70, nos. 1–3, pp. 489–501, Dec. 2006.

[24] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, “Extreme learning machine for regression and multiclass classification,” IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 42, no. 2, pp. 513–529, Apr. 2012.

[25] A. Hart, J. Hook, and J. Davies, “Embedding and approximation theorems for echo state networks,” Neural Netw., vol. 128, pp. 234–247, Aug. 2020.

[26] R. I. Jennrich, “Asymptotic properties of non-linear least squares estimators,” Ann. Math. Statist., vol. 40, no. 2, pp. 633–643, Apr. 1969.

[27] G. B. Folland, Real Analysis: Modern Techniques and Their Applications, 2nd ed. Hoboken, NJ, USA: Wiley, 1999.

[28] J. H. Shapiro, AA Fixed-Point Farrago. Cham, Switzerland: Springer, 2016.

[29] A. R. Barron, “Universal approximation bounds for superpositions of a sigmoidal function,” IEEE Trans. Inf. Theory, vol. 39, no. 3, pp. 930–945, May 1993.

[30] J. M. Klusowski and A. R. Barron, “Approximation by combinations of ReLU and squared ReLU ridge functions with $l^1$ and $l^0$ controls,” IEEE Trans. Inf. Theory, vol. 64, no. 12, pp. 7649–7656, Dec. 2018.

[31] E. Weinan, C. Ma, and L. Wu, “A priori estimates of the population risk for two-layer neural networks,” Commun. Math. Sci., vol. 17, no. 5, pp. 1407–1425, 2019.

[32] E. Weinan, C. Ma, and L. Wu, “The Barron space and the flow-induced function spaces for neural network models,” Constructive Approximation, vol. 55, no. 1, pp. 369–406, Feb. 2022.

[33] E. Weinan and S. Wojtowytsch, “Representation formulas and pointwise properties for Barron functions,” Calculus Variat. Partial Differ. Equ., vol. 61, no. 2, pp. 1–37, Feb. 2022.

[34] S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[35] S. Boucheron, G. Lugosi, and P. Massart, Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford, U.K.: Oxford Univ. Press, 2013.

[36] P. L. Bartlett and S. Mendelson, “Rademacher and Gaussian complexities: Risk bounds and structural results,” J. Mach. Learn. Res., vol. 3, pp. 463–482, Nov. 2002.

Zhen Li received the Ph.D. degree in mathematics from Tsinghua University, Beijing, China, in 2016. He is currently a Senior Researcher of Theory Laboratory, Huawei Technologies Company Ltd. His current research interests include theory of deep learning and applied harmonic analysis.

Yunfei Yang received the Ph.D. degree in mathematics from The Hong Kong University of Science and Technology, Hong Kong, in 2022. He is currently a Post-Doctoral Researcher with the City University of Hong Kong, Hong Kong. His research interests include theory of deep learning, applied harmonic analysis, and approximation theory.