Error Bounds for a Matrix-Vector Product Approximation with Deep ReLU Neural Networks

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY-NC-SA 4.0

SUBMISSION DATE / POSTED DATE

22-08-2020 / 24-08-2020

CITATION

Getu, Tilahun (2020): Error Bounds for a Matrix-Vector Product Approximation with Deep ReLU Neural Networks. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.12845753.v1>

DOI

[10.36227/techrxiv.12845753.v1](https://doi.org/10.36227/techrxiv.12845753.v1)

Error Bounds for a Matrix-Vector Product Approximation with Deep ReLU Neural Networks

Tilahun M. Getu, *Member, IEEE*

Abstract—Inspired by the depth and breadth of developments on the theory of deep learning, we pose these fundamental questions: can we accurately approximate an arbitrary matrix-vector product using deep rectified linear unit (ReLU) feedforward neural networks (FNNs)? If so, can we bound the resulting approximation error? Attempting to answer these questions, we derive error bounds in Lebesgue and Sobolev norms for a matrix-vector product approximation with deep ReLU FNNs. Since a matrix-vector product models several problems in wireless communications and signal processing; network science and graph signal processing; and network neuroscience and brain physics, we discuss various applications that are motivated by an accurate matrix-vector product approximation with deep ReLU FNNs. Toward this end, the derived error bounds offer a theoretical insight and guarantee in the development of algorithms based on deep ReLU FNNs.

Index Terms—Deep ReLU FNNs, the theory of deep learning, error bounds, modeling via a matrix-vector product, applications, algorithm development.

I. INTRODUCTION

A. Related Works and Motivation

Understanding natural and artificial intelligence has been one of the most prominent inquiries of mankind. In this vein, the neuroscience, machine learning (ML), and artificial intelligence (AI) research communities have been pursuing this inquiry by advancing *whole brain emulation*, *biological cognition*, *human-machine interfaces*, and *AI* [1, Ch. 2]. Referring to the quest on the latter that aims to innovate intelligent and autonomous machines, feedforward neural networks (FNNs) have been the workhorse models since the Rosenblatt’s perceptron of the 1950s [2, Ch. 1]. Following the crucial discovery of the *backpropagation algorithm* in around 1986 [3], research on the representation and approximation power of (shallow) FNNs was intensified in late 1980s and 1990s [4]. During this era, several works established that FNNs with one hidden layer are universal approximators [4]–[9].

As an enabler of deep networks such as deep FNNs (D-FNNs), deep learning (DL) [10]–[12] can disentangle and learn [13] the various implicit discriminative features of an input data through depth of non-linear (also linear¹) transformations. This developing ML technique has become the norm of ML and AI research, especially after the super-human performance of *AlexNet* [17] demonstrated in the

2012 *ImageNet* [18] recognition challenge. Consequently, DL (and deep reinforcement learning) are applied in numerous research fields as diverse as clustering, information retrieval, dimensionality reduction, natural language processing [11], [12]; computer vision, speech recognition, image processing, object recognition [10], [11], [19]; wireless communications and networking [20]–[22]; gaming, finance, energy, healthcare [23], [24]; and alike.

Despite the broadly demonstrated significant performance of the existing DL-based ML/AI algorithms, one of the fundamental questions worth addressing rigorously is interpreting the operations of the respective ML/AI “blackbox” [25]. Motivated in part by the theories on (shallow) FNNs, the intense advancements and wide adoption of DL, and the need to have an *interpretable/explainable AI*, the *theory of deep learning (ToDL)* has emerged as a promising and vibrant research field since recently. Research in ToDL attempts to demystify the various hidden transformations of deep architectures to provide a theoretical guarantee and understanding on the learning, approximation, optimization, and generalization capability of deep networks—and their variants—such as FNNs, convolutional neural networks (CNNs) [26], [27], recurrent neural networks (RNNs) [28], [29], autoencoders (AEs) [30]–[32], generative adversarial networks (GANs) [33]–[35], *ResNet* [36], and *DenseNet* [37], [38]. To interpret/explain learning [39], approximation [40], optimization [41], [42], and generalization [43] in these deep networks employed for classification [44], [45] and regression problems [46], advancements in ToDL have been made via numerous frameworks such as mean field theory [47]–[49], random matrix theory [50], [51], tensor factorization [52], [53], optimization theory [54]–[57], kernel learning [58]–[60], linear algebra [61], [62], spline theory [63], [64], theoretical neuroscience [65]–[67], high-dimensional probability and statistics [68]–[70], manifold theory [48], [71], Fourier analysis [72], and scattering networks (vis-à-vis a wavelet transform) [73], [74].

Several works on ToDL have established that deep FNNs have the power of *exponential expressivity* compared to FNNs with a single hidden layer which require an exponential number of hidden neurons [40], [48], [75]–[78]. Besides, recent ToDL research trends are on universal approximation capability of ResNet [79] and of slim (and sparse) networks [80], the duality of depth and width [81], the width-efficiency of rectified linear unit (ReLU) FNNs [82], the improving effect of depth and width [83], and *un-rectifying* signal representation [84]. Meanwhile, [85]–[89] provide approximation theories on various functions approximated with deep ReLU FNNs. Although these foundational works are interesting in their own rights and have the potential to inspire much more research,

T. M. Getu is with the National Institute of Standards and Technology (NIST), 100 Bureau Drive, Gaithersburg, MD 20899, USA and also with the École de Technologie Supérieure (ÉTS), Montréal, QC H3C 1K3, Canada (e-mail: tilahun.getu@nist.gov).

¹D-FNNs with linear layers can also learn several discriminative features of the input data successfully. In this respect, their convergence was characterized in [14], [15] and their nonlinear learning dynamics was analyzed in [16].

they mainly focus on a scalar input and a scalar output.

In this information age of *data deluge* and hence *big data*, ToDL (also *learning theory* and the *mathematics of data science*²) must be developed for a DL in high dimension. Besides, with respect to (w.r.t.) high-dimensional probability and statistics [90], [91] and empirical process theory [92]—core tools of ToDL, high-dimensional bounds can provide useful insights on approximation with deep ReLU FNNs fed with input data manifesting high-dimension. Toward this timely need, [93] and [94] present FNN-based approximations applicable for parametric partial differential equations. Therefore, following the lead of [93], [94] and [85]–[89], we set out to derive error bounds for a matrix-vector product approximation with deep ReLU FNNs. This is also motivated by the fact that a matrix-vector product models various research problems of *wireless communications and signal processing*; *network science and graph signal processing*; and *network neuroscience and brain physics* [95], [96].

B. Contributions

W.r.t. the presented motivation and the abstract's questions, we make these contributions: 1) we derive error bounds in *Lebesgue and Sobolev norms* for a matrix-vector product approximation with deep ReLU FNNs; 2) hoping to inspire more ToDL research by the IEEE SPS and IEEE ComSoc communities, we discuss applications in wireless communications and signal processing; network science and graph signal processing; and network neuroscience and brain physics.

Organization: Sec. II highlight the preliminaries. Sec. III present the problem formulation and FNN representation. Secs. IV and V, respectively (resp.), present the error bounds in Lebesgue and Sobolev norms. Sec. VI presents applications. Sec. VII provides concluding remarks and research outlook.

Notation: Italic letters represent scalars. Bold lowercase (uppercase) letters denote vectors (matrices). Calligraphic letters such as $\mathcal{D}$ (or $\mathcal{S}$) and $\mathcal{W}$, resp., implicate an arbitrary set and the *Sobolev spaces*. $\mathbb{C}^n(\mathbb{R}^n)$ and $\mathbb{C}^{m \times n}(\mathbb{R}^{m \times n})$, resp., represent the sets of n -dimensional vectors of complex(real) numbers and of $m \times n$ complex(real) matrices. $:=$, $\rightarrow$, $\in$, $\forall$, $|\cdot|$, and $(\cdot)^T$, resp., stand for equal by definition, to, element of, for all, cardinality, and transpose. $\text{vec}(\cdot)$, $\text{diag}(\cdot)$, $\text{Re}\{\cdot\}$, $\text{Im}\{\cdot\}$, $\max(\cdot, \cdot)$ (or $\max\{\cdot, \cdot\}$), and $\log_2(\cdot)$, resp., denote vectorization, (block) diagonal matrix, real part, imaginary part, maximum, and logarithm to the base 2.

$\mathbb{R}_+$, $\mathbb{N}$, and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, resp., denote the sets of positive real numbers, of natural numbers, and of natural numbers including zero. For $k \in \mathbb{N}_0$, $\mathbb{N}_{\geq k} := \{k, k+1, \dots\}$ denotes the sets of natural numbers that are greater than or equal to k . For $K, m, n \geq 2$, $[K] := \{1, \dots, K\}$, $[m] := \{1, \dots, m\}$, and $[n] := \{1, \dots, n\}$. $\mathbf{I}_n(\mathbf{I}_m)$, $\mathbf{0}_{m \times n}$, and $\mathbf{0}$, resp., denote an $n \times n$ ($m \times m$) identity matrix, an $m \times n$ zero matrix, and a zero vector (matrix) whose dimension is guided by the context. For $\mathbf{W} \in \mathbb{R}^{m \times n}$ and $\mathbf{x} \in \mathbb{R}^n$, $[\mathbf{W}]_{i,j}$ and $[\mathbf{x}]_i$, resp., denote the (i, j) -th element of $\mathbf{W}$ and the i -th entry of $\mathbf{x}$. Per the MATLAB[®] syntax,

$\mathbf{W}(i, :)$ and $\mathbf{W}(:, j)$, resp., denote the i -th row and j -th column of $\mathbf{W}$. The Python[®] syntax $\text{len}(\mathbf{W})$ denotes the number of rows of $\mathbf{W}$. $\|\mathbf{W}\|_{\ell_0} := |\{(i, j) : [\mathbf{W}]_{i,j} \neq 0\}|$; $\|\mathbf{x}\|_\infty := \max_{i=1, \dots, n} |[\mathbf{x}]_i|$; $\|\mathbf{W}\|_\infty := \max_{i,j} |[\mathbf{W}]_{i,j}|$. The horizontal concatenation of n vectors(matrices) is denoted as $[\mathbf{w}_1, \dots, \mathbf{w}_n](\mathbf{W}_1, \dots, \mathbf{W}_n)$. A sequence of K matrix-vector tuples is denoted as $[\mathbf{W}_1, \mathbf{b}_1], \dots, [\mathbf{W}_K, \mathbf{b}_K]$.

Throughout this paper, $\mathbb{R}^d$ implies a standard topology. For $\mathbf{a} \subset \mathbb{R}^d$, $\bar{\mathbf{a}}$ denotes the closure of $\mathbf{a}$. If $\mathbf{a}, \mathbf{b} \subset \mathbb{R}^d$, we then write $\mathbf{a} \subset \subset \mathbf{b}$ provided that $\bar{\mathbf{a}}$ is compact in $\mathbf{b}$. For the multi-index $\alpha \in \mathbb{N}_0^d$, we employ the notation $|\alpha| := \alpha_1 + \dots + \alpha_d$. W.r.t. L^p (Lebesgue) spaces, $\Omega \subset \mathbb{R}^d$, and $f(x) : \mathbb{R}^d \rightarrow \mathbb{R}$, the L^∞ -norm of f is defined as $\|f\|_{L^\infty(\Omega)} := \inf\{C \geq 0 : |f(x)| \leq C, \forall x \in \Omega\} = \text{ess sup}_{x \in \Omega} |f(x)|$, $f : \Omega \rightarrow \mathbb{R}$ measurable [86], [88]. For a function $f : X \rightarrow \mathbb{R}$, $\text{supp } f$ and $\text{ran } f$, resp., stand for the support of f and the range of f . If $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ are two functions, their composition is denoted as $g \circ f : X \rightarrow Z$. A function $f : \Omega \rightarrow \mathbb{R}^m$ for $\Omega \subset \mathbb{R}^d$ is said to be L -Lipschitz continuous for a constant $L > 0$ provided that $|f(x) - f(y)| \leq L|x - y|$.

II. PRELIMINARIES

A. The Mathematics of FNNs

Based on [86, Definition II.1] and [89, Definition 1.1], we state the following definition on the mathematics of FNNs.³

Definition 1. Suppose $K, N_0, N_1, N_2, \dots, N_K \in \mathbb{N}$ and $K \geq 2$. A map $\Phi : \mathbb{R}^{N_0} \rightarrow \mathbb{R}^{N_K}$ characterized as

$$\Phi(\mathbf{x}) := \mathbf{A}_K(\rho(\mathbf{A}_{K-1}(\rho(\dots \rho(\mathbf{A}_1(\mathbf{x})))))), \mathbf{x} \in \mathbb{R}^{N_0}, \quad (1)$$

with affine linear maps $\mathbf{A}_k : \mathbb{R}^{N_{k-1}} \rightarrow \mathbb{R}^{N_k}$, $k \in [K]$, and the ReLU activation function $\rho(a) := \max(a, 0)$, $a \in \mathbb{R}$, acting element-wise is called a ReLU FNN. Hereinafter, N_0 is the dimension of the 0-th layer (considered as input layer), K is the depth of the FNN ($\mathcal{L}(\Phi) := K$), N_k , $k \in [K-1]$, is the dimension of the k -th (hidden) layer, and N_K is the dimension of the output layer. The affine linear map corresponding to the k -th layer, $k \in [K-1]$, is given by $\mathbf{A}_k(\mathbf{x}_{k-1}) := \mathbf{W}_k \mathbf{x}_{k-1} + \mathbf{b}_k$ with $\mathbf{W}_k \in \mathbb{R}^{N_k \times N_{k-1}}$, $\mathbf{x}_{k-1} := \rho(\mathbf{A}_k(\mathbf{x}_{k-2})) \in \mathbb{R}^{N_{k-1}}$, and $\mathbf{b}_k \in \mathbb{R}^{N_k}$. To this end, $[\mathbf{A}_k]_{i,j}$ denotes the weight associated with the edge between the j -th node of the $(k-1)$ -th layer and the i -th node of the k -th layer, whereas $[\mathbf{b}_k]_i$ is the weight associated with the i -th node of the k -th layer. The total number of neurons, the network connectivity, the maximum width of the FNN, and the maximum absolute value of the FNN weights are, resp., defined as $\mathcal{N}(\Phi) := \sum_{k=0}^K N_k$, $\mathcal{M}(\Phi) := \sum_{k=1}^K (\|\mathbf{W}_k\|_{\ell_0} + \|\mathbf{b}_k\|_{\ell_0})$, $\mathcal{W}(\Phi) := \max_{k=0, \dots, K} N_k$, and $\mathcal{B}(\Phi) := \max_{k \in [K]} \{\|\mathbf{W}_k\|_\infty, \|\mathbf{b}_k\|_\infty\}$. The class of FNNs $\Phi : \mathbb{R}^{N_0} \rightarrow \mathbb{R}^{N_K}$ with no more than K layers, connectivity no more than M , input dimension N_0 , output dimension N_K , and activation function ρ is denoted by $\mathcal{NN}_{K, M, \rho}^{N_0, N_K}$. Toward this end, $\mathcal{NN}_{K, M, \rho}^{N_0, N_K} := \bigcup_{K \in \mathbb{N}} \mathcal{NN}_{K, M, \rho}^{N_0, N_K}$, $\mathcal{NN}_{K, \infty, \rho}^{N_0, N_K} := \bigcup_{M \in \mathbb{N}} \mathcal{NN}_{K, M, \rho}^{N_0, N_K}$, and $\mathcal{NN}_{\infty, \infty, \rho}^{N_0, N_K} := \bigcup_{K \in \mathbb{N}} \mathcal{NN}_{K, \infty, \rho}^{N_0, N_K}$.

Stating Definition 1 differently, the underneath definition follows from [87, Definition 2.2] and [88, Definition 2.1].

²ToDL, learning theory, and the mathematics of data science are not given an explicit distinction in scope by the ML/AI and math research communities.

³To stick to our notation, we assume a vector of FNN inputs and outputs.

Definition 2. Suppose $K, N_0, N_1, \dots, N_K \in \mathbb{N}$ and $K \geq 2$. A FNN denoted by Φ is a sequence of matrix-vector tuples

$$\Phi := [[\mathbf{W}_1, \mathbf{b}_1], \dots, [\mathbf{W}_K, \mathbf{b}_K]], \quad (2)$$

where $\mathbf{W}_k \in \mathbb{R}^{N_k \times N_{k-1}}$ and $\mathbf{b}_k \in \mathbb{R}^{N_k}$, $k \in [K]$, $\mathcal{N}(\Phi) = \sum_{k=0}^K N_k$, and $\mathcal{M}(\Phi) = \sum_{k=1}^K (\|\mathbf{W}_k\|_{\ell_0} + \|\mathbf{b}_k\|_{\ell_0})$. For $\mathbf{x} \in \mathbb{R}^{N_0}$ and $\rho(a) := \max(a, 0)$, a map $\Phi : \mathbb{R}^{N_0} \rightarrow \mathbb{R}^{N_K}$ is defined as $\Phi(\mathbf{x}) = \mathbf{x}_K$, where $\mathbf{x}_K$ is obtained recursively as

$$\mathbf{x}_0 := \mathbf{x}; \quad \mathbf{x}_k := \rho(\mathbf{W}_k \mathbf{x}_{k-1} + \mathbf{b}_k), \quad k \in [K-1] \quad (3a)$$

$$\mathbf{x}_K := \mathbf{W}_K \mathbf{x}_{K-1} + \mathbf{b}_K, \quad (3b)$$

where $\rho(\mathbf{y}) = [\rho(y_1), \dots, \rho(y_m)]^T$ for $\mathbf{y} = [y_1, \dots, y_m]^T$.

Expressed via (2), the concatenation of two FNNs follows.

Definition 3. Let $K_1, N_0^1, \dots, N_{K_1}^1 \in \mathbb{N}$ and $K_2, N_0^2, \dots, N_{K_2}^2 \in \mathbb{N}$. Define two FNNs as $\Phi^1 := [[\mathbf{W}_1^1, \mathbf{b}_1^1], \dots, [\mathbf{W}_{K_1}^1, \mathbf{b}_{K_1}^1]]$ and $\Phi^2 := [[\mathbf{W}_1^2, \mathbf{b}_1^2], \dots, [\mathbf{W}_{K_2}^2, \mathbf{b}_{K_2}^2]]$. For $N_0^1 = N_{K_2}^2$, the concatenation of Φ^1 and Φ^2 is given by [88, Definition 2.2]

$$\Phi^1 \bullet \Phi^2 := [[\mathbf{W}_1^1, \mathbf{b}_1^1], \dots, (\mathbf{W}_{K_2-1}^2, \mathbf{b}_{K_2-1}^2), (\mathbf{W}_1^1 \mathbf{W}_{K_2}^2, \mathbf{W}_1^1 \mathbf{b}_{K_2}^2 + \mathbf{b}_1^1), (\mathbf{W}_2^1, \mathbf{b}_2^1), \dots, (\mathbf{W}_{K_1}^1, \mathbf{b}_{K_1}^1)], \quad (4)$$

where $\mathcal{L}(\Phi^1 \bullet \Phi^2) = \sum_{i=1}^2 K_i - 1$.

Note that $(\Phi^1 \bullet \Phi^2)(\mathbf{x}) = \Phi^1(\Phi^2(\mathbf{x}))$ and (4) follows from (1). For a ReLU activation $\rho(\cdot)$ and $\mathbf{x} \in \mathbb{R}$, $\mathbf{x} = \rho(\mathbf{x}) - \rho(-\mathbf{x})$ [86]. This identity is exploited to prove the following lemma.

Lemma 1 ([88, Remark 2.4]). For $N_0 \geq 2$, $K \in \mathbb{N}_{\geq 2}$, and $\mathbf{x} \in \mathbb{R}^{N_0}$, one can construct a ReLU identity FNN (I-FNN) Φ_K^{I, N_0} with K layers and at most $2N_0 K \{-1, 1\}$ -valued weights such that $\Phi_K^{I, N_0}(\mathbf{x}) = \mathbf{x}$. Thus, one can choose:

$$\Phi_K^{I, N_0} := [[[\mathbf{I}_{N_0}, -\mathbf{I}_{N_0}]^T, \mathbf{0}], \overbrace{[\mathbf{I}_{2N_0}, \mathbf{0}], \dots, [\mathbf{I}_{2N_0}, \mathbf{0}]}^{(K-2) \text{ times}}, [\mathbf{I}_{N_0}, -\mathbf{I}_{N_0}]]]. \quad (5)$$

For $K = 1$, an I-FNN is constructed as $\Phi_1^{I, N_0} := [[\mathbf{I}_{N_0}, \mathbf{0}]]$.

Proof. Please refer to Appendix A.

Meanwhile, the parallelization of FNNs is formalized below.

Lemma 2. Consider n FNNs $\Phi^i := [[\mathbf{W}_1^i, \mathbf{b}_1^i], \dots, [\mathbf{W}_K^i, \mathbf{b}_K^i]]$, $i \in [n]$. Let $K, n \geq 2$ and $K, n \in \mathbb{N}$. For these K -layer FNNs with input $\mathbf{x} \in \mathbb{R}^{N_0}$, the parallelization of $\Phi^1, \dots, \Phi^n$ be denoted and defined as $\mathcal{P}(\Phi^1, \dots, \Phi^n) := [[\tilde{\mathbf{W}}_1, \tilde{\mathbf{b}}_1], \dots, [\tilde{\mathbf{W}}_K, \tilde{\mathbf{b}}_K]]$. If $\mathcal{P}(\Phi^1, \dots, \Phi^n)(\mathbf{x}) = [(\Phi^1(\mathbf{x}))^T, \dots, (\Phi^n(\mathbf{x}))^T]^T$, then for $\tilde{\Phi} := [\Phi^1, \dots, \Phi^n]$ and $2 \leq k \leq K$

$$\tilde{\mathbf{W}}_1 = [(\mathbf{W}_1^1)^T, \dots, (\mathbf{W}_1^n)^T]^T; \tilde{\mathbf{b}}_1 = [(\mathbf{b}_1^1)^T, \dots, (\mathbf{b}_1^n)^T]^T \quad (6)$$

$$\tilde{\mathbf{W}}_k = \text{diag}(\mathbf{W}_k^1, \dots, \mathbf{W}_k^n); \tilde{\mathbf{b}}_k = [(\mathbf{b}_k^1)^T, \dots, (\mathbf{b}_k^n)^T]^T. \quad (7)$$

As a result, $\mathcal{L}(\mathcal{P}(\tilde{\Phi})) = K$, $\mathcal{M}(\mathcal{P}(\tilde{\Phi})) = \sum_{i=1}^n \mathcal{M}(\Phi^i)$, and $\mathcal{N}(\mathcal{P}(\tilde{\Phi})) = \sum_{i=1}^n \mathcal{N}(\Phi^i) - (n-1)N_0$.

Proof. Please refer to Appendix B.

Moreover, w.r.t. the superposition and parallelization of n FNNs, we adopt the following lemma.

Lemma 3 ([86, Lemma II.7]). Suppose $\rho(x) := \max(x, 0)$, $a_i \in \mathbb{R}$, $n, d_i, K_i, M_i \in \mathbb{N}$, $\Phi_i \in \mathcal{NN}_{K_i, M_i, \rho}^{d_i, 1}$, $i \in [n]$, and $d = \sum_{i=1}^n d_i$. Let $\mathbf{x} = [\mathbf{x}_1^T, \dots, \mathbf{x}_n^T]^T \in \mathbb{R}^d$ with $\mathbf{x}_i \in \mathbb{R}^{d_i}$, $i \in [n]$. There exist FNNs $\Phi^1 \in \mathcal{NN}_{K, M, \rho}^{d, n}$ and $\Phi^2 \in \mathcal{NN}_{K, M+n, \rho}^{d, 1}$ with $K = \max_i K_i$, $\mathcal{W}(\Phi^1) = \mathcal{W}(\Phi^2) \leq \sum_{i=1}^n \max\{2, \mathcal{W}(\Phi^i)\}$, and $M = \sum_{i=1}^n (M_i + \mathcal{W}(\Phi_i) + 2(K - K_i) + 1)$ fulfilling $\Phi^1(\mathbf{x}) = [a_1 \Phi^1(\mathbf{x}_1), \dots, a_n \Phi^1(\mathbf{x}_n)]^T$ and $\Phi^2(\mathbf{x}) = \sum_{i=1}^n a_i \Phi_i(\mathbf{x}_i)$. Furthermore, the weights of Φ^1 and Φ^2 comprise the weights of Φ_i , $i \in [n]$, $\{a_1, \dots, a_n\}$, and ± 1 's.

B. Functional Norms and Spaces

Concerning the L^p -norm of a composition function, we adopt the following lemma proved in [88, p. 325-326].

Lemma 4 ([88, Lemma 5.2]). Suppose $m, n \in \mathbb{N}$ with $n \leq m$. Let $U \subset \mathbb{R}^m$ be open, $t : U \rightarrow \mathbb{R}^n$ be continuously differentiable, and $\emptyset \neq K \subset U$ be compact. Assume that $Dt(x) \in \mathbb{R}^{n \times m}$ has a full rank $\forall x \in K$. Then, there is a constant C fulfilling $\int_K (f \circ t)(x) dx \leq C \cdot \int_{t(K)} f(y) dy$ for all Borel measurable functions such that $f : t(K) \rightarrow \mathbb{R}_+$. For all $0 < p < \infty$ and $f : t(K) \rightarrow \mathbb{R}$ being measurable

$$\|f \circ t\|_{L^p(K)} \leq C^{1/p} \cdot \|f\|_{L^p(t(K))}. \quad (8)$$

For $\Omega \in \mathbb{R}^d$ being open and $n \in \mathbb{N}_0$, $C^n(\Omega)$ denotes the set of n times continuously differentiable functions on Ω . The space of test functions is denoted as $C_c^\infty(\Omega) := \{f \in C^\infty(\Omega) | \text{supp } f \subset \subset \Omega\}$. For $f \in C^n(\Omega)$ and $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq n$, $D^\alpha f := \partial^{|\alpha|} f / \partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}$. We hereby define the Sobolev space [4], [87].

Definition 4 ([87, Definition 3.1]). Suppose $n, d \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^d$ denote an open subset of $\mathbb{R}^d$. Let $L^p(\Omega)$ denote the standard Lebesgue spaces on Ω for $1 \leq p \leq \infty$. The Sobolev space $\mathcal{W}^{n,p}(\Omega)$ is defined $\forall \alpha \in \mathbb{N}_0^d$ as $\mathcal{W}^{n,p}(\Omega) := \{f \in L^p(\Omega) : D^\alpha f \in L^p(\Omega) \text{ with } |\alpha| \leq n\}$ equipped with the norm defined for $f \in \mathcal{W}^{n,p}(\Omega)$ as $\|f\|_{\mathcal{W}^{n,p}(\Omega)} := (\sum_{0 \leq |\alpha| \leq n} \|D^\alpha f\|_{L^p(\Omega)}^p)^{1/p}$, $1 \leq p < \infty$, and

$$\|f\|_{\mathcal{W}^{n,\infty}(\Omega)} := \max_{0 \leq |\alpha| \leq n} \|D^\alpha f\|_{L^\infty(\Omega)}. \quad (9)$$

W.r.t. (9), we will use the notation $\|f\|_{\mathcal{W}^{n,p}(\Omega)} = \|f\|_{\mathcal{W}^{n,p}} = \|f(x)\|_{\mathcal{W}^{n,p}(\Omega; dx)}$. Meanwhile, Definition 4 is generalized below for m vector-valued functions.

Definition 5 ([87, Definition B.1]). Let $n \in \mathbb{N}_0$, $m \in \mathbb{N}_{\geq 2}$, and $1 \leq p \leq \infty$. Then, the Sobolev spaces of m vector-valued functions are defined as $\mathcal{W}^{n,p}(\Omega; \mathbb{R}^m) := \{(f_1, \dots, f_m) : f_i \in \mathcal{W}^{n,p}(\Omega)\}$ with $\|f\|_{\mathcal{W}^{n,p}(\Omega; \mathbb{R}^m)} := (\sum_{i=1}^m \|f_i\|_{\mathcal{W}^{n,p}(\Omega)}^p)^{1/p}$, $1 \leq p < \infty$ and

$$\|f\|_{\mathcal{W}^{n,\infty}(\Omega; \mathbb{R}^m)} := \max_{i=1, \dots, m} \|f_i\|_{\mathcal{W}^{n,\infty}(\Omega)}. \quad (10)$$

Concerning some analyses of this paper, we also adopt the definition—stated below—of the Sobolev semi-norm.

Definition 6 ([87, Definition B.2]). For $f \in \mathcal{W}^{n,p}(\Omega; \mathbb{R}^m)$ with $n, k \in \mathbb{N}_0$, $k \leq n$, $m \in \mathbb{N}$, and $1 \leq p \leq \infty$, the Sobolev semi-norm is defined as $|f|_{\mathcal{W}^{k,p}(\Omega; \mathbb{R}^m)} := (\sum_{i=1, |\alpha|=k}^m \|D^\alpha f_i\|_{L^p(\Omega)}^p)^{1/p}$, $1 \leq p < \infty$, and

$$|f|_{\mathcal{W}^{k,\infty}(\Omega; \mathbb{R}^m)} := \max_{i=1, \dots, m; |\alpha|=k} \|D^\alpha f_i\|_{L^\infty(\Omega)}. \quad (11)$$

For $m = 1$ and $k = 0$, Definition 6 corroborates that $|f|_{\mathcal{W}^{k,p}(\Omega)} = \|f\|_{L^p(\Omega)}$ [87]. Accordingly, we hereby define the Lebesgue spaces of m vector-valued functions.

Definition 7. Let $m \in \mathbb{N}_{\geq 2}$ and $1 \leq p \leq \infty$. Then, the Lebesgue spaces of m vector-valued functions are defined as $L^p(\Omega; \mathbb{R}^m) := \{(f_1, \dots, f_m) : f_i \in L^p(\Omega)\}$ with $\|f\|_{L^p(\Omega; \mathbb{R}^m)} := (\sum_{i=1}^m \|f_i\|_{L^p(\Omega)}^p)^{1/p}$, $1 \leq p < \infty$ and

$$\|f\|_{L^\infty(\Omega; \mathbb{R}^m)} := \max_{i=1, \dots, m} \|f_i\|_{L^\infty(\Omega)}. \quad (12)$$

To facilitate our subsequent analyses, we will exploit the following corollary which was proved in [87, p. 23-24].

Corollary 1 ([87, Corollary B.5]). For $d, m \in \mathbb{N}$, suppose $\Omega_1 \subset \mathbb{R}^d$ and $\Omega_2 \subset \mathbb{R}^m$ are both open, bounded, and convex. Then, there is a constant $C = C(d, m) > 0$ with regard to the following property: if $f \in \mathcal{W}^{1,\infty}(\Omega_1; \mathbb{R}^m)$ and $g \in \mathcal{W}^{1,\infty}(\Omega_2)$ are Lipschitz continuous functions such that $\text{ran } f \subset \Omega_2$, then $g \circ f \in \mathcal{W}^{1,\infty}(\Omega_1)$ and it holds that

$$|g \circ f|_{\mathcal{W}^{1,\infty}(\Omega_1)} \leq C |g|_{\mathcal{W}^{1,\infty}(\Omega_2)} |f|_{\mathcal{W}^{1,\infty}(\Omega_1; \mathbb{R}^m)} \quad (13a)$$

$$\|g \circ f\|_{\mathcal{W}^{1,\infty}(\Omega_1)} \leq C \max\{\|g\|_{L^\infty(\Omega_2)}, \tilde{g}\tilde{f}\}, \quad (13b)$$

where $\tilde{g} = |g|_{\mathcal{W}^{1,\infty}(\Omega_2)}$ and $\tilde{f} = |f|_{\mathcal{W}^{1,\infty}(\Omega_1; \mathbb{R}^m)}$.

III. PROBLEM FORMULATION AND FNN REPRESENTATION

A. Problem Formulation

Let $\mathbf{W} \in \mathbb{R}^{m \times n}$ and $\mathbf{x} \in \mathbb{R}^n$, resp., be a matrix and a vector whose product is going to be approximated by a deep ReLU FNN denoted by $\Phi_{D,\varepsilon}^\Pi$. Let $\mathbf{w}_i^T := (\mathbf{W}(i, :))^T \in \mathbb{R}^n$ and $\mathbf{W} := [\mathbf{w}_1^T, \dots, \mathbf{w}_m^T]^T$. As a result,

$$\mathbf{W}\mathbf{x} = [\mathbf{w}_1\mathbf{x}, \dots, \mathbf{w}_m\mathbf{x}]^T \in \mathbb{R}^m. \quad (14)$$

Considering that $\Phi_{D,\varepsilon}^\Pi$ is fed with inputs which are all the elements of $\mathbf{W}$ and $\mathbf{x}$ to produce their approximated product, we want to derive the error bounds for $\|\Phi_{D,\varepsilon}^\Pi(\mathbf{W}, \mathbf{x}) - \mathbf{W}\mathbf{x}\|_{L^\infty([-D,D]^{2n}; \mathbb{R}^m)}$ and $\|\Phi_{D,\varepsilon}^\Pi(\mathbf{W}, \mathbf{x}) - \mathbf{W}\mathbf{x}\|_{\mathcal{W}^{1,\infty}((-D,D)^{2n}; \mathbb{R}^m)}$. Using the error bounds to be derived, we will derive an error bound for a matrix-vector product when all the elements of $\mathbf{W}$ and $\mathbf{x}$ are complex.

B. FNN Representation

FNNs are universal approximators that can represent any affine transformation. Toward this end, the following proposition follows.

Proposition 1. Let $\mathbf{W} \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^n$, and $\rho(x) := \max(x, 0)$, $x \in \mathbb{R}$. For the affine transformation $\mathbf{A} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ defined as $\mathbf{A} = \mathbf{W}\mathbf{x}$, the following holds w.r.t. Definition 1.

- 1) $\mathbf{A}$ can be represented by $\mathcal{NN}_{2,M,\rho}^{n,m}$ with $M = 2\|\mathbf{W}\|_{\ell_0} + 2m$.
- 2) For $K \gg 2$, $\mathbf{A}$ can be represented by $\mathcal{NN}_{K,M,\rho}^{n,m}$ with $M = 2m + 2(K-2)n + 4\|\mathbf{W}\|_{\ell_0}$.
- 3) For $K \gg 2$, $\mathbf{A}$ can also be represented by $\mathcal{NN}_{K,M,\rho}^{n,m}$ with $M = 2Km + 2\|\mathbf{W}\|_{\ell_0}$.

Proof. The proof is provided in Appendix C.

W.r.t. the questions posed in the abstract, error bounds in Lebesgue and Sobolev norms are derived subsequently.

IV. ERROR BOUNDS IN LEBESGUE NORMS

First, we state the following proposition proved in [86] using piecewise linear spline interpolation of $f(x) = x^2$ achieved through the ‘‘sawtooth’’ construction of [97] and [98].⁴

Proposition 2 ([86, Proposition III.1]). For all $\varepsilon \in (0, 1/2)$ and $\rho(x) := \max(x, 0)$, there exists a constant $C > 0$ and a FNN $\Phi_\varepsilon^{\text{sq}} \in \mathcal{NN}_{\infty,\infty,\rho}^{1,1}$ satisfying $\mathcal{L}(\Phi_\varepsilon^{\text{sq}}) \leq C \log_2(\varepsilon^{-1})$, $\mathcal{W}(\Phi_\varepsilon^{\text{sq}}) = 4$, $\mathcal{B}(\Phi_\varepsilon^{\text{sq}}) \leq 4$, $\Phi_\varepsilon^{\text{sq}}(0) = 0$, and $\|\Phi_\varepsilon^{\text{sq}}(x) - x^2\|_{L^\infty([0,1])} \leq \varepsilon$.

Building from Proposition 2, the following theorem follows.

Theorem 1. Let $D \in \mathbb{R}_+$, $\mathbf{W} \in \mathbb{R}^{m \times n}$, and $\mathbf{x} \in \mathbb{R}^n$. For all $\varepsilon \in (0, 1/2)$ and $\rho(x) := \max(x, 0)$, there exists a constant $C > 0$ such that there is a FNN $\Phi_{D,\varepsilon}^\Pi \in \mathcal{NN}_{\infty,\infty,\rho}^{n(m+1),m}$ satisfying $\mathcal{L}(\Phi_{D,\varepsilon}^\Pi) \leq C \log_2(nD^2\varepsilon^{-1})$, $\mathcal{W}(\Phi_{D,\varepsilon}^\Pi) \leq 12mn$, $\mathcal{B}(\Phi_{D,\varepsilon}^\Pi) \leq \max\{4, 2D^2\}$, $\Phi_{D,\varepsilon}^\Pi(0, \mathbf{x}) = \Phi_{D,\varepsilon}^\Pi(\mathbf{W}, 0) = 0$ and

$$\|\Phi_{D,\varepsilon}^\Pi(\mathbf{W}, \mathbf{x}) - \mathbf{W}\mathbf{x}\|_{L^\infty([-D,D]^{2n}; \mathbb{R}^m)} \leq \varepsilon. \quad (15)$$

Proof. The proof is provided in Appendix D.

Based on Theorem 1, the following corollary follows.

Corollary 2. Let $D \in \mathbb{R}_+$, $\mathbf{W} \in \mathbb{C}^{m \times n}$, $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{W}_1 := \text{Re}\{\mathbf{W}\} \in \mathbb{R}^{m \times n}$, $\mathbf{W}_2 := \text{Im}\{\mathbf{W}\} \in \mathbb{R}^{m \times n}$, $\mathbf{x}_1 := \text{Re}\{\mathbf{x}\} \in \mathbb{R}^n$, and $\mathbf{x}_2 := \text{Im}\{\mathbf{x}\} \in \mathbb{R}^n$. Let $\mathbf{p}_1 := \text{Re}\{\mathbf{W}\mathbf{x}\} \in \mathbb{R}^m$, $\mathbf{p}_2 := \text{Im}\{\mathbf{W}\mathbf{x}\} \in \mathbb{R}^m$, $\mathbf{W}_{1,2} := [\mathbf{W}_1, \mathbf{W}_2] \in \mathbb{R}^{m \times 2n}$, and $\mathbf{x}_{1,2} := [\mathbf{x}_1, \mathbf{x}_2] \in \mathbb{R}^{n \times 2}$. For all $\varepsilon \in (0, 1/2)$ and $\rho(x) := \max(x, 0)$, there exists a constant $C > 0$ such that there is a FNN $\Phi_{D,\varepsilon}^{\Pi_c} \in \mathcal{NN}_{\infty,\infty,\rho}^{2n(m+1),2m}$ satisfying $\mathcal{L}(\Phi_{D,\varepsilon}^{\Pi_c}) \leq C \log_2(4nD^2\varepsilon^{-1})$, $\mathcal{W}(\Phi_{D,\varepsilon}^{\Pi_c}) \leq 48mn$, $\mathcal{B}(\Phi_{D,\varepsilon}^{\Pi_c}) \leq \max\{4, 2D^2\}$, $\Phi_{D,\varepsilon}^{\Pi_c}(0, \mathbf{x}_{1,2}) = \Phi_{D,\varepsilon}^{\Pi_c}(\mathbf{W}_{1,2}, 0) = 0$ and

$$\|\Phi_{D,\varepsilon}^{\Pi_c}(\mathbf{W}_{1,2}, \mathbf{x}_{1,2}) - [\mathbf{p}_1^T, \mathbf{p}_2^T]^T\|_{L^\infty([-D,D]^{2n}; \mathbb{R}^{2m})} \leq \varepsilon. \quad (16)$$

Proof. Please refer to Appendix E.

⁴Similar interesting proof is also reported in [94, p. 30-31].

In light of Theorem 1 and Corollary 2, the concepts that are remarked underneath follow.

Remark 1. 1) As $\varepsilon \rightarrow 0$, $\mathcal{L}(\Phi_{D,\varepsilon}^\Pi) \rightarrow 0$, i.e., as an FNN gets deeper, it gives lower approximation error. 2) For $\Phi_{D,\varepsilon}^\Pi$ —considered as a teacher D-FNN—fed and trained with inputs that are elements of $\mathbf{W}$ and $\mathbf{x}$, one can deploy knowledge distillation (KD)⁵ [101] to train a shallow student FNN $\Phi_{D,\varepsilon}^{\text{eq}}$ fed with only elements of $\mathbf{x}$ such that $\Phi_{D,\varepsilon}^\Pi(\mathbf{W}, \mathbf{x}) \approx \Phi_{D,\varepsilon}^{\text{eq}}(\mathbf{x})$.

V. ERROR BOUNDS IN SOBOLEV NORMS

First, we state the proposition proved in [87, p. 29-30].

Proposition 3 ([87, Proposition C.1]). For all $\varepsilon \in (0, 1/2)$ and $\rho(x) := \max(x, 0)$, there exist constants $C_1, C_2, C_3, C_4 > 0$ such that there is a FNN $\Phi_\varepsilon^{\text{sq}} \in \mathcal{NN}_{\infty,\infty,\rho}^{1,1}$ with $\mathcal{M}(\Phi_\varepsilon^{\text{sq}}) \leq C_1 \log_2(\varepsilon^{-1})$, $\mathcal{L}(\Phi_\varepsilon^{\text{sq}}) \leq C_2 \log_2(\varepsilon^{-1})$, $\mathcal{N}(\Phi_\varepsilon^{\text{sq}}) \leq C_3 \log_2(\varepsilon^{-1})$, and $\Phi_\varepsilon^{\text{sq}}(0) = 0$ such that

$$\|\Phi_\varepsilon^{\text{sq}}(x) - x^2\|_{\mathbf{W}^{1,\infty}((0,1);dx)} \leq \varepsilon; \quad (17)$$

$$\|\Phi_\varepsilon^{\text{sq}}\|_{\mathbf{W}^{1,\infty}((0,1))} \leq C_4. \quad (18)$$

Based on Proposition 3, the following result follows.

Proposition 4 ([87, Proposition C.2]). For all $\varepsilon \in (0, 1/2)$, $D \in \mathbb{R}_+$, and $\rho(x) := \max(x, 0)$, there exist constants $\bar{C}, C_1 > 0$ and $C_2 = C_2(D) > 0$ such that there is a FNN $\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)} \in \mathcal{NN}_{\infty,\infty,\rho}^{2,1}$ that manifest the following properties:

$$\|\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w, x) - wx\|_{\mathbf{W}^{1,\infty}((-D,D)^2;dw dx)} \leq \varepsilon \quad (19a)$$

$$\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(0, x) = \Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(x, 0) = 0, \forall x \in \mathbb{R} \quad (19b)$$

$$\|\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}\|_{\mathbf{W}^{1,\infty}((-D,D)^2)} \leq \bar{C}D \quad (19c)$$

$$\begin{aligned} \mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) &\leq C_1 \log_2(\varepsilon^{-1}) + C_2; \mathcal{M}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) \leq C_1 \times \\ &\log_2(\varepsilon^{-1}) + C_2; \mathcal{N}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) \leq C_1 \log_2(\varepsilon^{-1}) + C_2. \end{aligned} \quad (20)$$

Proof. The proof is provided in Appendix F.

Employing (19a)-(20), the following theorem follows.

Theorem 2. Let $D \in \mathbb{R}_+$, $\mathbf{W} \in \mathbb{R}^{m \times n}$, and $\mathbf{x} \in \mathbb{R}^n$. For all $\varepsilon \in (0, 1/2)$ and $\rho(x) := \max(x, 0)$, there exist constants $\bar{C}_1, \bar{C}_2 > 0$ such that there is a FNN $\Phi_{D,\varepsilon}^\Pi \in \mathcal{NN}_{\infty,\infty,\rho}^{n(m+1),m}$ satisfying $\mathcal{L}(\Phi_{D,\varepsilon}^\Pi) \leq \bar{C}_1 \log_2(\varepsilon^{-1}) + \bar{C}_2$, $\mathcal{M}(\Phi_{D,\varepsilon}^\Pi) \leq \bar{C}_1 \log_2(\varepsilon^{-1}) + \bar{C}_2$, $\mathcal{N}(\Phi_{D,\varepsilon}^\Pi) \leq \bar{C}_1 \log_2(\varepsilon^{-1}) + \bar{C}_2$, $\Phi_{D,\varepsilon}^\Pi(0, \mathbf{x}) = \Phi_{D,\varepsilon}^\Pi(\mathbf{W}, 0) = \mathbf{0}$ and

$$\|\Phi_{D,\varepsilon}^\Pi(\mathbf{W}, \mathbf{x}) - \mathbf{W}\mathbf{x}\|_{\mathbf{W}^{1,\infty}((-D,D)^{2n};\mathbb{R}^m)} \leq \varepsilon. \quad (21)$$

Proof. The proof is deferred to Appendix G.

⁵For classification problems, KD is about one classifier being trained on the outputs of another classifier [99]. However, the “dark knowledge” offered by *soft labels* of a teacher FNN cannot be used for regression problems and [100] surmounted this issue by employing a teacher FNN’s loss as a confidence score that places variable relative importance on a teacher FNN’s prediction.

VI. APPLICATIONS

We highlight the underneath applications of a matrix-vector product approximation, as established in Theorems 1 and 2.

1) *Applications in Wireless Communications and Signal Processing:* apart from the contaminating Gaussian noise vector, an arbitrary matrix-vector product models, amongst others, these problems: a multiple-input multiple-output (MIMO) received signal [102]; a received signal over a *doubly selective* orthogonal frequency division multiplexing (OFDM) channel [103] and a doubly selective MIMO-OFDM channel [104]; a massively concurrent non-orthogonal multiple-access (MC-NOMA) [105] received signal; a multi-carrier code division multiple access (MC-CDMA) [106], [107]—also known as OFDM/CDMA—received signal; and so forth. Therefore, with regard to Theorem 1, Corollary 2, and Remark 1, deep ReLU FNNs can be trained for a blind detection of a MIMO received signal; a blind estimation of a doubly selective (MIMO-)OFDM channel; a blind detection of a (MIMO-)OFDM signal received through a doubly selective (MIMO-)OFDM channel; a blind detection of a MC-NOMA received signal and a MC-CDMA received signal; a blind estimation of a doubly selective MC-CDMA channel; a blind estimation of a *triply-selective* MIMO channel [108]; and so on.

2) *Applications in Network Science and Graph Signal Processing:* aside from the vector signifying error terms, an arbitrary matrix-vector product models, among others, these problems: non-parametric regression for graph-aware signal reconstruction; joint inference of signals and graphs; identification of directed graph topologies and tracking of dynamic networks; and so forth [109], [110]. Besides, a matrix-vector product models the graph Fourier transforms [111]. Accordingly, concerning Theorem 1, Corollary 2, Remark 1, and Proposition 1, algorithms and representations based on deep ReLU FNNs are applicable for solving these problems.

3) *Applications in Network Neuroscience and Brain Physics:* being one of the various interdisciplinary fields that comprise *brain physics* [112], network neuroscience aims to build a fundamental mechanistic understanding of how neuron level processes contribute to the structure and function of large-scale circuits, brain systems, and whole-brain structure and function [95]. It also inquires perception-action coupling, brain-behavior interactions, and social networks [95], [96]. To model *brain network structure*, an *adjacency matrix* is constructed from experimental data specifying the physical interconnections between neurons or brain regions [96, Fig. 1, p. 8] [113]. To model *brain network function*, a *similarity matrix* is constructed from experimental data on the activity of neurons or brain regions (e.g., fMRI⁶ data of blood oxygen level in different parts of the brain) [96, Fig. 2, p. 17]. If $\mathbf{W} \in \mathbb{R}^{m \times m}$ denotes the adjacency or similarity matrix, Theorem 1 corroborates that a deep ReLU FNN can accurately approximate $\text{diag}(\mathbf{W}, \dots, \mathbf{W})\text{vec}(\mathbf{I}_m)$ for $\text{diag}(\mathbf{W}, \dots, \mathbf{W}) \in \mathbb{R}^{m^2 \times m^2}$. By concatenating and training AEs with this approximating FNN, efficient *dimensionality reduction* can be achieved for a further insight into a brain network.

⁶fMRI: functional magnetic resonance imaging.

VII. CONCLUDING REMARKS AND RESEARCH OUTLOOK

We derive error bounds in Lebesgue and Sobolev norms for a matrix-vector product approximation with deep ReLU FNNs. These bounds offer a theoretical insight and guarantee in the development of algorithms based on deep ReLU FNNs. They can also serve as quantifiable guidance on the selection of a deep network's hyperparameters carried out during the training of deep ReLU FNNs. As highlighted in Section VI, this paper inspires research in wireless communications and signal processing; network science and graph signal processing; and network neuroscience and brain physics. It also inspires ToDL-based signal processing research that might be pursued through the lenses of the emerging tools of data science such as high-dimensional probability and statistics [90], [91].

Referring to the lead of [114], tighter error bounds for a matrix-vector product approximation with deep ReLU FNNs might be derived using *sparse grids*. At last, this paper also inspires ToDL research on error bounds for a matrix-matrix, a tensor-matrix, and a tensor-tensor product approximation—applicable in interference detection and excision, and channel estimation [115] and brain imaging [116]—with deep ReLU FNNs.

APPENDIX A PROOF OF LEMMA 1

If $\Phi_K^{I_{N_0}}(\mathbf{x}) = \mathbf{x} \in \mathbb{R}^{N_0}$ and $\rho(\mathbf{x}) = [\max(0, [\mathbf{x}]_1), \dots, \max(0, [\mathbf{x}]_{N_0})]^T$, then (1) leads to

$$\mathbf{A}_K(\rho(\mathbf{A}_{K-1}(\rho(\dots\rho(\mathbf{A}_1(\mathbf{x})))))) = \mathbf{x}, \quad (22)$$

where $\mathbf{x}_0 = \mathbf{x}$, $\mathbf{A}_k(\mathbf{x}_{k-1}) = \mathbf{W}_k \mathbf{x}_{k-1} + \mathbf{b}_k$, $\mathbf{x}_{k-1} := \rho(\mathbf{A}_k(\mathbf{x}_{k-2}))$, and $[\mathbf{W}_k, \mathbf{b}_k]$ —comprising $\Phi_K^{I_{N_0}}$ —is to be inferred for all $k \in [K]$. In the sequel, we will exploit these identities: $\mathbf{x} = \rho(\mathbf{x}) - \rho(-\mathbf{x})$ and $\rho(\rho(\mathbf{x})) = \rho(\mathbf{x})$. For (22) to be valid, here is one FNN configuration setting: all layers except the last layer has an output equal to $[(\rho(\mathbf{x}))^T, (\rho(-\mathbf{x}))^T]^T$. At the last layer which is without an activation function as per (1), we will add these two components so that we will obtain $\rho(\mathbf{x}) - \rho(-\mathbf{x}) = \mathbf{x}$. W.r.t. this setting, $\mathbf{x}_1 = \rho(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) = \rho(\mathbf{W}_1[\rho(\mathbf{x}) - \rho(-\mathbf{x})] + \mathbf{b}_1) = \rho(\mathbf{W}_1 \rho(\mathbf{x}) - \mathbf{W}_1 \rho(-\mathbf{x}) + \mathbf{b}_1) = [(\rho(\mathbf{x}))^T, (\rho(-\mathbf{x}))^T]^T$. W.r.t. the last equality, $\mathbf{W}_1 = [\mathbf{I}_{N_0}, \mathbf{I}_{N_0}]^T$ and $\mathbf{b}_1 = \mathbf{0}_{2N_0 \times 1}$. Thus, $[\mathbf{W}_1, \mathbf{b}_1] = [[\mathbf{I}_{N_0}, -\mathbf{I}_{N_0}]^T, \mathbf{0}]$.

W.r.t. the aforementioned setting, $\mathbf{x}_k = [(\rho(\mathbf{x}))^T, (\rho(-\mathbf{x}))^T]^T$, $k = 2, \dots, K-1$. Accordingly, $\mathbf{x}_2 = \rho(\mathbf{W}_2 \mathbf{x}_1 + \mathbf{b}_2) = \rho(\mathbf{W}_2[(\rho(\mathbf{x}))^T, (\rho(-\mathbf{x}))^T]^T + \mathbf{b}_2) = [(\rho(\mathbf{x}))^T, (\rho(-\mathbf{x}))^T]^T$. Concerning the last equality, $\mathbf{W}_2 = \mathbf{I}_{2N_0}$, $\mathbf{b}_2 = \mathbf{0}_{2N_0}$, and hence $[\mathbf{W}_2, \mathbf{b}_2] = [\mathbf{I}_{2N_0}, \mathbf{0}]$. Repeating identical procedure for the k -th, $k \in \{3, \dots, K-1\}$, layer, $[\mathbf{W}_k, \mathbf{b}_k] = [\mathbf{I}_{2N_0}, \mathbf{0}]$. On the last layer, $\mathbf{x}_K$ should be $\mathbf{x}$ for $\Phi_K^{I_{N_0}}$ to be an I-FNN. Thus, $\mathbf{x}_K = \mathbf{W}_K \mathbf{x}_{K-1} + \mathbf{b}_K = \mathbf{W}_K[(\rho(\mathbf{x}))^T, (\rho(-\mathbf{x}))^T]^T + \mathbf{b}_K = \mathbf{x}$. From the last equality, $[\mathbf{W}_K, \mathbf{b}_K] = [[\mathbf{I}_{N_0}, -\mathbf{I}_{N_0}], \mathbf{0}]$. Therefore, recalling that the ReLU I-FNN is a sequence of matrix-vector tuples as in (5), employing the inferred $[\mathbf{W}_k, \mathbf{b}_k]$, $k \in [K]$, leads to (5).

For $K = 1$, (22) leads to $\mathbf{x}_1 = \mathbf{W}_1 \mathbf{x} + \mathbf{b}_1 = \mathbf{x}$. Thus, $\Phi_K^{I_{N_0}} = [[\mathbf{I}_{N_0}, \mathbf{0}]]$. This completes the proof of Lemma 1. ■

APPENDIX B PROOF OF LEMMA 2

We first assume that $K \geq 2$. From (3a)-(3b), the first layer output of the parallelized FNN will be $\mathbf{x}_1^p = \rho(\tilde{\mathbf{W}}_1 \mathbf{x} + \tilde{\mathbf{b}}_1)$. For the n FNNs, the respective first layer outputs are given by $\mathbf{x}_1^i = \rho(\mathbf{W}_1^i \mathbf{x} + \mathbf{b}_1^i)$, $i \in [n]$. If $\mathcal{P}(\Phi^1, \dots, \Phi^n)(\mathbf{x})$ should be equal to $[(\Phi^1(\mathbf{x}))^T, \dots, (\Phi^n(\mathbf{x}))^T]^T$, this should be valid for the outputs of all layers. Hence, $\rho(\tilde{\mathbf{W}}_1 \mathbf{x} + \tilde{\mathbf{b}}_1) = [(\rho(\mathbf{W}_1^1 \mathbf{x} + \mathbf{b}_1^1))^T, \dots, (\rho(\mathbf{W}_1^n \mathbf{x} + \mathbf{b}_1^n))^T]^T$. Since ρ is applied element-wise, the following relationship follows as a first condition: $\rho(\tilde{\mathbf{W}}_1 \mathbf{x} + \tilde{\mathbf{b}}_1) = \rho([(W_1^1)^T, \dots, (W_1^n)^T]^T \mathbf{x} + [(b_1^1)^T, \dots, (b_1^n)^T]^T)$. Thus, $\tilde{\mathbf{W}}_1 = [(W_1^1)^T, \dots, (W_1^n)^T]^T$; $\tilde{\mathbf{b}}_1 = [(b_1^1)^T, \dots, (b_1^n)^T]^T$. This completes the proof of (6).

From the above-mentioned first condition, the following relation is valid w.r.t. the outputs of the respective second layers: $\rho(\tilde{\mathbf{W}}_2 \mathbf{x}_1^p + \tilde{\mathbf{b}}_2) = [(\rho(\mathbf{W}_2^1 \mathbf{x}_1^1 + \mathbf{b}_2^1))^T, \dots, (\rho(\mathbf{W}_2^n \mathbf{x}_1^n + \mathbf{b}_2^n))^T]^T$. Equivalently, $\rho(\tilde{\mathbf{W}}_2 \mathbf{x}_1^p + \tilde{\mathbf{b}}_2) = \rho(\text{diag}(\mathbf{W}_2^1, \dots, \mathbf{W}_2^n)[(x_1^1)^T, \dots, (x_1^n)^T]^T + [(b_2^1)^T, \dots, (b_2^n)^T]^T)$. From the parallelization of the outputs of the first layers of the n FNNs, $\mathbf{x}_1^p = [(x_1^1)^T, \dots, (x_1^n)^T]^T$. Accordingly, the second parallelization conditions emanating from the outputs of the respective second layers of the n FNNs are these: $\tilde{\mathbf{W}}_2 = \text{diag}(\mathbf{W}_2^1, \dots, \mathbf{W}_2^n)$ and $\tilde{\mathbf{b}}_2 = [(b_2^1)^T, \dots, (b_2^n)^T]^T$. Meanwhile, employing the exact procedure and continuing recursively leads to these conditions: $\tilde{\mathbf{W}}_k = \text{diag}(\mathbf{W}_k^1, \dots, \mathbf{W}_k^n)$, $\tilde{\mathbf{b}}_k = [(b_k^1)^T, \dots, (b_k^n)^T]^T$, $3 \leq k \leq K$. This completes the proof of (7).

As we are parallelizing n FNNs of equal depth, $\mathcal{L}(\mathcal{P}(\tilde{\Phi})) = K$. To continue, $\mathcal{M}(\mathcal{P}(\tilde{\Phi})) = \sum_{k=1}^K (\|\tilde{\mathbf{W}}_k\|_{\ell_0} + \|\tilde{\mathbf{b}}_k\|_{\ell_0})$. From the preceding derivations, $\|\tilde{\mathbf{W}}_k\|_{\ell_0} = \sum_{i=1}^n \|\mathbf{W}_k^i\|_{\ell_0}$ and $\tilde{\mathbf{b}}_k = \sum_{i=1}^n \|\mathbf{b}_k^i\|_{\ell_0}$. Thus, $\mathcal{M}(\mathcal{P}(\tilde{\Phi})) = \sum_{k=1}^K \sum_{i=1}^n (\|\mathbf{W}_k^i\|_{\ell_0} + \|\mathbf{b}_k^i\|_{\ell_0}) = \sum_{i=1}^n \sum_{k=1}^K (\|\mathbf{W}_k^i\|_{\ell_0} + \|\mathbf{b}_k^i\|_{\ell_0})$. Following Definition 1, $\mathcal{M}(\Phi^i) = \sum_{k=1}^K (\|\mathbf{W}_k^i\|_{\ell_0} + \|\mathbf{b}_k^i\|_{\ell_0})$ and hence $\mathcal{M}(\mathcal{P}(\tilde{\Phi})) = \sum_{i=1}^n \mathcal{M}(\Phi^i)$. Moreover, recalling that the total number of neurons is the sum of the number of rows of the weight matrices plus the total input neurons, $\mathcal{N}(\Phi^i) = N_0 + \sum_{k=1}^K \text{len}(\mathbf{W}_k^i)$ and $\mathcal{N}(\mathcal{P}(\tilde{\Phi})) = N_0 + \sum_{k=1}^K \text{len}(\tilde{\mathbf{W}}_k)$. Noting that $\text{len}(\tilde{\mathbf{W}}_k) = \sum_{i=1}^n \text{len}(\mathbf{W}_k^i)$ and $\sum_{k=1}^K \text{len}(\mathbf{W}_k^i) = \mathcal{N}(\Phi^i) - N_0$, $\mathcal{N}(\mathcal{P}(\tilde{\Phi})) = N_0 + \sum_{i=1}^n \sum_{k=1}^K \text{len}(\mathbf{W}_k^i) = N_0 + \sum_{i=1}^n (\mathcal{N}(\Phi^i) - N_0) = \sum_{i=1}^n \mathcal{N}(\Phi^i) - nN_0 + N_0 = \sum_{i=1}^n \mathcal{N}(\Phi^i) - (n-1)N_0$. This completes the proof of Lemma 2. ■

APPENDIX C PROOF OF PROPOSITION 1

Via a ReLU operation acting component-wise, $\mathbf{x} = \rho(\mathbf{x}) - \rho(-\mathbf{x})$ [86]. Exploiting this identity, $\mathbf{A} = \mathbf{I}_m \rho(\mathbf{A}) - \mathbf{I}_m \rho(-\mathbf{A})$ for $\mathbf{A} = \mathbf{W}\mathbf{x}$. Accordingly, $\mathbf{A} = \mathbf{I}_m \rho(\mathbf{W}\mathbf{x}) - \mathbf{I}_m \rho(-\mathbf{W}\mathbf{x})$.

Realizing $\rho(\mathbf{W}\mathbf{x})$ is a transformation with $\mathbf{x}$ as an input of a FNN, $\mathbf{W}$ as the weight matrix of a FNN, and $\rho(\cdot)$ as an element-wise ReLU activation, $\rho(\mathbf{W}\mathbf{x})$ is represented by $\mathcal{NN}_{1, M', \rho}^{n, m}$ with $M' = \|\mathbf{W}\|_{\ell_0}$. W.r.t. this 1-layer FNN, $\mathbf{I}_m \rho(\mathbf{W}\mathbf{x})$ can be realized as adding a linear operation on every corresponding output, i.e., multiplying every respective output by 1 w.r.t. an additional layer of no activation function. Hence, $\mathbf{I}_m \rho(\mathbf{W}\mathbf{x})$ can be represented by $\mathcal{NN}_{2, M_1, \rho}^{n, m}$ with

$M_1 = \|\mathbf{W}\|_{\ell_0} + m$ and no activation at the output layer. By the same token, $-\mathbf{I}_m \rho(-\mathbf{W}\mathbf{x})$ can also be represented by $\mathcal{NN}_{2,M_2,\rho}^{n,m}$ with $M_2 = \|\mathbf{W}\|_{\ell_0} + m = \|\mathbf{W}\|_{\ell_0} + m$ and also without any activation at the output layer. Similar to the concepts of [117, Lemma D.2], addition operation using two FNNs can be conceived as putting two ReLU FNNs parallelly and superimposing the outputs of the corresponding output neurons. Doing so results in a 2-layer FNN with a hidden layer double the size of the FNNs representing the summands. As a result, $\mathbf{A} = \mathbf{I}_m \rho(\mathbf{W}\mathbf{x}) - \mathbf{I}_m \rho(-\mathbf{W}\mathbf{x})$ can be represented by $\mathcal{NN}_{2,M,\rho}^{n,m}$ with $M = 2\|\mathbf{W}\|_{\ell_0} + 2m$ and without any activation at the output layer. This completes the proof of 1).

Utilizing the result of 1), let $\Phi^1 \in \mathcal{NN}_{2,M,\rho}^{n,m}$ and $\Phi^2 := \Phi_{K-1}^{I_m}$, an I-FNN defined via (5). As $\Phi_{K-1}^{I_m}(\mathbf{x}) = \mathbf{x}$, $\mathbf{A} = \mathbf{A}(\Phi_{K-1}^{I_m}) = \Phi^1(\Phi_{K-1}^{I_m}) = \Phi^1 \bullet \Phi_{K-1}^{I_m}$, the concatenation of Φ^1 and $\Phi_{K-1}^{I_m}$. To this end, deploying (4) and (5) leads to

$$\Phi^1 \bullet \Phi_{K-1}^{I_m} = \left[\left[[I_n, -I_n]^T, \mathbf{0} \right], \overbrace{[I_{2n}, \mathbf{0}], \dots, [I_{2n}, \mathbf{0}]}^{(K-1)-2 \text{ times}}, \right. \\ \left. [B_1, B_2], \mathbf{0} \right], [I_m, -I_m], \mathbf{0} \Big], \quad (23)$$

where $B_1 = [\mathbf{W}^T, -\mathbf{W}^T]^T$ and $B_2 = [-\mathbf{W}^T, \mathbf{W}^T]^T$. Hence, $\mathcal{L}(\Phi^1 \bullet \Phi_{K-1}^{I_m}) = K$ and $\mathcal{M}(\Phi^1 \bullet \Phi_{K-1}^{I_m}) = 2m + 2(K-2)n + 4\|\mathbf{W}\|_{\ell_0}$. Accordingly, $\mathbf{A}$ can also be represented by $\mathcal{NN}_{K,M,\rho}^{n,m}$ with $M = 2m + 2(K-2)n + 4\|\mathbf{W}\|_{\ell_0}$ and no activation at the output layer. This completes the proof of 2).

Similarly, with regard to the concatenation of (23), if we let $\Phi^1 := \Phi_{K-1}^{I_m}$ —an I-FNN defined via (5)—and $\Phi^2 \in \mathcal{NN}_{2,M,\rho}^{n,m}$, then $\mathbf{A} = \Phi_{K-1}^{I_m}(\mathbf{A}) = \Phi_{K-1}^{I_m}(\Phi^2) = \Phi_{K-1}^{I_m} \bullet \Phi^2$, the concatenation of $\Phi_{K-1}^{I_m}$ and Φ^2 . Consequently, employing (4) and (5) gives

$$\Phi_{K-1}^{I_m} \bullet \Phi^2 = \left[\left[[\mathbf{W}^T, -\mathbf{W}^T]^T, \mathbf{0} \right], [I(A_1), I(A_2)], \mathbf{0} \right], \\ \overbrace{[I_{2m}, \mathbf{0}], \dots, [I_{2m}, \mathbf{0}]}^{(K-1)-2 \text{ times}}, [I_m, -I_m], \mathbf{0} \Big], \quad (24)$$

where $I(A_1) = [I_m, -I_m]^T$, $I(A_2) = [-I_m, I_m]^T$. Accordingly, $\mathcal{L}(\Phi_{K-1}^{I_m} \bullet \Phi^2) = K$ and $\mathcal{M}(\Phi_{K-1}^{I_m} \bullet \Phi^2) = 2Km + 2\|\mathbf{W}\|_{\ell_0}$. Therefore, $\mathbf{A}$ can also be represented by $\mathcal{NN}_{K,M,\rho}^{n,m}$ with $M = 2Km + 2\|\mathbf{W}\|_{\ell_0}$ and no activation at the output layer. This completes the proof of 3). This also completes the proof of Proposition 1. ■

APPENDIX D PROOF OF THEOREM 1

To prove Theorem 1, we start by proving the following proposition.

Proposition 5. For all $\varepsilon \in (0, 1/2)$, $D \in \mathbb{R}_+$, and $\rho(x) := \max(x, 0)$, there exists a constant $C > 0$ such that there is a FNN $\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)} \in \mathcal{NN}_{\infty,\infty,\rho}^{2,1}$ satisfying $\mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq C \log_2(D^2 \varepsilon^{-1})$, $\mathcal{W}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq 12$, $\mathcal{B}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq \max\{4, 2D^2\}$, $\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(0, x) = \Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(x, 0) = 0$, $\forall x \in \mathbb{R}$, and

$$\|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w, x) - wx\|_{L^\infty([-D, D]^2)} \leq \varepsilon. \quad (25)$$

Proof. Similar proof is available in [86, p. 8-9]. For the sake of rigor and completeness, we henceforth provide our proof which will exploit Lemma 4.

We begin with the identity $wx = \frac{1}{2}((w+x)^2 - w^2 - x^2)$, $\forall x, y \in \mathbb{R}$. Hence, $wx = 4D^2(w/2D)(x/2D)$ and

$$wx = 4D^2 \frac{1}{2} \left(\left(\frac{w}{2D} + \frac{x}{2D} \right)^2 - \left(\frac{w}{2D} \right)^2 - \left(\frac{x}{2D} \right)^2 \right) \quad (26a)$$

$$= 2D^2 \left(\left(\frac{|w+x|}{2D} \right)^2 - \left(\frac{|w|}{2D} \right)^2 - \left(\frac{|x|}{2D} \right)^2 \right). \quad (26b)$$

W.r.t. (26b) and the identity $|x| = \rho(x) + \rho(-x)$, [97, Proposition 3] guarantees that $\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}$ can be expressed via the square FNN Φ_δ^{sq} characterized in Proposition 2. Consequently,

$$\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w, x) = 2D^2 \left(\Phi_\delta^{\text{sq}} \left(\frac{|w+x|}{2D} \right) - \Phi_\delta^{\text{sq}} \left(\frac{|w|}{2D} \right) - \Phi_\delta^{\text{sq}} \left(\frac{|x|}{2D} \right) \right), \quad (27)$$

such that

$$\|\Phi_\delta^{\text{sq}}(x) - x^2\|_{L^\infty([0,1])} \leq \delta. \quad (28)$$

Exploiting (27) and Lemma 3, the parameters of $\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}$ can be inferred from the parameters of Φ_δ^{sq} stated in Proposition 2. Thus, $\mathcal{W}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq \sum_{i=1}^3 \max\{2, 4\} = 12$; $\mathcal{B}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq \max\{4, 2D^2\}$, as each FNN in (27) can be implemented via the square FNN by adding one extra layer with weights $2D^2$; and $\mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq C \log_2(\delta^{-1})$, as asserted by (28) and Proposition 2. It also follows directly from (27) that $\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(0, x) = \Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(x, 0) = 0$, $\forall x \in \mathbb{R}$.

To continue, we define the following transformations:

$$t_w : [-D, D]^2 \rightarrow [0, 1], (w, x) \mapsto |w|/(2D) \quad (29a)$$

$$t_x : [-D, D]^2 \rightarrow [0, 1], (w, x) \mapsto |x|/(2D) \quad (29b)$$

$$t_{wx} : [-D, D]^2 \rightarrow [0, 1], (w, x) \mapsto |w+x|/(2D) \quad (29c)$$

$$f : [0, 1] \rightarrow \mathbb{R}, x \mapsto x^2. \quad (29d)$$

Using (29a)-(29d), (26b) can also be expressed as

$$wx = 2D^2(f \circ t_{wx} - f \circ t_w - f \circ t_x). \quad (30)$$

Similarly, deploying (29a)-(29d) in (27):

$$\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w, x) = 2D^2(\Phi_\delta^{\text{sq}} \circ t_{wx} - \Phi_\delta^{\text{sq}} \circ t_w - \Phi_\delta^{\text{sq}} \circ t_x). \quad (31)$$

Hence, employing (30) and (31) in the left-hand side of (25), and collecting similar terms give

$$\|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w, x) - wx\|_{L^\infty([-D, D]^2)} = \|2D^2(\Phi_\delta^{\text{sq}} - f) \circ t_{wx} \\ - 2D^2(\Phi_\delta^{\text{sq}} - f) \circ t_w - 2D^2(\Phi_\delta^{\text{sq}} - f) \circ t_x\|_{L^\infty([-D, D]^2)}. \quad (32)$$

Applying the properties of L^∞ -norm to (32),

$$\|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w, x) - wx\|_{L^\infty([-D, D]^2)} \leq \\ 2D^2 \sum_{u \in \{w, x, wx\}} \|(\Phi_\delta^{\text{sq}} - f) \circ t_u\|_{L^\infty([-D, D]^2)} \\ \stackrel{(i)}{\leq} 2D^2 \sum_{u \in \{w, x, wx\}} \|\Phi_\delta^{\text{sq}}(x) - x^2\|_{L^\infty(t_u([-D, D]^2))}, \quad (33)$$

where (i) follows from Lemma 4 through (8). From (29a)-(29d), $t_u([-D, D]^2) = [0, 1]$. Consequently,

$$\|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w, x) - wx\|_{L^\infty([-D, D]^2)} \leq 6D^2 \|\Phi_\delta^{\text{sq}}(x) - x^2\|_{L^\infty([0,1])}. \quad (34)$$

At last, exploiting (28) in (34) leads to

$$\|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w, x) - wx\|_{L^\infty([-D, D]^2)} \leq 6D^2 \delta = \varepsilon, \quad (35)$$

where we chose $\varepsilon = 6D^2\delta$. This leads to (25) and the depth constraint $\mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq C \log_2(\delta^{-1}) \leq C' \log_2(D^2\varepsilon^{-1})$, where C' is chosen as $C' \gg C \log_2 6$. This completes the proof of Proposition 5. ■

To proceed further toward the proof of Theorem 1, we proof the following proposition inspired by Proposition 5.

Proposition 6. *Let $\mathbf{w}^T = [w_1, \dots, w_n]^T \in \mathbb{R}^n$ and $\mathbf{x} = [x_1, \dots, x_n]^T \in \mathbb{R}^n$. For all $\varepsilon \in (0, 1/2)$, $D \in \mathbb{R}_+$, and $\rho(x) := \max(x, 0)$, there exists a constant $C > 0$ such that there is a FNN $\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)} \in \mathcal{NN}_{\infty,\infty,\rho}^{2n,1}$ satisfying $\mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq C \log_2(nD^2\varepsilon^{-1})$, $\mathcal{W}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq 12n$, $\mathcal{B}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq \max\{4, 2D^2\}$, $\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(\mathbf{0}, \mathbf{x}) = \Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(\mathbf{x}, \mathbf{0}) = 0$, $\forall \mathbf{x} \in \mathbb{R}^n$, and*

$$\|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(\mathbf{w}, \mathbf{x}) - \mathbf{w}\mathbf{x}\|_{L^\infty([-D, D]^{2n})} \leq \varepsilon. \quad (36)$$

Proof. Because $\mathbf{w}\mathbf{x} = \sum_{i=1}^n w_i x_i$, this sum of products can be realized via a ReLU FNN $\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}$ characterized in Proposition 5. Specifically, $\mathbf{w}\mathbf{x}$ can be implemented via the superposition of n $\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}$ that are, resp., fed with input tuples (w_i, x_i) , $i \in [n]$. Consequently,

$$\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(\mathbf{w}, \mathbf{x}) = \sum_{i=1}^n \Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w_i, x_i), \quad (37)$$

where $\|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w_i, x_i) - w_i x_i\|_{L^\infty([-D, D]^2)} \leq \tilde{\varepsilon}$, $\forall i$, $\mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w_i, x_i)}) \leq C \log_2(D^2\tilde{\varepsilon}^{-1})$, $\mathcal{W}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w_i, x_i)}) \leq 12$, and $\mathcal{B}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w_i, x_i)}) \leq \max\{4, 2D^2\}$. Moreover, $\Phi_{D,\varepsilon}^{\tilde{\chi}(w_i, x_i)}(\mathbf{0}, x_i) = \Phi_{D,\varepsilon}^{\tilde{\chi}(w_i, x_i)}(x_i, \mathbf{0}) = 0$, $\forall x_i \in \mathbb{R}$, and $i \in [n]$. Deploying these FNN constraints, we can infer from (37) that $\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(\mathbf{0}, \mathbf{x}) = \Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(\mathbf{x}, \mathbf{0}) = 0$, $\forall \mathbf{x} \in \mathbb{R}^n$. Besides, deploying Lemma 3 and using the aforementioned constraints, $\mathcal{W}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq 12n = \sum_{i=1}^n \max\{2, 12\}$ and $\mathcal{B}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w_i, x_i)}) \leq \max\{4, 2D^2\}$, as (37) can be implemented as a superposition of n FNNs with a weight matrix equals to an identity matrix and a zero bias vector at the last layer of the superposition FNN.

W.r.t. our subsequent analyses, we define these parameters for $\mathcal{D} := \{(w_1, x_1), \dots, (w_i, x_i), \dots, (w_n, x_n)\}$, $i \in [n]$:

$$t_{w_i x_i} : [-D, D]^{2n} \rightarrow [-D, D]^2, \quad \mathcal{D} \mapsto (w_i, x_i) \quad (38a)$$

$$g : [-D, D]^2 \rightarrow \mathbb{R}, \quad (w, x) \mapsto wx. \quad (38b)$$

Deploying (38a)-(38b) in (37) and $\mathbf{w}\mathbf{x} = \sum_{i=1}^n w_i x_i$,

$$(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(\mathbf{w}, \mathbf{x}), \mathbf{w}\mathbf{x}) = \sum_{i=1}^n (\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}, g) \circ t_{w_i x_i}. \quad (39)$$

Using (39) in the left-hand side of (36); collecting similar terms; and applying the properties of L^∞ -norm for $\Delta := \Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)} - g$

$$\begin{aligned} \|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(\mathbf{w}, \mathbf{x}) - \mathbf{w}\mathbf{x}\|_{L^\infty([-D, D]^{2n})} &= \left\| \sum_{i=1}^n \Delta \circ t_{w_i x_i} \right\|_{L^\infty([-D, D]^{2n})} \\ &\leq \sum_{i=1}^n \|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)} - g\|_{L^\infty(t_{w_i x_i}([-D, D]^{2n}))} \\ &\stackrel{(i)}{\leq} \sum_{i=1}^n \|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)} - g\|_{L^\infty(t_{w_i x_i}([-D, D]^{2n}))} \\ &\stackrel{(ii)}{=} \sum_{i=1}^n \|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w, x) - wx\|_{L^\infty([-D, D]^2)} \stackrel{(iii)}{=} \sum_{i=1}^n \tilde{\varepsilon}, \quad (40) \end{aligned}$$

where (i), (ii), and (iii), resp., follow from Lemma 4 via (8), (38a)-(38b), and (25)—w.r.t. the $\tilde{\varepsilon}$ constraint in $\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}$. Consequently, realizing that the superposition of n FNNs needs one additional layer and choosing $\varepsilon = n\tilde{\varepsilon}$ lead to (36) with $\mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq C \log_2(D^2\tilde{\varepsilon}^{-1}) + 1 = C[\log_2(nD^2\varepsilon^{-1}) + 1/C]$, as $\tilde{\varepsilon} = \varepsilon/n$. Choosing C to be large, $1/C \rightarrow 0$ and hence $\mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq C \log_2(nD^2\varepsilon^{-1})$. This completes the proof of Proposition 6. ■

Deploying Proposition 6, Lemma 2 on the parallelization of FNNs, and a part of Lemma 3 on the parallelization of FNNs, we prove Theorem 1 in the sequel.

From (14), $\mathbf{W}\mathbf{x} = [w_1 x, \dots, w_m x]^T \in \mathbb{R}^m$. Per Proposition 6, there are m FNNs approximating the m affine transformations with regard to $\mathbf{x}$ and the weights $w_1, \dots, w_m$. Hence, it can be inferred from (36) that

$$\|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w_i, x) - w_i x\|_{L^\infty([-D, D]^{2n})} \leq \varepsilon, \quad (41)$$

where $\mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq C \log_2(nD^2\varepsilon^{-1})$, $\mathcal{W}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq 12n$, $\mathcal{B}(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}) \leq \max\{4, 2D^2\}$, and $\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(\mathbf{0}, x) = \Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(x, \mathbf{0}) = 0$, $\forall x \in \mathbb{R}^n$. Accordingly,

$$\mathbf{W}\mathbf{x} \approx [\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w_1, x), \dots, \Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w_m, x)]^T, \quad (42)$$

where $(\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w_i, x))^T = \Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w_i, x)$, $i \in [m]$.

Per Lemma 2 and Lemma 3 on the parallelization of FNNs, there exist a FNN which exactly implements the parallelization of the FNNs that produce the right-hand side of (42). Therefore, under the constraints of Lemma 3 and assuming that these FNNs have the same depth, (42) can also be expressed for $\tilde{\Phi} := [\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}, \dots, \Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}]$ as

$$\mathcal{P}(\tilde{\Phi})(\mathbf{W}, \mathbf{x}) = [\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w_1, x), \dots, \Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w_m, x)]^T, \quad (43)$$

where—by Lemma 3— $\mathcal{L}(\mathcal{P}(\tilde{\Phi})) \leq C \log_2(nD^2\varepsilon^{-1})$, $\mathcal{W}(\mathcal{P}(\tilde{\Phi})) \leq \sum_{i=1}^m 12n = 12mn$, and $\mathcal{B}(\mathcal{P}(\tilde{\Phi})) \leq \max\{4, 2D^2\}$.

Because $\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}$ has inputs that are all the elements of $\mathbf{W}$ and $\mathbf{x}$ so as to produce their approximated product realizable via the parallelized FNN of (43), $\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(\mathbf{W}, \mathbf{x}) =$

$\mathcal{P}(\tilde{\Phi})(W, x)$. Thus, (43) corroborates w.r.t. every scalar output that

$$\Phi_{D,\varepsilon}^\Pi(W, x) = [\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w_1, x), \dots, \Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w_m, x)]^T, \quad (44)$$

where $\Phi_{D,\varepsilon}^\Pi(0, x) = \Phi_{D,\varepsilon}^\Pi(W, 0) = 0$. Meanwhile, employing Definition 7 and (12) in the left-hand side of (15),

$$\begin{aligned} \|\Phi_{D,\varepsilon}^\Pi(W, x) - Wx\|_{L^\infty([-D,D]^{2n}; \mathbb{R}^m)} &= \max_{i=1, \dots, m} \\ \|\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w_i, x) - w_i x\|_{L^\infty([-D,D]^{2n})} &\stackrel{(i)}{\leq} \varepsilon, \end{aligned} \quad (45)$$

where (i) follows from (41) and the bound of (15) is obtained. This completes the proof of (15). As $\Phi_{D,\varepsilon}^\Pi(W, x) = \mathcal{P}(\tilde{\Phi})(W, x)$, the aforementioned network constraints of $\mathcal{P}(\tilde{\Phi})$ are also the network constraints of $\Phi_{D,\varepsilon}^\Pi$ mentioned in Theorem 1. This concludes the proof of Theorem 1. ■

APPENDIX E PROOF OF COROLLARY 2

The proof follows from Theorem 1, the superposition of FNNs as in Lemma 3, the parallelization of FNNs as in Lemma 2 (also Lemma 3), and the following identities:

$$p_1 := \text{Re}\{Wx\} = \text{Re}\{W\}\text{Re}\{x\} - \text{Im}\{W\}\text{Im}\{x\} \quad (46a)$$

$$p_2 := \text{Im}\{Wx\} = \text{Re}\{W\}\text{Im}\{x\} + \text{Im}\{W\}\text{Re}\{x\}. \quad (46b)$$

Since $W_1 := \text{Re}\{W\} \in \mathbb{R}^{m \times n}$, $W_2 := \text{Im}\{W\} \in \mathbb{R}^{m \times n}$, $x_1 := \text{Re}\{x\} \in \mathbb{R}^n$, $x_2 := \text{Im}\{x\} \in \mathbb{R}^n$, (46a) and (46b) can be approximated via the superposition of two FNNs characterized via Theorem 1. Thus,

$$p_1 \approx \Phi_{D,\tilde{\varepsilon}}^\Pi(W_1, x_1) - \Phi_{D,\tilde{\varepsilon}}^\Pi(W_2, x_2) \quad (47a)$$

$$p_2 \approx \Phi_{D,\tilde{\varepsilon}}^\Pi(W_1, x_2) + \Phi_{D,\tilde{\varepsilon}}^\Pi(W_2, x_1). \quad (47b)$$

Per the superposition of FNNs in Lemma 3, (47a) and (47b) can be represented via FNNs $\Phi_{D,\tilde{\varepsilon}}^{\Pi_{s_1}}$ and $\Phi_{D,\tilde{\varepsilon}}^{\Pi_{s_2}}$ such that

$$[p_1, p_2] \approx [\Phi_{D,\tilde{\varepsilon}}^{\Pi_{s_1}}(W_{1,2}, x_{1,2}), \Phi_{D,\tilde{\varepsilon}}^{\Pi_{s_2}}(W_{1,2}, x_{1,2})], \quad (48)$$

where $W_{1,2} := [W_1, W_2]$ and $x_{1,2} := [x_1, x_2]$; $\mathcal{L}(\Phi_{D,\tilde{\varepsilon}}^{\Pi_{s_i}}) \leq C \log_2(nD^2\tilde{\varepsilon}^{-1}) + 1 \leq \bar{C} \log_2(nD^2\tilde{\varepsilon}^{-1})$ — $\bar{C} > C \gg 1$, $\mathcal{W}(\Phi_{D,\tilde{\varepsilon}}^{\Pi_{s_i}}) \leq 24mn$, $\mathcal{B}(\Phi_{D,\tilde{\varepsilon}}^{\Pi_{s_i}}) \leq \max\{4, 2D^2\}$, and $\Phi_{D,\tilde{\varepsilon}}^{\Pi_{s_i}}(0, x_{1,2}) = \Phi_{D,\tilde{\varepsilon}}^{\Pi_{s_i}}(W_{1,2}, 0) = 0$, $i = 1, 2$. Meanwhile, per Lemma 2 (also Lemma 3), the parallelization of the FNNs of (48) fed with all elements of $W_{1,2}$ and $x_{1,2}$ give $[p_1^T, p_2^T]^T$. This is equivalent to $\Phi_{D,\varepsilon}^{\Pi_c}$ and hence

$$\begin{aligned} \Phi_{D,\varepsilon}^{\Pi_c}(W_{1,2}, x_{1,2}) &= \left[\left(\Phi_{D,\tilde{\varepsilon}}^{\Pi_{s_1}}(W_{1,2}, x_{1,2}) \right)^T, \right. \\ &\quad \left. \left(\Phi_{D,\tilde{\varepsilon}}^{\Pi_{s_2}}(W_{1,2}, x_{1,2}) \right)^T \right]^T, \end{aligned} \quad (49)$$

where $\Phi_{D,\varepsilon}^{\Pi_c} \in \mathcal{NN}_{\infty,\infty,\rho}^{2n(m+1),2m}$ satisfying $\mathcal{L}(\Phi_{D,\varepsilon}^{\Pi_c}) \leq \bar{C} \log_2(nD^2\tilde{\varepsilon}^{-1})$, $\mathcal{W}(\Phi_{D,\varepsilon}^{\Pi_c}) \leq 48mn$, $\mathcal{B}(\Phi_{D,\varepsilon}^{\Pi_c}) \leq \max\{4, 2D^2\}$, and $\Phi_{D,\varepsilon}^{\Pi_c}(0, x_{1,2}) = \Phi_{D,\varepsilon}^{\Pi_c}(W_{1,2}, 0) = 0$. Meanwhile, deploying (49) and (46a)-(46b) in the left-hand side of (16),

$$\begin{aligned} &\|\Phi_{D,\varepsilon}^{\Pi_c}(W_{1,2}, x_{1,2}) - [p_1^T, p_2^T]^T\|_{L^\infty([-D,D]^{2n}; \mathbb{R}^m)} \stackrel{(i)}{\leq} \\ &\sum_{i=1}^2 \|\Phi_{D,\tilde{\varepsilon}}^\Pi(W_i, x_i) - W_i x_i\|_{L^\infty([-D,D]^{2n}; \mathbb{R}^m)} + \sum_{i=1}^2 \\ &\|\Phi_{D,\tilde{\varepsilon}}^\Pi(W_i, x_{i+(-1)^{i+1}}) - W_i x_{i+(-1)^{i+1}}\|_{L^\infty([-D,D]^{2n}; \mathbb{R}^m)} \\ &\stackrel{(ii)}{\leq} 4\tilde{\varepsilon}, \end{aligned} \quad (50)$$

where (i) follows from (46a)-(46b), (47a)-(48), and the properties of L^∞ -norm; (ii) follows from Theorem 1 and (15). Therefore, choosing $\varepsilon = 4\tilde{\varepsilon}$ gives (16) and $\mathcal{L}(\Phi_{D,\varepsilon}^{\Pi_c}) \leq \bar{C} \log_2(4nD^2\varepsilon^{-1})$. Choosing $C = \bar{C}$ leads to the depth bound stated in Corollary 2 whose proof is just completed. ■

APPENDIX F PROOF OF PROPOSITION 4

Similar proof is provided in [87, p. 31-33]. However, for completeness and clarity, we provide our version of the proof which our subsequent developments are based upon.

W.r.t. the identity $wx = \frac{1}{2}((w+x)^2 - w^2 - x^2)$ and (26b),

$$wx = 2D^2 \left(\left(\frac{|w+x|}{2D} \right)^2 - \left(\frac{|w|}{2D} \right)^2 - \left(\frac{|x|}{2D} \right)^2 \right). \quad (51)$$

In line with (51) and the identity $|x| = \rho(x) + \rho(-x)$, [97, Proposition 3] asserts that $\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}$ can be expressed in terms of a square ReLU FNN Φ_δ^{sq} which is characterized in Proposition 3. Consequently,

$$\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w, x) = 2D^2 \left(\Phi_\delta^{\text{sq}} \left(\frac{|w+x|}{2D} \right) - \Phi_\delta^{\text{sq}} \left(\frac{|w|}{2D} \right) - \Phi_\delta^{\text{sq}} \left(\frac{|x|}{2D} \right) \right), \quad (52)$$

such that $\|\Phi_\delta^{\text{sq}}(x) - x^2\|_{\mathcal{W}^{1,\infty}((0,1); dx)} \leq \delta$; cf. (17).

Meanwhile, implementing $|w+x|$, $|w|$, and $|x|$ requires 3 additional layers. Hence, employing this fact and [97, Proposition 3], $\mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) \leq C_0 \log_2(\varepsilon^{-1}) + C_0 \log_2(6D^2) + 3 \leq C_0 \log_2(\varepsilon^{-1}) + C_1$, where C_0 is a constant that emanates from the depth constraint of Proposition 3. As there would be 9 additional neurons in the implementations of the absolute value expressions of (52), [97, Proposition 3] also asserts that $\mathcal{N}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) \leq 3C_0 \log_2(\varepsilon^{-1}) + 3C_0 \log_2(6D^2) + 9 \leq C' \log_2(\varepsilon^{-1}) + C_2$. Moreover, since there would be 17 additional non-zero weights to implement $|w+x|$, $|w|$, and $|x|$ using the aforementioned additional three layers, $\mathcal{M}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) \leq 3C_0 \log_2(\varepsilon^{-1}) + 3C_0 \log_2(6D^2) + 17 \leq C'' \log_2(\varepsilon^{-1}) + C_3$. As $C', C'' > 0$, choosing $C_1 = C_1(D)$, $C_2 = C_2(D)$, $C_3 = C_3(D) > 0$ satisfy (20). Employing (52), $\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(0, x) = \Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(x, 0) = 0$, $\forall x \in \mathbb{R}$. Thus, (19b) is satisfied.

To continue, we will employ the following transformations:

$$t_w : (-D, D)^2 \rightarrow (0, 1), (w, x) \mapsto |w|/(2D) \quad (53a)$$

$$t_x : (-D, D)^2 \rightarrow (0, 1), (w, x) \mapsto |x|/(2D) \quad (53b)$$

$$t_{wx} : (-D, D)^2 \rightarrow (0, 1), (w, x) \mapsto |w + x|/(2D) \quad (53c)$$

$$f : (0, 1) \rightarrow \mathbb{R}, x \mapsto x^2. \quad (53d)$$

Using (53a)-(53d), (51) can be expressed as

$$wx = 2D^2(f \circ t_{wx} - f \circ t_w - f \circ t_x). \quad (54)$$

Similarly, using (53a)-(53d) in (52):

$$\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w, x) = 2D^2(\Phi_\delta^{\text{sq}} \circ t_{wx} - \Phi_\delta^{\text{sq}} \circ t_w - \Phi_\delta^{\text{sq}} \circ t_x). \quad (55)$$

Hence, using (54) and (55), and collecting similar terms give

$$\begin{aligned} & \|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w, x) - wx\|_{\mathcal{W}^{1,\infty}((-D,D)^2; dwdx)} \\ &= \|2D^2(\Phi_\delta^{\text{sq}} - f) \circ t_{wx} - 2D^2(\Phi_\delta^{\text{sq}} - f) \circ t_w \\ &\quad - 2D^2(\Phi_\delta^{\text{sq}} - f) \circ t_x\|_{\mathcal{W}^{1,\infty}((-D,D)^2)}. \end{aligned} \quad (56)$$

In line with Definition 4 and (9), applying the properties of L^∞ -norm to (56),

$$\begin{aligned} & \|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w, x) - wx\|_{\mathcal{W}^{1,\infty}((-D,D)^2; dwdx)} \leq \\ & 2D^2 \sum_{u \in \{w, x, wx\}} \|(\Phi_\delta^{\text{sq}} - f) \circ t_u\|_{\mathcal{W}^{1,\infty}((-D,D)^2)}. \end{aligned} \quad (57)$$

Applying Corollary 1—via (13b)—to (57) leads to

$$\begin{aligned} & \|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w, x) - wx\|_{\mathcal{W}^{1,\infty}((-D,D)^2; dwdx)} \leq \\ & 2D^2 C \sum_{u \in \{w, x, wx\}} \max \left\{ \|\Phi_\delta^{\text{sq}} - f\|_{L^\infty((0,1)^2)}, \right. \\ & \left. |\Phi_\delta^{\text{sq}} - f|_{\mathcal{W}^{1,\infty}((0,1)^2)} |t_u|_{\mathcal{W}^{1,\infty}((-D,D)^2)} \right\}. \end{aligned} \quad (58)$$

Employing Definition 6—via (11)—gives

$$|t_u|_{\mathcal{W}^{1,\infty}((-D,D)^2)} = 1/2D. \quad (59)$$

Substituting (59) into (58),

$$\begin{aligned} & \|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}(w, x) - wx\|_{\mathcal{W}^{1,\infty}((-D,D)^2; dwdx)} \leq 2D^2 C \\ & \sum_{u \in \{w, x, wx\}} \max \left\{ \|\Phi_\delta^{\text{sq}} - f\|_{L^\infty((0,1)^2)}, (2D)^{-1} |\Phi_\delta^{\text{sq}} - f|_{\mathcal{W}^{1,\infty}((0,1)^2)} \right\} \\ & \leq 6D^2 C \|\Phi_\delta^{\text{sq}} - x^2\|_{L^\infty((0,1)^2)}. \end{aligned} \quad (60)$$

Via Proposition 2, $\|\Phi_\delta^{\text{sq}}(x) - x^2\|_{L^\infty((0,1)^2)} \leq \delta$. Therefore, setting $\varepsilon = 6D^2 C \delta$, (60) leads to (19a).

Moreover, employing Definition 6—via (11)—in (55)

$$\begin{aligned} & |\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}|_{\mathcal{W}^{1,\infty}((-D,D)^2)} \leq 2D^2 \times \\ & \sum_{u \in \{w, x, wx\}} |\Phi_\delta^{\text{sq}}(x) \circ t_u|_{\mathcal{W}^{1,\infty}((-D,D)^2)} \stackrel{(i)}{\leq} \\ & 2D^2 C \sum_{u \in \{w, x, wx\}} |\Phi_\delta^{\text{sq}}(x)|_{\mathcal{W}^{1,\infty}((0,1))} |t_u|_{\mathcal{W}^{1,\infty}((-D,D)^2)} \\ & \stackrel{(ii)}{\leq} 2D^2 C 3C_4 (2D)^{-1} = 3CC_4 D, \end{aligned} \quad (61)$$

where (i) follows from (13a) and (ii) follows from (18) and (59). Choosing $\bar{C} = 3CC_4 > 0$, (61) asserts that $|\Phi_{D,\varepsilon}^{\tilde{\chi}(w,x)}|_{\mathcal{W}^{1,\infty}((-D,D)^2)} \leq \bar{C}D$. This proves (19c) and completes the proof of Proposition 4. ■

APPENDIX G PROOF OF THEOREM 2

On our way to the proof of Theorem 2, we begin by proving the following proposition.

Proposition 7. *Let $\mathbf{w}^T = [w_1, \dots, w_n]^T \in \mathbb{R}^n$ and $\mathbf{x} = [x_1, \dots, x_n]^T \in \mathbb{R}^n$. For all $\varepsilon \in (0, 1/2)$, $D \in \mathbb{R}_+$, and $\rho(x) := \max(x, 0)$, there exist constants $\bar{C}_1, \bar{C}_2, \bar{C}_3 > 0$ such that there is a FNN $\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})} \in \mathcal{NN}_{\infty, \infty, \rho}^{2n, 1}$ satisfying $\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}(\mathbf{0}, \mathbf{x}) = \Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}(\mathbf{x}, \mathbf{0}) = 0$, $\forall \mathbf{x} \in \mathbb{R}^n$, and*

$$\|\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}(\mathbf{w}, \mathbf{x}) - \mathbf{w}\mathbf{x}\|_{\mathcal{W}^{1,\infty}((-D,D)^{2n}; d\mathbf{w}d\mathbf{x})} \leq \varepsilon \quad (62)$$

$$\begin{aligned} \mathcal{M}(\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}) &\leq \bar{C}_1 \log_2(\varepsilon^{-1}) + \bar{C}_2; \mathcal{N}(\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}) \leq \bar{C}_1 \\ &\times \log_2(\varepsilon^{-1}) + \bar{C}_2; \mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}) \leq \bar{C}_1 \log_2(\varepsilon^{-1}) + \bar{C}_2 \end{aligned} \quad (63)$$

$$|\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}|_{\mathcal{W}^{1,\infty}((-D,D)^{2n})} \leq \bar{C}_3 D. \quad (64)$$

Proof. Since $\mathbf{w}\mathbf{x} = \sum_{i=1}^n w_i x_i$, it can be implemented via the superposition of n $\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}$ characterized via Proposition 4 and fed, resp., with input tuples $(w_1, x_1), \dots, (w_n, x_n)$. Thus, the output of a ReLU FNN— $\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}$ —that approximates $\mathbf{w}\mathbf{x}$ can be expressed as

$$\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}(\mathbf{w}, \mathbf{x}) = \sum_{i=1}^n \Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}(w_i, x_i), \quad (65)$$

where—via (19a)-(19c) and (20)— $\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}(\mathbf{0}, \mathbf{x}) = 0 = \Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}(\mathbf{x}, \mathbf{0})$, $\forall \mathbf{x} \in \mathbb{R}^n$;

$$\|\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}(\mathbf{w}, \mathbf{x}) - \mathbf{w}\mathbf{x}\|_{\mathcal{W}^{1,\infty}((-D,D)^{2n}; d\mathbf{w}d\mathbf{x})} \leq \tilde{\varepsilon} \quad (66)$$

$$\begin{aligned} \mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}) &\leq C_1 \log_2(\tilde{\varepsilon}^{-1}) + C_2; \mathcal{M}(\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}) \leq C_1 \times \\ &\log_2(\tilde{\varepsilon}^{-1}) + C_2; \mathcal{N}(\Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}) \leq C_1 \log_2(\tilde{\varepsilon}^{-1}) + C_2. \end{aligned} \quad (67)$$

Employing (52) in (65),

$$\begin{aligned} & \Phi_{D,\varepsilon}^{\tilde{\chi}(\mathbf{w}, \mathbf{x})}(\mathbf{w}, \mathbf{x}) = 2D^2 \\ & \times \sum_{i=1}^n \left(\Phi_\delta^{\text{sq}}\left(\frac{|w_i + x_i|}{2D}\right) - \Phi_\delta^{\text{sq}}\left(\frac{|w_i|}{2D}\right) - \Phi_\delta^{\text{sq}}\left(\frac{|x_i|}{2D}\right) \right). \end{aligned} \quad (68)$$

With respect to (51),

$$\mathbf{w}\mathbf{x} = 2D^2 \sum_{i=1}^n \left(\left(\frac{|w_i + x_i|}{2D}\right)^2 - \left(\frac{|w_i|}{2D}\right)^2 - \left(\frac{|x_i|}{2D}\right)^2 \right). \quad (69)$$

For our upcoming simplifications, we define these parameters for $\mathcal{S} := \{(w_1, x_1), \dots, (w_i, x_i), \dots, (w_n, x_n)\}$, $i \in [n]$:

$$t_{w_i} : (-D, D)^{2n} \rightarrow (0, 1), \mathcal{S} \mapsto |w_i|/(2D) \quad (70a)$$

$$t_{x_i} : (-D, D)^{2n} \rightarrow (0, 1), \mathcal{S} \mapsto |x_i|/(2D) \quad (70b)$$

$$t_{w_i x_i} : (-D, D)^{2n} \rightarrow (0, 1), \mathcal{S} \mapsto |w_i + x_i|/(2D) \quad (70c)$$

$$f : [0, 1] \rightarrow \mathbb{R}, x \mapsto x^2. \quad (70d)$$

Employing (70a)-(70d), (68) and (69) can be expressed as

$$\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w, x) = 2D^2 \times \sum_{i=1}^n (\Phi_{\delta}^{\text{sq}} \circ t_{w_i x_i} - \Phi_{\delta}^{\text{sq}} \circ t_{w_i} - \Phi_{\delta}^{\text{sq}} \circ t_{x_i}) \quad (71)$$

$$wx = 2D^2 \times \sum_{i=1}^n (f \circ t_{w_i x_i} - f \circ t_{w_i} - f \circ t_{x_i}). \quad (72)$$

Deploying (71) and (72) in the left-hand side of (62), collecting similar terms, and applying the properties of L^∞ -norm—via Definition 4 and (9)—result in

$$\|\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w, x) - wx\|_{\mathcal{W}^{1,\infty}((-D,D)^{2n}; dwdx)} \leq 2D^2 \times \sum_{i=1}^n \sum_{u \in \{w_i, x_i, w_i x_i\}} \|(\Phi_{\delta}^{\text{sq}} - f) \circ t_u\|_{\mathcal{W}^{1,\infty}((-D,D)^{2n})}. \quad (73)$$

Applying Corollary 1—via (13b)—to each summand of (73),

$$\|(\Phi_{\delta}^{\text{sq}} - f) \circ t_u\|_{\mathcal{W}^{1,\infty}((-D,D)^{2n})} \leq C \max \left\{ \|\Phi_{\delta}^{\text{sq}} - f\|_{L^\infty((0,1)^2)}, \|\Phi_{\delta}^{\text{sq}} - f\|_{\mathcal{W}^{1,\infty}((0,1)^2)} |t_u|_{\mathcal{W}^{1,\infty}((-D,D)^{2n})} \right\}. \quad (74)$$

Deploying (11) for $t_u \in \{t_{w_i}, t_{x_i}, t_{w_i x_i}\}$,

$$|t_u|_{\mathcal{W}^{1,\infty}((-D,D)^{2n})} = 1/2D. \quad (75)$$

Substituting (75) into (74) and, in turn, into (73) lead to

$$\begin{aligned} \|\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w, x) - wx\|_{\mathcal{W}^{1,\infty}((-D,D)^{2n}; dwdx)} &\leq 2D^2 C \times \\ &\sum_{i=1}^n \sum_{u \in \{w_i, x_i, w_i x_i\}} \max \left\{ \|\Phi_{\delta}^{\text{sq}} - f\|_{L^\infty((0,1)^2)}, (2D)^{-1} \times \right. \\ &\left. \|\Phi_{\delta}^{\text{sq}} - f\|_{\mathcal{W}^{1,\infty}((0,1)^2)} \right\} \leq 6D^2 C n \|\Phi_{\delta}^{\text{sq}}(x) - x^2\|_{L^\infty((0,1)^2)} \\ &\stackrel{(i)}{\leq} (6D^2 C \delta) n, \end{aligned} \quad (76)$$

where (i) follows from Proposition 2. W.r.t. (60) and the $\tilde{\varepsilon}$ constraint of (66), $\tilde{\varepsilon} = 6D^2 C \delta$. As a result, setting $\varepsilon = \tilde{\varepsilon} n$, (76) simplifies to (62). This completes the proof of (62).

Concerning the superposition of (65), $\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w, x)$ is implemented by adding the outputs of n $\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}$ —fed with (w_i, x_i) —via an additional output layer that comprises one neuron. Accordingly,

$$\begin{aligned} \mathcal{M}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) &= \sum_{i=1}^n \mathcal{M}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) + n \stackrel{(i)}{\leq} n[C_1 \log_2(n\varepsilon^{-1}) \\ &+ (C_2 + 1)] \stackrel{(ii)}{=} \bar{C}_1 \log_2(\varepsilon^{-1}) + \bar{C}_2, \end{aligned} \quad (77)$$

where (i) follows from (67) and the assignment $\varepsilon = \tilde{\varepsilon} n$; (ii) follows from these assignments: $\bar{C}_1 = nC_1$ and $\bar{C}_2 = n[C_1 \log_2 n + C_2 + 1]$. Similarly,

$$\begin{aligned} \mathcal{N}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) &= \sum_{i=1}^n \mathcal{N}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) + 1 \stackrel{(i)}{\leq} n[C_1 \log_2(n\varepsilon^{-1}) \\ &+ C_2] + 1 \leq \bar{C}_1 \log_2(\varepsilon^{-1}) + \bar{C}_2; \mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) = \mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) + 1 \\ &\stackrel{(i)}{\leq} C_1 \log_2(n\varepsilon^{-1}) + C_2 + 1 \leq \bar{C}_1 \log_2(\varepsilon^{-1}) + \bar{C}_2, \end{aligned} \quad (78)$$

where (i) follows from (67) and the assignment $\varepsilon = \tilde{\varepsilon} n$. Thus, the results in (77) and (78) complete the proof of (63).

To continue to the proof of (64), at last, deploying Definition 6—via (11)—and applying the properties of L^∞ -norm to (71)

$$\begin{aligned} |\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}|_{\mathcal{W}^{1,\infty}((-D,D)^{2n})} &\leq 2D^2 \sum_{i=1}^n \sum_{u \in \{w_i, x_i, w_i x_i\}} \\ |\Phi_{\delta}^{\text{sq}}(x_i) \circ t_u|_{\mathcal{W}^{1,\infty}((-D,D)^{2n})} &\stackrel{(i)}{\leq} 2D^2 C \sum_{i=1}^n \sum_{u \in \{w, x, wx\}} \\ |\Phi_{\delta}^{\text{sq}}(x)|_{\mathcal{W}^{1,\infty}((0,1))} |t_u|_{\mathcal{W}^{1,\infty}((-D,D)^{2n})} &\stackrel{(ii)}{\leq} 3nCC_4 D, \end{aligned} \quad (79)$$

where (i) follows from (13a); (ii) follows from (18) and (75). Setting $\bar{C}_3 = 3nCC_4 > 0$, (79) asserts that $|\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}|_{\mathcal{W}^{1,\infty}((-D,D)^{2n})} \leq \bar{C}_3 D$. This proves (64) and also completes the proof of Proposition 7. ■

Exploiting Proposition 7 which characterizes an approximating FNN on a vector-vector product and the parallelization of FNNs as in Lemma 2, we can now advance to the proof of Theorem 2 detailed in the sequel.

From (14), $Wx = [w_1 x, \dots, w_m x]^T \in \mathbb{R}^m$. Per Proposition 7, $w_i x - i \in [m]$ —can be approximated via a ReLU FNN such that

$$\|\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w_i, x) - w_i x\|_{\mathcal{W}^{1,\infty}((-D,D)^{2n}; dwdx)} \leq \varepsilon, \quad (80)$$

where the ReLU FNN constraints of (63) are valid. Thus, Wx can be approximated via the outputs of m ReLU FNNs as

$$[\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w_1, x), \dots, \Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w_m, x)]^T \approx Wx, \quad (81)$$

where $(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w_i, x))^T = \Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w_i, x)$, $i \in [m]$. Therefore, under the depth, connectivity, and neuronal number constraints of Lemma 2, (81) is also equated for $\tilde{\Phi} := [\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}, \dots, \Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}]$ as

$$\mathcal{P}(\tilde{\Phi})(W, x) = [\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w_1, x), \dots, \Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w_m, x)]^T, \quad (82)$$

where $\mathcal{M}(\mathcal{P}(\tilde{\Phi})) = \sum_{i=1}^m \mathcal{M}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) \leq m[\bar{C}_1 \log_2(\varepsilon^{-1}) + \bar{C}_2]$; $\mathcal{N}(\mathcal{P}(\tilde{\Phi})) = \sum_{i=1}^m \mathcal{N}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) - (m-1)N_0 \leq \sum_{i=1}^m \mathcal{N}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) \leq m[\bar{C}_1 \log_2(\varepsilon^{-1}) + \bar{C}_2]$; and $\mathcal{L}(\mathcal{P}(\tilde{\Phi})) = \mathcal{L}(\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}) \leq [\bar{C}_1 \log_2(\varepsilon^{-1}) + \bar{C}_2] \leq m[\bar{C}_1 \log_2(\varepsilon^{-1}) + \bar{C}_2]$.

Since $\Phi_{D,\varepsilon}^\Pi$ has inputs that are all the elements of W and x to produce their approximated product which is also realizable by the parallelized FNN of (82), $\Phi_{D,\varepsilon}^\Pi(W, x) = \mathcal{P}(\tilde{\Phi})(W, x)$. Thus, (82) corroborates w.r.t. every scalar output that

$$\Phi_{D,\varepsilon}^\Pi(W, x) = [\Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w_1, x), \dots, \Phi_{D,\varepsilon}^{\tilde{\times}(w,x)}(w_m, x)]^T. \quad (83)$$

As a result, $\Phi_{D,\varepsilon}^\Pi(\mathbf{0}, \mathbf{x}) = \Phi_{D,\varepsilon}^\Pi(\mathbf{W}, \mathbf{0}) = \mathbf{0}$. Therefore, substituting (83) and (14) into (21); collecting similar terms; and exploiting (10)

$$\begin{aligned} \|\Phi_{D,\varepsilon}^\Pi(\mathbf{W}, \mathbf{x}) - \mathbf{W}\mathbf{x}\|_{\mathcal{W}^{1,\infty}((-D,D)^{2n};\mathbb{R}^m)} &= \max_{i=1,\dots,m} \\ \|\Phi_{D,\varepsilon}^{\tilde{\mathbf{x}}(\mathbf{w},\mathbf{x})}(\mathbf{w}_i, \mathbf{x}) - \mathbf{w}_i\mathbf{x}\|_{\mathcal{W}^{1,\infty}((-D,D)^{2n};d\mathbf{w}d\mathbf{x})} &\stackrel{(i)}{\leq} \varepsilon, \end{aligned} \quad (84)$$

where (i) follows from (80). This completes the proof of (21).

Because $\Phi_{D,\varepsilon}^\Pi(\mathbf{W}, \mathbf{x}) = \mathcal{P}(\tilde{\Phi})(\mathbf{W}, \mathbf{x})$, the aforementioned network constraints of $\mathcal{P}(\tilde{\Phi})$ become the constraints of $\Phi_{D,\varepsilon}^\Pi$. As a result, letting $m\bar{C}_1 = \bar{C}_1$ and $m\bar{C}_2 = \bar{C}_2$ lead to the network constraints mentioned in Theorem 2. This concludes the proof of Theorem 2. ■

ACKNOWLEDGMENTS AND DEDICATION

The author acknowledges the US Department of Commerce and NIST for the provided funding. He also acknowledges NIST's stuff for facilitating his Guest Researcher-ship.

Along with his NIST colleagues, the author mourned the sudden loss of Michael Souryal who was his thoughtful NIST mentor. This paper is, thus, dedicated to the memory of Michael Souryal.

REFERENCES

- [1] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. Oxford, UK: Oxford Univ. Press, 2014.
- [2] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2009.
- [3] D. E. Rumelhart *et al.*, "Learning Representations by Back-propagating Errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [4] A. Pinkus, "Approximation theory of the MLP model in neural networks," *Acta Numerica*, vol. 8, pp. 143–195, 1999.
- [5] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems*, 1989.
- [6] K. Hornik *et al.*, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, p. 359–366, Jul. 1989.
- [7] K. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural Netw.*, vol. 2, no. 3, p. 183–192, May 1989.
- [8] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Mach. Learn.*, vol. 14, no. 1, p. 115–133, Jan 1994.
- [9] H. N. Mhaskar, "Neural networks for optimal approximation of smooth and analytic functions," *Neural Comput.*, vol. 8, pp. 164–177, 1996.
- [10] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
- [11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [12] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, no. 3–4, pp. 197–387, 2014.
- [13] Y. Bengio *et al.*, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1798–1828, 2013.
- [14] K. Kawaguchi, "Deep learning without poor local minima," in *Proc. NIPS*, 2016, pp. 586–594.
- [15] S. Arora *et al.*, "A convergence analysis of gradient descent for deep linear neural networks." [Online]. Available: <https://arxiv.org/pdf/1810.02281.pdf>
- [16] A. M. Saxe *et al.*, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," 2014. [Online]. Available: <https://arxiv.org/pdf/1312.6120.pdf>
- [17] A. Krizhevsky *et al.*, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, vol. 25, Jan. 2012, pp. 1097–1105.
- [18] J. Deng *et al.*, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. CVPR09*, 2009, p. 248–255.
- [19] L. Liu *et al.*, "Deep learning for generic object detection: A survey." [Online]. Available: <http://arxiv.org/abs/1809.02165>
- [20] Q. Mao, F. Hu, and Q. Hao, "Deep learning for intelligent wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2595–2621, 4th Quart. 2018.
- [21] M. Chen *et al.*, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3039–3071, 2019.
- [22] N. C. Luong *et al.*, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 2019.
- [23] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [24] Y. Li, "Deep reinforcement learning," 15 Oct. 2018. [Online]. Available: <https://arxiv.org/pdf/1810.06339.pdf>
- [25] D. Castelvetti, "The Black Box of AI," *Nature*, vol. 538, pp. 20–23, Oct. 2016.
- [26] J. Gu *et al.*, "Recent advances in convolutional neural networks," 19 Oct. 2017. [Online]. Available: <https://arxiv.org/pdf/1512.07108.pdf>
- [27] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
- [28] R. Jozefowicz *et al.*, "An empirical exploration of recurrent network architectures," in *Proc. ICML*, vol. 37, 2015, pp. 2342–2350.
- [29] K. Greff *et al.*, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, 2017.
- [30] M. Tschannen *et al.*, "Recent advances in autoencoder-based representation learning." [Online]. Available: <https://arxiv.org/pdf/1812.05069.pdf>
- [31] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," 2020. [Online]. Available: <https://arxiv.org/pdf/2003.05991.pdf>
- [32] S. Yu and J. C. Principe, "Understanding autoencoders with information theoretic concepts." [Online]. Available: <https://arxiv.org/pdf/1804.00057.pdf>
- [33] Z. Wang *et al.*, "Generative adversarial networks in computer vision: A survey and taxonomy." [Online]. Available: <https://arxiv.org/pdf/1906.01529.pdf>
- [34] A. Creswell *et al.*, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, 2018.
- [35] Z. Hu *et al.*, "On unifying deep generative models," 11 Jul. 2018. [Online]. Available: <https://arxiv.org/pdf/1706.00550.pdf>
- [36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.
- [37] G. Huang *et al.*, "Densely connected convolutional networks," in *Proc. CVPR*, 2017, pp. 2261–2269.
- [38] G. Huang *et al.*, "Convolutional networks with dense connectivity," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.
- [39] J. Fan, C. Ma, and Y. Zhong, "A selective overview of deep learning," 15 Apr. 2019. [Online]. Available: <https://arxiv.org/pdf/1904.05526.pdf>
- [40] H. Mhaskar and T. Poggio, "Deep vs. shallow networks : An approximation theory perspective." [Online]. Available: <https://arxiv.org/pdf/1608.03287.pdf>
- [41] R. Sun, "Optimization for deep learning: theory and algorithms," 19 Dec. 2019. [Online]. Available: <https://arxiv.org/pdf/1912.08957.pdf>
- [42] P. L. Bartlett *et al.*, "Representing smooth functions as compositions of near-identity functions with implications for deep network optimization." [Online]. Available: <https://arxiv.org/pdf/1804.05012.pdf>
- [43] K. Kawaguchi *et al.*, "Generalization in deep learning," 10 May 2019. [Online]. Available: <https://arxiv.org/pdf/1710.05468.pdf>
- [44] R. Giryes *et al.*, "Deep neural networks with random gaussian weights: A universal classification strategy?" *IEEE Trans. Signal Process.*, vol. 64, no. 13, pp. 3444–3457, 2016.
- [45] R. Giryes *et al.*, "Corrections to "deep neural networks with random gaussian weights: A universal classification strategy?,"" *IEEE Trans. Signal Process.*, vol. 68, pp. 529–531, 2020.
- [46] T. Poggio, A. Banburski, and Q. Liao, "Theoretical issues in deep networks," *Proc. Natl. Acad. Sci. U.S.A.*, Jun. 2020.
- [47] S. Mei, A. Montanari, and P.-M. Nguyen, "A mean field view of the landscape of two-layer neural networks," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 115, no. 33, pp. E7665–E7671, 2018.
- [48] B. Poole *et al.*, "Exponential expressivity in deep neural networks through transient chaos," in *Proc. NIPS*, 2016, pp. 3360–3368.
- [49] S. S. Schoenholz *et al.*, "Deep information propagation," Apr. 2017. [Online]. Available: <https://arxiv.org/pdf/1611.01232.pdf>
- [50] J. Pennington and Y. Bahri, "Geometry of neural network loss surfaces via random matrix theory," in *Proc. ICML*, 2017, pp. 2798–2806.
- [51] S. Becker *et al.*, "Geometry of energy landscapes and the optimizability of deep neural networks," *Phys. Rev. Lett.*, vol. 124, no. 10, 2020.
- [52] B. D. Haeffele and R. Vidal, "Global optimality in tensor factorization, deep learning, and beyond," 24 Jun. 2015. [Online]. Available: <https://arxiv.org/pdf/1506.07540.pdf>

- [53] N. Cohen *et al.*, “On the expressive power of deep learning: A tensor analysis,” in *Proc. COLT*, 2016, pp. 698–728.
- [54] Y. Bengio, N. L. Roux, P. Vincent, O. Delalleau, and P. Marcotte, “Convex neural networks,” in *Proc. NIPS*, 2006, pp. 123–130.
- [55] F. Bach, “Breaking the curse of dimensionality with convex neural networks,” [Online]. Available: <https://arxiv.org/pdf/1412.8690.pdf>
- [56] S. Feizi *et al.*, “Porcupine neural networks: (almost) all local optima are global,” [Online]. Available: <https://arxiv.org/pdf/1710.02196.pdf>
- [57] J. Huang, Q. Qiu, R. Calderbank, and G. Sapiro, “Geometry-aware deep transform,” in *Proc. ICCV*, 2015, pp. 4139–4147.
- [58] A. Jacot *et al.*, “Neural tangent kernel: Convergence and generalization in neural networks,” in *Proc. NIPS*, 2018, pp. 8571–8580.
- [59] S. Arora *et al.*, “On exact computation with an infinitely wide neural net,” [Online]. Available: <https://arxiv.org/pdf/1904.11955.pdf>
- [60] M. Belkin *et al.*, “To understand deep learning we need to understand kernel learning,” [Online]. Available: <https://arxiv.org/pdf/1802.01396.pdf>
- [61] G. Georgiev, “Linear algebra and duality of neural networks,” [Online]. Available: <https://arxiv.org/pdf/1809.04711.pdf>
- [62] N. Tsapanos *et al.*, “Neurons with paraboloid decision boundaries for improved neural network classification performance,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 284–294, 2019.
- [63] R. Balestriero and R. Baraniuk, “A spline theory of deep learning,” in *Proc. ICML*, Stockholm, Sweden, 10–15 Jul 2018, pp. 374–383.
- [64] R. Balestriero and R. Baraniuk, “Mad max: Affine spline insights into deep learning,” [Online]. Available: <https://arxiv.org/pdf/1805.06576.pdf>
- [65] B. A. Richards *et al.*, “A deep learning framework for neuroscience,” *Nat Neurosci.*, vol. 22, pp. 1761–1770, 2019.
- [66] A. H. Marblestone *et al.*, “Toward an integration of deep learning and neuroscience,” *Front. Comput. Neurosci.*, vol. 10, 2016.
- [67] C. Pehlevan and D. B. Chklovskii, “Neuroscience-inspired online unsupervised learning algorithms: Artificial neural networks,” *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 88–96, 2019.
- [68] Z. Allen-Zhu *et al.*, “A convergence theory for deep learning via over-parameterization,” [Online]. Available: <https://arxiv.org/pdf/1811.03962.pdf>
- [69] D. Zou *et al.*, “Stochastic gradient descent optimizes over-parameterized deep ReLU networks,” [Online]. Available: <https://arxiv.org/pdf/1811.08888.pdf>
- [70] S. S. Du *et al.*, “Gradient descent finds global minima of deep neural networks,” [Online]. Available: <https://arxiv.org/pdf/1811.03804.pdf>
- [71] W. Zhu *et al.*, “LDMNet: Low dimensional manifold regularized neural networks,” in *Proc. CVPR*, 2018, pp. 2743–2751.
- [72] Z. J. Xu *et al.*, “Frequency principle: Fourier analysis sheds light on deep neural networks,” [Online]. Available: <https://arxiv.org/pdf/1901.06523.pdf>
- [73] T. Wiatowski *et al.*, “Energy propagation in deep convolutional neural networks,” *IEEE Trans. Inf. Theory*, pp. 4819–4842, 2018.
- [74] T. Wiatowski and H. Bölcskei, “A mathematical theory of deep convolutional neural networks for feature extraction,” *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1845–1866, 2018.
- [75] D. Rolnick and M. Tegmark, “The power of deeper networks for expressing natural functions,” in *Proc. ICLR*, 2018.
- [76] H. W. Lin *et al.*, “Why does deep and cheap learning work so well?” *J Stat Phys.*, vol. 168, no. 6, p. 1223–1247, 2017.
- [77] R. Eldan and O. Shamir, “The power of depth for feedforward neural networks,” in *Proc. COLT*, 2016, pp. 907–940.
- [78] T. Poggio *et al.*, “Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review,” *Int. J. Autom. Comput.*, vol. 14, pp. 503–519, 2017.
- [79] H. Lin and S. Jegelka, “ResNet with one-neuron hidden layers is a universal approximator,” [Online]. Available: <https://arxiv.org/pdf/1806.10909.pdf>
- [80] F. Fan *et al.*, “Slim, sparse, and shortcut networks,” Apr. 2020. [Online]. Available: <https://arxiv.org/pdf/1811.09003.pdf>
- [81] F.-L. Fan and G. Wang, “Duality of width and depth of neural networks,” [Online]. Available: <https://arxiv.org/pdf/2002.02515.pdf>
- [82] Z. Lu *et al.*, “The expressive power of neural networks: A view from the width,” in *Proc. NIPS*, 2017.
- [83] K. Kawaguchi, J. Huang, and L. P. Kaelbling, “Effect of depth and width on local minima in deep learning,” *Neural Comput.*, vol. 31, no. 7, pp. 1462–1498, 2019.
- [84] W. Hwang and A. Heinecke, “Un-rectifying non-linear networks for signal representation,” *IEEE Trans. Signal Process.*, vol. 68, 2020.
- [85] B. Hanin, “Universal function approximation by deep neural nets with bounded width and ReLU activations,” *Mathematics*, vol. 7, 2019.
- [86] P. Grohs, D. Perekrestenko, D. Elbrächter, and H. Boelskei, “Deep neural network approximation theory,” Jan. 2019. [Online]. Available: <https://arxiv.org/pdf/1901.02220.pdf>
- [87] I. Gühring *et al.*, “Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms,” *Analysis and Applications*, 2019.
- [88] P. Petersen and F. Voigtlaender, “Optimal approximation of piecewise smooth functions using deep ReLU neural networks,” *Neural Netw.*, vol. 108, pp. 296–330, Dec. 2018.
- [89] H. Boelskei *et al.*, “Optimal approximation with sparsely connected deep neural networks,” *SIAM J. Math. Data Sci.*, vol. 1, p. 8–45, 2019.
- [90] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Univ. Press, 2018.
- [91] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Univ. Press, 2019.
- [92] B. Sen, “A gentle introduction to empirical process theory and applications,” Apr. 2018. [Online]. Available: <http://www.stat.columbia.edu/~bodhi/Talks/Emp-Proc-Lecture-Notes.pdf>
- [93] G. Kutyniok *et al.*, “A theoretical analysis of deep neural networks and parametric PDEs,” 2019. [Online]. Available: <https://arxiv.org/pdf/1904.00377.pdf>
- [94] C. Schwab and J. Zech, “Deep learning in high dimension: Neural network expression rates for generalized polynomial chaos expansions in UQ,” *Analysis and Applications*, vol. 17, no. 1, pp. 19–55, 2019.
- [95] D. Bassett and O. Sporns, “Network neuroscience,” *Nat Neurosci.*, vol. 20, p. 353–364, 2017.
- [96] C. W. Lynn and D. S. Bassett, “The physics of brain network structure, function, and control,” [Online]. Available: <https://arxiv.org/pdf/1809.06441.pdf>
- [97] D. Yarotsky, “Error bounds for approximations with deep ReLU networks,” *Neural Networks*, vol. 94, p. 103–114, 2017.
- [98] M. Telgarsky, “Representation benefits of deep feedforward networks,” Sep. 2015. [Online]. Available: <https://arxiv.org/pdf/1509.08101.pdf>
- [99] M. Phuong and C. H. Lampert, “Towards understanding knowledge distillation,” in *Proc. ICML*, 2019, pp. 5142–5151.
- [100] M. R. U. Saputra *et al.*, “Distilling knowledge from a deep pose regressor network,” [Online]. Available: <https://arxiv.org/pdf/1908.00858.pdf>
- [101] G. E. Hinton *et al.*, “Distilling the knowledge in a neural network,” in *Proc. of NIPS Wksp. Deep Learning*, Montreal, QC, Canada, 2014.
- [102] E. Biglieri *et al.*, *MIMO Wireless Communications*. New York, NY, USA: Cambridge Univ. Press, 2007.
- [103] P. Schniter, “Low-complexity equalization of OFDM in doubly selective channels,” *IEEE Trans. Signal Process.*, vol. 52, no. 4, pp. 1002–1011, Apr. 2004.
- [104] Y. Xiaoyan *et al.*, “Doubly selective fading channel estimation in MIMO OFDM systems,” *Science in China*, vol. 48, 2005.
- [105] R.-A. Stoica and G. T. F. de Abreu, “6G: the wireless communications network for collaborative and AI applications,” [Online]. Available: <https://arxiv.org/pdf/1904.03413.pdf>
- [106] L. Hanzo and T. Keller, *OFDM and MC-CDMA: A Primer*. West Sussex, UK: Wiley, 2006.
- [107] J.-F. Hélaré *et al.*, “Multicarrier CDMA: A very promising multiple access scheme for future wideband wireless networks,” in *Proc. European Wksp. Integrated Radio Commun. Syst.*, 2002.
- [108] C. Xiao *et al.*, “A discrete-time model for triply selective MIMO Rayleigh fading channels,” *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1678–1688, Sep. 2004.
- [109] G. B. Giannakis, Y. Shen, and G. V. Karanikolas, “Topology identification and learning over graphs: Accounting for nonlinearities and dynamics,” *Proc. IEEE*, vol. 106, no. 5, pp. 787–807, 2018.
- [110] Y. Shen, B. Baingana, and G. B. Giannakis, “Tensor decompositions for identifying directed graph topologies and tracking dynamic networks,” *IEEE Trans. Signal Process.*, vol. 65, no. 14, pp. 3675–3687, 2017.
- [111] A. Ortega *et al.*, “Graph signal processing: Overview, challenges, and applications,” *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, 2018.
- [112] C. W. Lynn and D. S. Bassett, “The physics of brain network structure, function and control,” *Nat Rev Phys.*, vol. 1, p. 318–332, 2019.
- [113] A. Fornito, A. Zalesky, and E. T. Bullmore, *Fundamentals of Brain Network Analysis*. San Diego, CA, USA: Academic Press, 2016.
- [114] H. Montanelli and Q. Du, “New error bounds for deep ReLU networks using sparse grids,” *SIAM J. Math. Data Sci.*, vol. 1, pp. 78–92, 2019.
- [115] T. M. Getu, “Advanced RFI detection, RFI excision, and spectrum sensing: Algorithms and performance analyses,” Ph.D. dissertation, École de Technologie Supérieure (ÉTS), Montréal, QC, Canada, 2019.
- [116] H. Becker *et al.*, “Brain-source imaging: From sparse to tensor models,” *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 100–112, Nov. 2015.
- [117] R. Arora *et al.*, “Understanding deep neural networks with rectified linear units,” in *Proc. ICLR*, 2018.