Analyzing Upper Bounds on Mean Absolute Errors for Deep Neural Network-Based Vector-to-Vector Regression

Jun Qi , *Student Member, IEEE*, Jun Du , *Member, IEEE*, Sabato Marco Siniscalchi , *Senior Member, IEEE*, Xiaoli Ma , *Fellow, IEEE*, and Chin-Hui Lee , *Fellow, IEEE*

Abstract—In this paper, we show that, in vector-to-vector regression utilizing deep neural networks (DNNs), a generalized loss of mean absolute error (MAE) between the predicted and expected feature vectors is upper bounded by the sum of an approximation error, an estimation error, and an optimization error. Leveraging upon error decomposition techniques in statistical learning theory and non-convex optimization theory, we derive upper bounds for each of the three aforementioned errors and impose necessary constraints on DNN models. Moreover, we assess our theoretical results through a set of image de-noising and speech enhancement experiments. Our proposed upper bounds of MAE for DNN based vector-to-vector regression are corroborated by the experimental results and the upper bounds are valid with and without the "over-parametrization" technique.

Index Terms—Deep neural network, mean absolute error, vector-to-vector regression, non-convex optimization, image de-noising, speech enhancement.

I. INTRODUCTION

VECTOR-TO-VECTOR regression, also known as multi-variate regression, provides an effective way to find underlying relationships between input vectors and their corresponding output ones at the same time. The problems of vector-to-vector regression are of great interest in signal processing, wireless communication, and machine learning communities. For example, speech enhancement aims at finding a vector-to-vector mapping to convert noisy speech spectral vectors to the clean ones [1], [2]. Similarly, clean images can be extracted from the corrupted ones by leveraging upon image de-noising techniques [3]. Besides, wireless communication systems are designed to transmit local encrypted and corrupted codes to targeted receivers with decrypted information as correct as possible [4], [5]. Moreover, the vector-to-vector regression tasks are also commonly seen in ecological modeling, natural gas demand forecasting, and drug efficacy prediction domains [6].

The vector-to-vector regression can be theoretically formulated as follows: given a d-dimensional input vector space $\mathbb{R}^d$ and a measurable q-dimensional output vector space $\mathbb{R}^q$, the goal of vector-to-vector regression is to learn a functional relationship $f : \mathbb{R}^d \rightarrow \mathbb{R}^q$ such that the output vectors can approximate desirable target ones. The regression process is described as:

$$\mathbf{y} = f(\mathbf{x}) + \mathbf{e}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{y} \in \mathbb{R}^q$, $\mathbf{e}$ is an error vector, and f refers to the regression function to be exploited. To implement the regression function f, linear regression [7] was the earliest approach and several other methods, such as support vector regression [8] and decision tree regressions [9], were further proposed to enhance regression performance. However, deep neural networks (DNN) [10], [11] with multiple hidden layers offer a more efficient and robust solution to dealing with large-scale regression problems. For example, our previous experimental study [12] demonstrated that DNNs outperform shallow neural networks on speech enhancement. Similarly, auto-encoders with deep learning architectures can achieve better results on image de-noising [13].

Although most endeavors on DNN based vector-to-vector regression focus on the experimental gain in terms of mapping accuracy, the related theoretical performance of DNN has not been fully developed. Our recent work [14] tried to bridge the gap by analyzing the representation power of DNN based vector-to-vector regression and deriving upper bounds for different DNN architectures. However, those bounds particularly target experiments with consistent training and testing conditions, and they may not be adapted to the experimental tasks where unseen testing data are involved. Therefore, in this work, we focus on an analysis of the generalization power and investigate upper bounds on a generalized loss of mean absolute error (MAE) for DNN based vector-to-vector regression with mismatched training and testing scenarios. Moreover, we associate the required constraints with DNN models to attain the upper bounds.

The remainder of this paper is organized as follows: Section II highlights the contribution of our work and its relationship with the related work. Section III underpins concepts and notations

used in this work. Section IV discusses the upper bounds on MAE for DNN based vector-to-vector regression by analyzing the approximation, estimation, and optimization errors, respectively. Section V presents how to utilize our derived upper bounds to estimate practical MAE values. Section VI shows the experiments of image de-noising and speech enhancement to validate our theorems. Finally, Section VII concludes our work.

II. RELATED WORK AND CONTRIBUTION

The recent success of deep learning has inspired many studies on the expressive power of DNNs [15]–[18], which extended the original universal approximation theory on shallow artificial neural networks (ANNs) [19]–[23] to DNNs. As discussed in [24], the approximation error is tightly associated with the DNN expressive power. Moreover, the estimation error and optimization error jointly represent the DNN generalization power, which can be reflected by error bounds on the out-of-sample error or the testing error. The methods of analyzing DNN generalization power are mainly divided into two classes: one refers to algorithm-independent controls [25]–[27] and another one denotes algorithm-dependent approaches [28], [29]. In the class of algorithm-independent controls, the upper bounds for the estimation error are based on the empirical Rademacher complexity [30] for a functional family of certain DNNs. In practice, those approaches concentrate on techniques of how weight regularization affects the generalization error without considering advanced optimizers and the configuration of hyper-parameters. As for the algorithm-dependent approaches [28], [29], several theoretical studies focus on the "over-parametrization" technique [31]–[34], and they suggest that a global optimal point can be ensured if parameters of a neural network significantly exceed the amount of training data during the training process.

We notice that the generalization capability of deep models can also be investigated through the stability of the optimization algorithms. More specifically, an algorithm is stable if a small perturbation to the input does not significantly alter the output, and a precise connection between stability and generalization power can be found in [35], [36]. Besides, in [37]–[39], the authors investigate the stability and oscillations of various competitive neural networks from the perspective of equilibrium points. However, the analysis of the stability of the optimization algorithm is out of the scope of the present work, and we do not discuss it further in this study.

In this paper, the aforementioned issues are taken into account by employing the error decomposition technique [40] with respect to an empirical risk minimizer (ERM) [41], [42] using three error terms: an approximation error, an estimation error, and an optimization error. Then, we analyze generalized error bounds on MAE for DNN based vector-to-vector regression models. More specifically, the approximation error can be upper bounded by modifying our previous bound on the representation power of DNN based vector-to-vector regression [14]. The upper bound on the estimation error relies on the empirical Rademacher complexity [30] and necessary constraints imposed upon DNN parameters. The optimization error can be upper bounded by assuming γ-Polyak-Lojasiewicz (γ-PL) [43] condition under the

"over-parameterization" configuration for neural networks [44], [45]. Putting together all pieces, we attain an aggregated upper bound on MAE by summing the three upper bounds. Furthermore, we exploit our derived upper bounds to estimate practical MAE values in experiments of DNN based vector-to-vector regression.

We use image de-noising and speech enhancement experiments to validate the theoretical results in this work. Image de-noising is a simple regression task from $[0, 1]^d$ to $[0, 1]^d$, where the configuration of "over-parametrization" can be simply satisfied on datasets like MNIST [46]. Speech enhancement is another useful illustration of the general theoretical analysis because it is an unbounded conversion from $\mathbb{R}^d \to \mathbb{R}^q$. Although the "over-parametrization" technique could not be employed in the speech enhancement task due to a significantly huge amount of training data, we can relax the "over-parametrization" setup and solely assume the γ-PL condition to attain the upper bound for MAE. In doing so, the upper bound can be adopted in experiments of speech enhancement.

III. PRELIMINARIES

A. Notations

- $f \circ g$: The composition of functions f and g.
- $||\mathbf{v}||_p$: L_p norm of the vector $\mathbf{v}$.
- $\langle \mathbf{x}, \mathbf{y} \rangle$ and $\mathbf{x}^T\mathbf{y}$: Inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$.
- $[q]$: An integer set $\{1, 2, 3, \ldots, q\}$.
- ∇f: A first-order gradient of function f.
- $\mathbb{E}[X]$: Expectation over a random variable X.
- w_j: The j-th element in the vector $\mathbf{w}$.
- f_v: DNN based vector-to-vector regression function.
- g_u: Smooth ReLU function.
- $\mathbf{1}$: A vector of all ones.
- $\mathbf{1}_m$: Indicator vector of zeros but with the m-th dimension assigned to 1.
- $\mathbb{R}^d$: d-dimensional real coordinate space.
- $\mathbb{F}$: A family of the DNN based vector-to-vector functions.
- $\mathbb{L}$: A family of generalized MAE loss functions.

B. Numerical Linear Algebra

- Hölder's inequality: Let $p, q \geq 1$ be conjugate: $\frac{1}{p} + \frac{1}{q} = 1$. Then, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq ||\mathbf{x}||_p ||\mathbf{y}||_q, \tag{2}$$

with equality when $|y_i| = |x_i|^{p-1}$ for all $i \in [N]$. In particular, when $p = q = 2$, Hölder's inequality becomes the Cauchy-Shwartz inequality.

C. Convex and Non-Convex Optimization

- A function f is β-Lipschitz continuous if $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$||f(\mathbf{x}) - f(\mathbf{y})|| \leq \beta ||\mathbf{x} - \mathbf{y}||. \tag{3}$$

- Let f be a β-smooth function on $\mathbb{R}^n$. Then, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \nabla f(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) + \frac{\beta}{2} ||\mathbf{x} - \mathbf{y}||_2^2. \tag{4}$$

- A function f satisfies the γ-Polyak-Lojasiewicz (γ-PL) condition [43]. Then, $\forall \mathbf{x} \in \mathbb{R}^n$,

$$||\nabla f(\mathbf{x})||_2^2 \geq \gamma(f(\mathbf{x}) - f^*). \tag{5}$$

where f^* refers to the optimal value over the input domain. The γ-PL condition is a significant property for a non-convex function f because a global minimization can be attained from $\nabla f(\mathbf{x}) = 0$, and a local minimum point corresponds to the global one. Furthermore, if a function is convex and also satisfies γ-PL condition, the function is strongly convex.

- Jensen's inequality: Let X be a random vector taking values in a non-empty convex set $K \subset \mathbb{R}^n$ with a finite expectation $\mathbb{E}[X]$, and f be a measurable convex function defined over K. Then, $\mathbb{E}[X]$ is in K, $\mathbb{E}[f(X)]$ is finite, and the following inequality holds

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]. \tag{6}$$

D. Empirical Rademacher Complexity

Empirical Rademacher complexity [30] is a measure of how well the function class correlates with the Rademacher random value. The references [24], [47], [48] show that a function class with a larger empirical Rademacher complexity is more likely to be overfitted to the training data.

Definition 1: A Rademacher random variable takes on values ± 1 and is defined by the uniform distribution as:

$$\sigma_i = \begin{cases} 1, & \text{with probability } \frac{1}{2} \\ -1, & \text{with probability } \frac{1}{2}. \end{cases} \tag{7}$$

Definition 2: The empirical Rademacher complexity of a hypothesis space $\mathbb{H}$ of functions $h : \mathbb{R}^n \to \mathbb{R}$ with respect to N samples $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ is:

$$\hat{\mathcal{R}}_S(\mathbb{H}) := \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathbb{H}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i h(\mathbf{x}_i) \right], \tag{8}$$

where $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$ indicates a set of N Rademacher random variables.

Lemma 1 (Talagrand's Lemma [49]): Let $\Phi_1, \ldots, \Phi_N$ be L-Lipschitz functions and $\sigma_1, \ldots, \sigma_N$ be Rademacher random variables. Then, for any hypothesis space $\mathbb{H}$ of functions $h : \mathbb{R}^n \to \mathbb{R}$ with respect to N samples $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, the following inequality holds

$$\frac{1}{N} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathbb{H}} \sum_{i=1}^{N} \sigma_i (\Phi_i \circ h)(\mathbf{x}_i) \right] \leq \frac{L}{N} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathbb{H}} \sum_{i=1}^{N} \sigma_i h(\mathbf{x}_i) \right]$$

$$= L\hat{\mathcal{R}}_S(\mathbb{H}). \tag{9}$$

E. MAE and MSE

Definition 3: MAE measures the average magnitude of absolute differences between N predicted vectors $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and N actual observations $S^* = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$, which is related to L_1 norm ($|| \cdot ||_1$) and the

corresponding loss function is defined as:

$$\mathcal{L}_{MAE}(S, S^*) = \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{y}_i||_1. \tag{10}$$

Mean Squared Error (MSE) [50] denotes a quadratic scoring rule that measures the average magnitude of N predicted vectors $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and N actual observations $S^* = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$, which is related to L_2 norm ($|| \cdot ||_2$) and the corresponding loss function is shown as:

$$\mathcal{L}_{MSE}(S, S^*) = \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{y}_i||_2^2. \tag{11}$$

IV. UPPER BOUNDING MAE FOR DNN BASED VECTOR-TO-VECTOR REGRESSION

This section derives the upper bound on a generalized loss of MAE for DNN based vector-to-vector regression. We first discuss the error decomposition technique for MAE. Then, we upper bound each decomposed error, and attain an aggregated upper bound on MAE.

A. Error Decomposition of MAE

Based on the traditional error decomposition approach [49], [51], we generalize the technique to the DNN based vector-to-vector regression, where the smooth ReLU activation function, the regression loss functions, and their associated hypothesis space are separately defined in Definition 4.

Definition 4: A smooth vector-to-vector regression function is defined as $f_v^* : \mathbb{R}^d \to \mathbb{R}^q$, and a family of DNN based vector-to-vector functions is represented as $\mathbb{F} = \{f_v : \mathbb{R}^d \to \mathbb{R}^q\}$, where a smooth ReLU activation is given as:

$$g_u(x) = \lim_{t \to \infty} \frac{\ln(1 + \exp(tx))}{t}. \tag{12}$$

Moreover, we assume $\mathbb{L} = \{\mathcal{L}(f_v, f_v^*) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, f_v \in \mathbb{F}\}$ as the family of generalized MAE loss functions. For simplicity, we denote $\mathcal{L}(f_v, f_v^*)$ as $\mathcal{L}(f_v)$. Besides, we denote $\mathcal{D}$ as a distribution over $\mathbb{R}^d$.

The following proposition bridges the connection of Rademacher complexity between the family $\mathbb{L}$ of generalized MAE loss functions and the family $\mathbb{F}$ of DNN based vector-to-vector functions.

Proposition 1: For any sample set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ drawn i.i.d. according to a given distribution $\mathcal{D}$, the Rademacher complexity of the family $\mathbb{L}$ is upper bounded as:

$$\hat{\mathcal{R}}_S(\mathbb{L}) \leq \hat{\mathcal{R}}_S(\mathbb{F}), \tag{13}$$

where $\hat{\mathcal{R}}_S(\mathbb{F})$ denotes the empirical Rademacher complexity over the family $\mathbb{F}$, and it is defined as:

$$\hat{\mathcal{R}}_S(\mathbb{F}) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\frac{1}{N} \sup_{f_v \in \mathbb{F}} \sum_{i=1}^{N} (\sigma_i \mathbf{1})^T f_v(\mathbf{x}_i) \right]. \tag{14}$$

Proof: We first show that MAE loss function is 1-Lipschitz continuous. For two vectors $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^q$ and a fixed vector $\mathbf{y} \in$

$\mathbb{R}^q$, the MAE loss difference is

$$|\mathcal{L}(\mathbf{y}_1, \mathbf{y}) - \mathcal{L}(\mathbf{y}_2, \mathbf{y})|$$
$$= |\,\|\mathbf{y}_1 - \mathbf{y}\|_1 - \|\mathbf{y}_2 - \mathbf{y}\|_1\,| \qquad (15)$$
$$\leq \|\mathbf{y}_1 - \mathbf{y}_2\|_1 \qquad \text{(triangle inequality)}.$$

Since the target function f_v^* is given, $\mathcal{L}(f_v) \in \mathbb{L}$ is 1-Lipschitz. By applying Lemma 1, we obtain that

$$\hat{\mathcal{R}}_S(\mathbb{L}) = \frac{1}{N} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f_v \in \mathbb{F}} \sum_{i=1}^N \sigma_i \mathcal{L}(f_v(\mathbf{x}_i)) \right]$$

$$= \frac{1}{N} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f_v \in \mathbb{F}} \sum_{i=1}^N \sigma_i \mathcal{L}\left(\sum_{m=1}^q \langle \mathbf{1}_m, f_v(\mathbf{x}_i)\rangle \mathbf{1}_m \right) \right]$$

$$\leq \frac{1}{N} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f_v \in \mathbb{F}} \sum_{i=1}^N (\sigma_i \mathbf{1})^T f_v(\mathbf{x}_i) \right] = \hat{\mathcal{R}}_S(\mathbb{F}). \quad (16)$$

■

Since $\hat{\mathcal{R}}_S(\mathbb{F})$ is an upper bound of $\hat{\mathcal{R}}_S(\mathbb{L})$, we can utilize the upper bound on $\hat{\mathcal{R}}_S(\mathbb{L})$ to derive the upper bound for $\hat{\mathcal{R}}_S(\mathbb{F})$. Next, we adopt the error decomposition technique to attain an aggregated upper bound which consists of three error components.

Theorem 1: Let $\hat{\mathcal{L}} \in \mathbb{L}$ denote the loss function for a set of samples S drawn i.i.d. according to a given distribution $\mathcal{D}$, and define $\hat{f}_v \in \mathbb{F}$ as an ERM for $\hat{\mathcal{L}}$. For a generalized MAE loss function $\mathcal{L} \in \mathbb{L}$, $\epsilon > 0$, and $0 < \delta < 1$, there exists $f_v^\epsilon \in \mathbb{F}$ such that $\mathcal{L}(f_v^\epsilon) \leq \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v) + \epsilon$. Then, with a probability of δ, we attain that

$$\mathcal{L}(\hat{f}_v)$$

$$\leq \underbrace{\inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v)}_{\text{Approx. error}} + \underbrace{2\sup_{f_v \in \mathbb{F}} |\mathcal{L}(f_v) - \hat{\mathcal{L}}(f_v)|}_{\text{Estimation error}} + \underbrace{\mathcal{L}(f_v^\epsilon) - \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v)}_{\text{Optimization error}}$$

$$\leq \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v) + 2\hat{\mathcal{R}}_S(\mathbb{F}) + \epsilon. \qquad (17)$$

Proof:

$$\mathcal{L}(\hat{f}_v) = \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v) + \mathcal{L}(\hat{f}_v) - \mathcal{L}(f_v^\epsilon) + \mathcal{L}(f_v^\epsilon) - \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v)$$

$$\leq \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v) + \mathcal{L}(\hat{f}_v) - \mathcal{L}(f_v^\epsilon) + \epsilon$$

$$\leq \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v) + \mathcal{L}(\hat{f}_v) - \hat{\mathcal{L}}(\hat{f}_v) + \hat{\mathcal{L}}(f_v^\epsilon) - \mathcal{L}(f_v^\epsilon) + \epsilon$$

$$\leq \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v) + 2\sup_{f_v \in \mathbb{F}} |\mathcal{L}(f_v) - \hat{\mathcal{L}}(f_v)| + \epsilon.$$

Then, we continue to upper bound the term $2\sup_{f_v \in \mathbb{F}} |\mathcal{L}(f_v) - \hat{\mathcal{L}}(f_v)|$. We first define μ as the expected value of $\sup_{f_v \in \mathbb{F}} |\mathcal{L}(f_v) - \hat{\mathcal{L}}(f_v)|$, and then introduce the fact that

$$\mu = \mathbb{E}\left[\sup_{f_v \in \mathbb{F}} \left| \mathcal{L}(f_v) - \hat{\mathcal{L}}(f_v) \right| \right] \leq 2\hat{\mathcal{R}}_S(\mathbb{L}), \qquad (18)$$

which is justified by Lemma 2 in Appendix A. Then, for a small δ $(0 < \delta < 1)$, we apply the Hoeffding's bound [52] as follows

$$\mathrm{P}\left(2\sup_{f_v \in \mathbb{F}} \left| \mathcal{L}(f_v) - \hat{\mathcal{L}}(f_v) \right| \leq \nu \right) \geq 1 - 2\exp\left(-2N(\nu - \mu)^2 \right)$$

$$\geq 1 - 2\exp\left(-2N(\nu - 2\hat{\mathcal{R}}_S(\mathbb{L}))^2 \right)$$

$$= \delta,$$

which can derive ν as:

$$\nu = 2\hat{\mathcal{R}}_S(\mathbb{L}) + \sqrt{\frac{1}{2N} \ln\left(\frac{2}{1-\delta} \right)},$$

and we thus obtain that

$$2\sup_{f_v \in \mathbb{F}} \left| \mathcal{L}(f_v) - \hat{\mathcal{L}}(f_v) \right| \leq 2\hat{\mathcal{R}}_S(\mathbb{L}) + \sqrt{\frac{1}{2N} \ln\left(\frac{2}{1-\delta} \right)}.$$

Therefore,

$$\mathcal{L}(\hat{f}_v) \leq \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v) + \left(2\hat{\mathcal{R}}_S(\mathbb{L}) + \sqrt{\frac{1}{2N} \ln\left(\frac{2}{1-\delta} \right)} \right) + \epsilon$$

$$\leq \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v) + 2\hat{\mathcal{R}}_S(\mathbb{F}) + \sqrt{\frac{1}{2N} \ln\left(\frac{2}{1-\delta} \right)} + \epsilon$$

$$\approx \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v) + 2\hat{\mathcal{R}}_S(\mathbb{F}) + \epsilon \quad \text{(for sufficiently large } N\text{)}.$$

■

Next, the remainder of this section presents how to upper bound the approximation error, approximation error, and optimization error, respectively.

B. An Upper Bound for Approximation Error

The upper bound for the approximation error is shown in Theorem 2, which is based on the modification of our previous theorem for the representation power of DNN based vector-to-vector regression [14].

Theorem 2: For a smooth vector-to-vector regression target function $f_v^*: \mathbb{R}^d \to \mathbb{R}^q$, there exists a DNN $\bar{f}_v \in \mathbb{F}$ with $k(k \geq 2)$ modified smooth ReLU based hidden layers, where the width of each hidden layer is at least $d + 2$ and the top hidden layer has $n_k(n_k \geq d + 2)$ units. Then, we derive the upper bound for the approximation error as:

$$\inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v) = \|f_v^* - \bar{f}_v\|_1 = \mathcal{O}\left(\frac{q}{(n_k + k - 1)^{\frac{r}{d}}} \right), \quad (19)$$

where a smooth ReLU function is defined in Eq. (12), and r refers to the differential order of f_v.

The smooth ReLU function in Eq. (12) is essential to derive the upper bound for the optimization error. Since Theorem 2 is a direct result from Lemma 2 in [53] where the standard ReLU is employed and does not consider Barron's bound for activation functions [21], the smooth ReLU function can be flexibly utilized in Theorem 2 because it is a close approximation to the standard ReLU function. Moreover, Theorem 2 requires at

least $d + 2$ neurons for a d-dimensional input vector to achieve the upper bound.

C. An Upper Bound for Estimation Error

Since the estimation error in Eq. (17) is upper bounded by the empirical Rademacher complexity $\hat{\mathcal{R}}_S(\mathbb{F})$, we derive Theorem 3 to present an upper bound on $\hat{\mathcal{R}}_S(\mathbb{F})$. The derived upper bound is explicitly controlled by the constraints of weights in the hidden layers, inputs, and the number of training data. In particular, the constraint of L_1 norm is set to the top hidden layer, and L_2 norm is imposed on the other hidden layers.

Theorem 3: For a DNN based vector-to-vector mapping function $f_v(\mathbf{x}) = \mathbf{W}_k \circ g_u \circ \mathbf{W}_{k-1} \circ \cdots \circ \mathbf{W}_2 \circ g_u \circ \mathbf{W}_1(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^q$ with a smooth ReLU function g_u as in Eq. (12) and $\forall i \in [k]$, $\mathbf{W}_i$ being the weight matrix of the i-th hidden layer, we obtain an upper bound for the empirical Rademacher complexity $\hat{\mathcal{R}}_S(\mathbb{F})$ with regularized constraints of the weights in each hidden layer, and the L_2 norm of input vectors $\mathbf{x}$ is bounded by s.

$$2 \sup_{f_v \in \mathbb{F}} |\mathcal{L}(f_v) - \mathcal{L}(\hat{f}_v)| \le 2\hat{\mathcal{R}}_S(\mathbb{F}) \le \frac{2q\Lambda'\Lambda^{k-1}s}{\sqrt{N}}$$

$$\text{s.t.,} \quad ||\mathbf{W}_k(i,:)||_1 \le \Lambda', \forall i \in [q]$$
$$||\mathbf{W}_j(a,:)||_2 \le \Lambda, \forall j \in [k-1], a \in [n_j]$$
$$||\mathbf{x}||_2 \le s, \tag{20}$$

where $\mathbf{W}_j(m,n)$ is an element associated with the j-th hidden layer of DNN where m is indexed to neurons in the j-th hidden layer and n is pointed to units of the $(j-1)$-th hidden layer, and $\mathbf{W}_j(m,:)$ contains all weights from the m-th neuron to all units in the $(j-1)$-th hidden layer.

Proof: We first consider an ANN with one hidden layer of n neuron units with the smooth ReLU function g_u as Eq. (12), and also denote $\hat{\mathbb{F}}$ as a family of ANN based vector-to-vector regression functions. $\hat{\mathbb{F}}$ can be decompoed into the sum of q subspaces $\sum_{i=1}^q \hat{\mathbb{F}}_i$ and each subspace $\hat{\mathbb{F}}_m$ is defined as:

$$\hat{\mathbb{F}}_m = \left\{ \mathbf{x} \to \sum_{j=1}^n w_j g_u(\mathbf{u}_j^T \mathbf{x}) \cdot \mathbf{1}_m : ||\mathbf{w}||_1 \le \Lambda', ||\mathbf{u}_j||_2 \le \Lambda \right\},$$

where n is the number of hidden neurons, $\forall j \in [n]$, $\mathbf{w}$ and $\mathbf{u}_j$ separately correspond to $\mathbf{W}_2(m,:)$ and $\mathbf{W}_1(j,:)$ in Eq. (20). Given N data samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, the empirical Rademacher complexity of $\hat{\mathbb{F}}_m$ is bounded as:

$$\hat{\mathcal{R}}_S(\hat{\mathbb{F}}_m) = \frac{1}{N}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{||\mathbf{w}||_1 \le \Lambda', ||\mathbf{u}_j||_2 \le \Lambda} \sum_{i=1}^N \sigma_i \sum_{j=1}^n w_j g_u(\mathbf{u}_j^T \mathbf{x}_i) \right]$$

$$= \frac{1}{N}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{||\mathbf{w}||_1 \le \Lambda', ||\mathbf{u}_j||_2 \le \Lambda} \sum_{j=1}^n w_j \sum_{i=1}^N \sigma_i g_u(\mathbf{u}_j^T \mathbf{x}_i) \right]$$

$$\le \frac{\Lambda'}{N}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{||\mathbf{u}_j||_2 \le \Lambda} \max_{j \in [n]} \left| \sum_{i=1}^N \sigma_i g_u(\mathbf{u}_j^T \mathbf{x}_i) \right| \right] \quad \text{(Hölder's ineq.)}$$

$$= \frac{\Lambda'}{N}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{||\mathbf{u}||_2 \le \Lambda} \left| \sum_{i=1}^N \sigma_i g_u(\mathbf{u}^T \mathbf{x}_i) \right| \right].$$

$$\le \frac{\Lambda'}{N}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{||\mathbf{u}||_2 \le \Lambda} \left| \sum_{i=1}^N \sigma_i \mathbf{u}^T \mathbf{x}_i \right| \right] \quad \text{(c.f. Lemma 1)}$$

$$\le \frac{\Lambda\Lambda'}{N}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\| \sum_{i=1}^N \sigma_i \mathbf{x}_i \right\|_2 \right] \quad \text{(Cauchy-Schwartz ineq.)}$$

$$\le \frac{\Lambda\Lambda'}{N}\sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\| \sum_{i=1}^N \sigma_i \mathbf{x}_i \right\|_2^2 \right]} \quad \text{(Jensen's inequality).} \tag{21}$$

The last term in the inequality (21) can be further simplified based on the independence of σ_i s. Thus, we finally derive the upper bound as:

$$\hat{\mathcal{R}}_S(\hat{\mathbb{F}}_m) \le \frac{\Lambda\Lambda'}{N}\sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\| \sum_{i=1}^N \sigma_i \mathbf{x}_i \right\|_2^2 \right]}$$

$$= \frac{\Lambda\Lambda'}{N}\sqrt{\sum_{i,j=1}^N \mathbb{E}_{\boldsymbol{\sigma}}[\sigma_i \sigma_j](\mathbf{x}_i^T \mathbf{x}_j)}$$

$$= \frac{\Lambda\Lambda'}{N}\sqrt{\sum_{i=1}^N ||\mathbf{x}_i||_2^2} \quad \text{(independence of } \sigma_i\text{s)}$$

$$\le \frac{\Lambda\Lambda's}{\sqrt{N}}. \tag{22}$$

The upper bound for $\hat{\mathcal{R}}_S(\hat{\mathbb{F}})$ is derived based on the fact that for q families of functions $\hat{\mathbb{F}}_m, m \in [q]$, there is $\hat{\mathcal{R}}_S(\mathbb{F}) = \hat{\mathcal{R}}_S(\sum_{m=1}^q \hat{\mathbb{F}}_m) = \sum_{m=1}^q \hat{\mathcal{R}}_S(\hat{\mathbb{F}}_m)$, and thus

$$\hat{\mathcal{R}}_S(\hat{\mathbb{F}}) = \sum_{m=1}^q \hat{\mathcal{R}}_S(\hat{\mathbb{F}}_m) \le \frac{q\Lambda\Lambda's}{\sqrt{N}}, \tag{23}$$

which is an extension of the empirical Rademacher identities [49], which is demonstrated in Lemma 3 of Appendix A.

Then, for the family of DNNs $\mathbb{F}$ with k hidden layers activated by the smooth ReLU function, we iteratively apply Lemma 1 and end up attaining the upper bound as:

$$\hat{\mathcal{R}}_S(\mathbb{F})$$

$$= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\forall l, w_{j_l} \in \mathbb{U}} \sum_{m=1}^q \sum_{i=1}^N \sigma_i \sum_{j_k=1}^{n_k} w_{j_k} g_u(\cdots \sum_{j_1=1}^{n_1} w_{j_1} g_u(\mathbf{u}_j^T \mathbf{x}_i)) \right]$$

$$\le \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\forall l, w_{j_l} \in \mathbb{U}} \sum_{m=1}^q \sum_{i=1}^N \sigma_i \sum_{j_k=1}^{n_k} w_{j_k} \cdots \sum_{j_1=1}^{n_1} w_{j_1} \mathbf{u}_j^T \mathbf{x}_i \right]$$

$$\le \frac{q\Lambda'\Lambda^{k-1}s}{\sqrt{N}},$$

where $w_{j_1}, \ldots, w_{j_k}$ are selected from the hypothesis space $\mathbb{U} = \{w_{j_1}, \ldots, w_{j_k} : \sum_{j_k=1}^{n_k} |w_{j_k}| \leq \Lambda', \sqrt{\sum_{j_i=1}^{n_i} w_{j_i}^2} \leq \Lambda, \forall i \in [k-1]\}$. ∎

D. An Upper Bound for Optimization Error

Next, we derive an upper bound for the optimization error. A recent work [54] has shown that the γ-PL property can be ensured if neural networks are configured with the setup of the "over-parametrization" [45], which is induced from the two facts as follows:

- Neural networks can satisfy γ-PL condition, when the weights of hidden layers are initialized near the global minimum point [45], [55].
- As the neural network involves more parameters, the update of parameters moves less, and there exists a global minimum point near the random initialization [33], [34].

Thus, the upper bound on the optimization error can be tractably derived in the context of the γ-PL condition for the generalized MAE loss $\mathcal{L}(\cdot) \in \mathbb{L}$. Since the smooth ReLU function admits smooth DNN based vector-to-vector functions, which can lead to an upper bound on the optimization error as:

$$\epsilon = \mathcal{L}(f_v^\epsilon) - \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v) \leq \frac{\mu M^2 \beta}{2\gamma}. \quad (24)$$

To achieve the upper bound in Eq. (24), we assume that the stochastic gradient descent (SGD) algorithm can result in an approximately equal optimization error for both the generalized MAE loss $\mathcal{L}(\cdot) \in \mathbb{L}$ and the empirical MAE loss $\hat{\mathcal{L}}(\cdot) \in \mathbb{L}$.

More specifically, for two DNN based vector-to-vector regression functions $f_v^\epsilon \in \mathbb{H}$ and $\hat{f}_v^\epsilon \in \mathbb{H}$, we have that

$$\epsilon = \mathcal{L}(f_v^\epsilon) - \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v) \approx \hat{\mathcal{L}}(\hat{f}_v^\epsilon) - \inf_{f_v \in \mathbb{F}} \hat{\mathcal{L}}(f_v). \quad (25)$$

Thus, we focus on analyzing $\hat{\mathcal{L}}(f_v)$ because it can be updated during the training process. We assume that $\hat{\mathcal{L}}(f_v)$ is β-smooth with $||\nabla \hat{\mathcal{L}}(f_v)||_2 \leq M$ and it also satisfies the γ-PL condition from an early iteration t_0. Besides, the learning rate of SGD is set to μ.

Moreover, we define $f_{v,\mathbf{w}_t} \in \mathbb{F}$ as the function with an updated parameter $\mathbf{w}_t$ at the iteration t, and denote $f_{v,\mathbf{w}_*} \in \mathbb{F}$ as the function with the optimal parameter $\mathbf{w}_*$. The smoothness of $\hat{\mathcal{L}}(\cdot)$ implies that

$$\hat{\mathcal{L}}(f_{v,\mathbf{w}_{t+1}}) - \hat{\mathcal{L}}(f_{v,\mathbf{w}_t}) - \langle \nabla \hat{\mathcal{L}}(f_{v,\mathbf{w}_t}), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle$$
$$\leq \frac{\beta}{2} ||\mathbf{w}_t - \mathbf{w}_{t+1}||_2^2. \quad (26)$$

Then, we apply the SGD algorithm to update model parameters at the iteration t as:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \mu \nabla \hat{\mathcal{L}}(f_{v,\mathbf{w}_t}). \quad (27)$$

Next, we substitute $-\mu \nabla \hat{\mathcal{L}}(f_{v,\mathbf{w}_t})$ in Eq. (27) for $\mathbf{w}_{t+1} - \mathbf{w}_t$ in the inequality (26), we have that

$$\hat{\mathcal{L}}(f_{v,\mathbf{w}_{t+1}}) - \hat{\mathcal{L}}(f_{v,\mathbf{w}_t}) + \mu ||\nabla \hat{\mathcal{L}}(f_{v,\mathbf{w}_t})||_2^2$$
$$\leq \frac{\beta \mu^2}{2} ||\nabla \hat{\mathcal{L}}(f_{v,\mathbf{w}_t})||_2^2. \quad (28)$$

By employing the condition $||\nabla \hat{\mathcal{L}}(f_{v,\mathbf{w}_t})||_2^2 \leq M^2$, we further derive that

$$\hat{\mathcal{L}}(f_{v,\mathbf{w}_{t+1}}) - \hat{\mathcal{L}}(f_{v,\mathbf{w}_t}) + \mu ||\nabla \hat{\mathcal{L}}(f_{v,\mathbf{w}_t})||_2^2 \leq \frac{\mu^2 M^2 \beta}{2}. \quad (29)$$

Furthermore, we employ the γ-PL condition to Eq. (29) and obtain the inequalities as:

$$\hat{\mathcal{L}}(f_{v,\mathbf{w}_{t+1}}) - \hat{\mathcal{L}}(f_{v,\mathbf{w}_*})$$
$$\leq \left(\hat{\mathcal{L}}(f_{v,\mathbf{w}_t}) - \hat{\mathcal{L}}(f_{v,\mathbf{w}_*}) - \gamma\mu(\hat{\mathcal{L}}(f_{v,\mathbf{w}_t}) - \hat{\mathcal{L}}(f_{v,\mathbf{w}_*})) \right)$$
$$+ \frac{\mu^2 M^2 \beta}{2}$$
$$\leq (1 - \mu\gamma) \left(\hat{\mathcal{L}}(f_{v,\mathbf{w}_t}) - \hat{\mathcal{L}}(f_{v,\mathbf{w}_*}) \right) + \frac{\mu^2 M^2 \beta}{2}$$
$$\leq (1 - \mu\gamma)^2 \left(\hat{\mathcal{L}}(f_{v,\mathbf{w}_{t-1}}) - \hat{\mathcal{L}}(f_{v,\mathbf{w}_*}) \right)$$
$$+ \sum_{i=0}^{1} (1 - \gamma\mu)^i \frac{\mu^2 M^2 \beta}{2} \leq \cdots$$
$$\leq (1 - \mu\gamma)^{t-t_0+1} \left(\hat{\mathcal{L}}(f_{v,\mathbf{w}_{t_0}}) - \hat{\mathcal{L}}(f_{v,\mathbf{w}_*}) \right)$$
$$+ \sum_{i=0}^{t-t_0} (1 - \gamma\mu)^i \frac{\mu^2 M^2 \beta}{2}$$
$$\leq (1 - \mu\gamma)^{t-t_0+1} \left(\hat{\mathcal{L}}(f_{v,\mathbf{w}_{t_0}}) - \hat{\mathcal{L}}(f_{v,\mathbf{w}_*}) \right) + \frac{\mu M^2 \beta}{2\gamma}$$
$$\leq \exp(-\mu\gamma(t - t_0 + 1)) \left(\hat{\mathcal{L}}(f_{v,\mathbf{w}_{t_0}}) - \hat{\mathcal{L}}(f_{v,\mathbf{w}_*}) \right) + \frac{\mu M^2 \beta}{2\gamma}. \quad (30)$$

By connecting the optimization error in Eq. (24) to our derived Eq. (30), we further have that

$$\epsilon = \mathcal{L}(f_v^\epsilon) - \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v)$$
$$\approx \hat{\mathcal{L}}(f_v^\epsilon) - \inf_{f_v \in \mathbb{F}} \hat{\mathcal{L}}(f_v)$$
$$\leq \exp(-\mu\gamma(T + 1)) \left(\hat{\mathcal{L}}(f_{v,\mathbf{w}_0}) - \inf_{f_v \in \mathbb{F}} \hat{\mathcal{L}}(f_v) \right) + \frac{\mu M^2 \beta}{2\gamma}$$
$$\approx \frac{\mu M^2 \beta}{2\gamma}, \quad (31)$$

where $T = t - t_0$ and $f_{v,\mathbf{w}_0} \in \mathbb{F}$ denotes a function with an initial parameter $\mathbf{w}_0$. The inequality in Eq. (31) suggests that when the number of iterations T is sufficiently large, we eventually attain the upper bound as Eq. (24).

Remark 1: The "over-parametrization" condition becomes difficult to be configured in practice when large datasets have to be dealt with. In such cases, the upper bound on the optimization error cannot be always guaranteed, but we can relax the configuration of "over-parametrization" for DNNs and assume the γ-PL condition to derive the upper bound on the optimization error. In doing so, our proposed upper bound can be applied to more general DNN based vector-to-vector regression functions.

E. An Aggregated Bound for MAE

Based on the upper bounds for the approximation error, estimation error and optimization error, we can derive an upper bound for $\mathcal{L}(h_S^{ERM})$. Besides, the constraints as shown in Eq. (32), which arise from the upper bounds on the approximation, estimation and optimization errors, are necessary conditions to derive the upper bound (with a probability $\delta \in (0,1)$) as:

$$\mathcal{L}(\hat{f}_v) \leq \inf_{f_v \in \mathbb{F}} \mathcal{L}(f_v) + 2\hat{\mathcal{R}}_S(\mathbb{F}) + \epsilon$$

$$\leq \mathcal{O}\left(\frac{q}{(n_k + k - 1)^{\frac{r}{d}}}\right) + \frac{2q\Lambda'\Lambda^{k-1}s}{\sqrt{N}} + \frac{\mu M^2 \beta}{2\gamma}$$

s.t., Smooth ReLU: $\lim_{t \to +\infty} \frac{1}{t} \ln(1 + \exp(tx))$

Hidden Layers: $n_j \geq d + 2, \forall j \in [k]$

Regularization: $\|\mathbf{W}_k(i,:)\|_1 \leq \Lambda', \forall i \in [q]$

$\|\mathbf{W}_j(m,:)\|_2 \leq \Lambda, \forall j \in [k-1], m \in [n_j]$

Bounded Inputs: $\|\mathbf{x}\|_2 \leq s$

Over-parametrization: The number of parameters exceeds

the amount of training data.
(32)

Eq. (32) suggests that several hyper-parameters are required to derive the upper bound, which makes it difficult to be utilized in practice because the prior setup of μ, M, β and γ are strong assumptions in use. Section V discusses how to estimate MAE values of ANN or DNN based vector-to-vector regression in practical experiments. Besides, the term $\frac{q\Lambda'\Lambda^{k-1}s}{\sqrt{N}}$ in Eq. (32) may become arbitrarily large when a large k and $\Lambda > 1$ are concerned. Thus, we set Λ as 1 to ensure normalized weights of the first $k-1$ layers, and the amount of training data N could be large enough to ensure a small estimation error.

The configuration of "over-parametrization" requires that the number of model parameters exceeds the amount of training data such that the γ-PL condition can be guaranteed and consequently the upper bound on the optimization error can be attained. However, when the setup of "over-parametrization" cannot be strictly satisfied, the γ-PL condition does not always hold. Then, we can still assume the γ-PL condition to obtain the upper bound (24), which allows the derived upper bound applicable for more general DNN based vector-to-vector regression functions.

Remark 2: Our work employs MAE as the loss function instead of MSE for the following reasons: (i) MSE does not satisfy the Lipschitz continuity such that the inequality Eq. (16) cannot be guaranteed; (ii) The MAE loss function for vector-to-vector regression tasks can achieve better performance than MSE in experiments [56].

V. ESTIMATION OF THE MAE UPPER BOUNDS

MAE can be employed as the loss function for training an ANN or DNN based vector-to-vector regression function. In this section, we discuss how to make use of the theorems in Section IV to estimate MAE upper bounds for the vector-to-vector regression models in our experiments.

Proposition 2 provides an upper bound on MAE based on our theorem in Eq. (32), where c and b are two non-negative hyper-parameters to be estimated from the experimental MAE losses of the ANN based vector-to-vector regression. An ANN with the smooth ReLU activation function in Eq. (12) is a convex and smooth function, which implies that the local optimum point returned by the SGD algorithm corresponds to a global one. Then, the estimated hyper-parameters c and b can be used to estimate the MAE values of DNN-based vector-to-vector regression.

Proposition 2: For a smooth target function $\hat{f}_v : \mathbb{R}^d \to \mathbb{R}^q$, we use N training data samples to obtain a DNN based vector-to-vector regression function $f_v \in \mathbb{F}$ with k smooth ReLU based hidden layers ($k \geq 2$), where the width of each hidden layer is at least $d + 2$. Then, we can derive an upper bound for MAE as:

$$MAE(\hat{f}_v, f_v) \leq \frac{cq}{(n_k + k - 1)^{\frac{r}{d}}} + \frac{2q\Lambda'\Lambda^{k-1}s}{\sqrt{N}} + b, \quad (33)$$

where the hyper-parameters c and b are separately set as:

$$c = \frac{(MAE_1 - MAE_2)l_1^{r/d}l_2^{r/d}}{q(l_2^{r/d} - l_1^{r/d})}, \quad (34)$$

and

$$b = \max\left(MAE_1 - \frac{(MAE_1 - MAE_2)l_2^{r/d}}{l_2^{r/d} - l_1^{r/d}} - \frac{2q\Lambda's}{\sqrt{N}}, 0\right). \quad (35)$$

Note that MAE_1 and MAE_2 are two practical MAE loss values associated with two ANNs with hidden units l_1 and l_2, respectively.

Proof: For two ANNs with hidden layers with units l_1 and l_2, we set k to 2 and then estimate their corresponding MAE losses as:

$$\frac{cq}{l_1^{r/d}} + \frac{2q\Lambda's}{\sqrt{N}} + b = MAE_1, \quad (36)$$

$$\frac{cq}{l_2^{r/d}} + \frac{2q\Lambda's}{\sqrt{N}} + b = MAE_2, \quad (37)$$

which can result in hyper-parameters c and b. In particular, we substitute $\frac{\mu M^2 \beta}{2\gamma}$ for b in Eq. (32) and then subtract two sides of Eq. (36) by Eq. (37), which can result in Eq. (34). By replacing c in Eq. (36) with Eq. (34), we finally obtain Eq. (35). ∎

Compared with our previous approaches to estimating practical MAE values in [14] where the DNN representation power is mainly considered, our new inequality (33) arises from the upper bound on the DNN generalization capability such that it can be used to estimate MAE values in more general experimental settings.

VI. Experiments

A. Experimental Goals

Our experiments separately employ the DNN based vector-to-vector regression for both image de-noising and speech enhancement with particular attention to linking empirical results with our proposed theorems. Unlike our analysis of the representation power of the DNN based regression tasks in [14], this work focuses on the generalization capability of the DNN based vector-to-vector regression based on our derived upper bounds. More specifically, we employ the tasks of image de-noising and speech enhancement, where inconsistent noisy conditions are mixed to the clean training and testing data, to validate our theorems by comparing the estimated MAE upper bound (MAE_ B) with the practical ones.

Moreover, the image de-noising experiment corresponds to the "over-parametrization" setting in which the number of DNN parameters is much larger than the amount of training data, but we cannot set up the "over-parametrization" for speech enhancement tasks due to a significantly large amount of training data. However, we assume the γ-PL condition and evaluate our derived upper bounds on the speech enhancement tasks.

Therefore, our experiments of image de-noising and speech enhancement aim at verifying the following points:

- The estimated MAE upper bound (MAE_ B) matches with experimental MAE values.
- A deeper DNN structure corresponds to a lower approximation error (AE).
- A significantly small optimization error can be achieved if the "over-parametrization" configuration is satisfied. Otherwise, the optimization error could be large enough to dominate MAE losses, even if the γ-PL condition is assumed.

B. Experiments of Image De-noising

1) Data Preparation: This section presents the image de-noising experiments on the MNIST dataset [46]. The MNIST dataset consists of 60000 images for training and 10000 ones for testing. We added additive Gaussian random noise (AGRN), with mean 0 and variance 1, to both training and testing data. The synthesized noisy data were then normalized such that for each image the condition $||\mathbf{x}_{noisy}||_2 \leq 1$ is satisfied.

2) Experimental Setup: The DNN based vector-to-vector regression in the experiments followed a feed-forward neural network architecture, where the inputs were 784-dimensional feature vectors of the noisy images and the outputs were 784-dimensional features of either clean or enhanced images. The reference of clean image features associated with the noisy inputs was assigned to the top layer of DNN in the training process, but the top layer corresponds to the features of the enhanced images during the testing stage. Table I exhibits the structures of neural networks used in the experiments. In more details, the vector-to-vector regression model was first built based on ANN. The width of the hidden layer of ANN1 was

TABLE I
Model Structures for Various Vector-to-Vector Regression

Models	Structures (Input − Hidden_layers − Output)
ANN1	784-1024-784
ANN2	784-2048-784
DNN1	784-1024-1024-1024-2048-784
DNN2	784-1024-1024-1024-1024-2048-784

TABLE II
Hyper-Parameters for the Estimation of MAE Upper Bounds

l_1	l_2	N	r	ANN1_MAE	ANN2_MAE
1024	2048	6×10^4	1176	0.1318	0.1292

TABLE III
The Evaluation Results Under the AGRN Noise

Models	MAE	AE	EE	OE	MAE_B
DNN1	0.1278	0.0172	0.0261	0.1005	0.1438
DNN2	0.1263	0.0168	0.0261	0.1005	0.1434

set to 1024, which satisfies the constraint of the number of neurons in hidden layers based on both the inequality Eq. (32) ($d = 784, d + 2 = 786 < 1024$) and the "over-parametrization" ($784 \times 1024 = 802816 > 60000$) condition. Whereas, ANN2 had a width of 2048 neurons, which is twice more than ANN1. Next, we studied the DNN based vector-to-vector regression by increasing the number of hidden layers of DNN1. Specifically, DNN1 was equipped with 4 hidden layers with widths 1024-1024-1024-2048. Additional two hidden layers of width 1024 were further appended to DNN2, which brings an architecture with 6 hidden layers 1024-1024-1024-1024-1024-2048.

Moreover, the SGD optimizer with a learning rate of 0.02 and a momentum rate of 0.2 was used to update model parameters based on the standard back-propagation (BP) algorithm [57]. The weights of the $k - 1$ hidden layers were normalized by dividing the L_2 norm, which corresponds to the term Λ^{k-1} configured to 1 in the inequality Eq. (33). The weights of the top hidden layer were normalized by dividing the L_1 norm such that Λ' is set to 1. Besides, MAE was employed as the evaluation metric in our experimental validation because the MAE metric is directly connected to the objective loss function of MAE.

3) Experimental Results: We present our experimental results on the noisy MNIST dataset, where AGRN was added to the clean images. Table II shows the setup of hyper-parameters l_1, l_2, N, and r in Eq. (33) to estimate MAE_B. Table III exhibits our estimated MAE values are in line with the practical MAE ones. Specifically, DNN2 attains a lower MAE (0.1278 vs. 0.1263) than DNN1. Moreover, our estimated MAE_B score for DNN2 is also lower than that for DNN1, namely 0.1438 vs. 0.1434, which arises from the decreasing AE score for DNN2 with a deeper architecture. Since we keep Λ and Λ' equal to 1, estimation error (EE) and optimization error (OE) for both DNN1 and DNN2 share the same values. Furthermore, although the OE values are comparatively larger than AE and EE, they also stay at a small

level because of the "over-parametrization" technique adopted in our experiments.

C. Experiments of Speech Enhancement

1) Data Preparation: Our experiments of speech enhancement were conducted on the Edinburgh noisy speech database [58], where the noisy backgrounds of the training data are inconsistent with the testing ones. More specifically, clean utterances were recorded from 56 speakers including 28 males and 28 females from different accent regions of both Scotland and the United States. Clean material was randomly split into 23075 training, and 824 testing waveforms, respectively. The noisy training waveforms at four SNR values, 15 dB, 10 dB, 5 dB, and 0 dB, were obtained using the following noises: a domestic noise (inside a kitchen), an office noise (in a meeting room), three public space noises (cafeteria, restaurant, subway station), two transportation noises (car and metro) and a street noise (busy traffic intersection). In sum, there were 40 different noisy conditions for synthesizing many noisy training data (ten noises × four SNRs). In the noisy test set, the noise types included: a domestic noise (living room), an office noise (office space), one transport (bus) and two street noises (open area cafeteria and a public square), and SNR values included: 17.5 dB, 12.5 dB, 7.5 dB, and 2.5 dB. Thus, there were 20 different noisy conditions for creating the testing dataset.

2) Experimental Setup: The DNN based vector-to-vector regression for speech enhancement also followed the feed-forward ANN architecture, where the input was a normalized log-power spectral (LPS) feature vector [59], [60] of noisy speech, and the output was LPS feature vector of either clean or enhanced speech. The references of clean speech feature vectors associated with the noisy inputs were assigned to the top layer of DNN in the training process, but the top layer of DNN corresponds to the feature vectors of the enhanced speech during the testing phase. The smooth ReLU function in Eq. (12) was employed in the hidden nodes of the neural architectures assessed in this work, whereas a linear function was used at the output layer. To improve the subjective perception in the speech enhancement tasks, the global variance equalization [61] was applied to alleviate the problem of over-smoothing by correcting a global variance between estimated features and clean reference targets [62]. In the training stage, the BP algorithm was adopted to update the model parameters, and the MAE loss was used to measure the difference between a normalized LPS vector, and the reference one. Noise-aware training (NAT) [63] was also employed to enable non-stationary noise awareness, and feature vectors of 3-frame size were obtained by concatenating frames within a sliding window [64]. Moreover, the SGD optimizer with a learning rate of 1×10^{-3} and a momentum rate of 0.4 was used for the update of parameters. The weights of the first $k-1$ hidden layers are normalized by dividing the L_2 norm of each row of weights, which correspond to the term Λ^{k-1} equal to 1 in Eq. (32). Moreover, we set s in Eq. (32) as the maximum value of L_2 norm of the input, and assume Λ' in Eq. (32)

TABLE IV
MODEL STRUCTURES FOR VARIOUS VECTOR-TO-VECTOR REGRESSION

Models	Structures (Input − Hidden_layers − Output)
ANN1	771-800-257
ANN2	771-1600-257
DNN1	771-800-800-800-1600-257
DNN2	771-800-800-800-800-800-1600-257

TABLE V
HYPER-PARAMETERS FOR THE ESTIMATION OF MAE UPPER BOUNDS

l_1	l_2	N	r
800	1600	1.04×10^{10}	771
ANN1_MAE	ANN2_MAE	Λ'(ANN1)	Λ' (ANN2)
0.7409	0.7328	8.9543	10.1542

TABLE VI
THE MAE RESULTS ON THE EDINBURGH SPEECH DATABASE

Models	MAE	PESQ	AE	EE	OE	MAE_B
DNN1	0.7060	2.82	0.0161	0.0579	0.6496	0.7236
DNN2	0.6859	2.85	0.0081	0.0728	0.6315	0.7124

as the maximum value of $(||\mathbf{W}_k(1,:)||_1, \ldots, ||\mathbf{W}_k(q,:)||_1)$, which are different from the setup of image de-noising.

Table IV exhibits the architectures of neural networks used in our experiments of speech enhancement. Two ANN models (ANN1 and ANN2) were utilized to estimate the hyper-parameters in Eq. (33), which were then used to estimate the MAE values of DNN models based on Eq. (33).

Two evaluation metrics, namely MAE and Perceptual Evaluation of Speech Quality (PESQ) [65], were employed in our experimental validation. Different from the MAE metric, PESQ is an indirect evaluation which is highly correlated with speech quality. The PESQ score, which ranges from -0.5 to 4.5, is calculated by comparing the enhanced speech with the clean one. A higher PESQ score corresponds to a higher quality of speech perception. All of the evaluation results on testing datasets are listed in Tables VI.

3) Experimental Results: We now present our experimental results on the Edinburgh speech database. Table V shows that the parameters used in the experiments to estimate the upper bound based on the inequality (33). The experimental results as shown in Table VI are in line with those observed in the consistent noisy conditions. Specifically, DNN2 attains a lower MAE (0.6859 vs. 0.7060) and higher PESQ values (2.85 vs. 2.82) than DNN1. Moreover, the MAE_ B score for DNN2 is also lower than that for DNN1, namely 0.7124 vs. 0.7236. Furthermore, DNN2 owns a better representation power in terms of AE scores (0.0081 vs. 0.0161) and a better power generalization capability because of a lower (EE + OE) score. More significantly, the OE term actually is the key contributor to the MAE_ B score, which suggests that the MAE loss is primarily from OE, as expected. In fact, optimization plays an important role when it comes to training large neural architectures [1], [12], which in turn shows

that the proposed upper bounds are in line with current research efforts [29], [33], [44], [45] on the optimization strategies.

D. Discussions

The experimental results of the image de-noising and speech enhancement suggest that our proposed upper bounds on the generalized loss of MAE can tightly estimate the practical MAE values. Unlike our previous work on the analysis of the representation power, which is strictly constrained to consistent noisy environments, our MAE bounds aim at the generalization power of DNN based vector-to-vector regression and can be generalized to more general noisy conditions.

Experimental results are based on our aggregated bound in Eq. (32), and the related practical methods in Eq. (33). The decreasing AE scores of DNN2 correspond to Eq. (19), where a deeper depth k can lead to smaller AE values. In the meanwhile, Eqs. (34) and (35) suggest that a smaller EE is associated with a larger OE, which also corresponds to our estimated results. Furthermore, deeper DNN structures can result in a larger Λ', which slightly escalates the AE scores and also decreases OE values. With the setup of "over-parametrization" for neural networks in image de-noising experiments, OE can be lowered to a small scale compared to AE and EE. However, OE becomes much lager than AE and EE without the "over-parametrization" configuration in the speech enhancement tasks.

VII. CONCLUSION

This study focuses on the theoretical analysis of an upper bound on a generalized loss of MAE for DNN based vector-to-vector regression and corroborates the theorems with image de-noising and speech enhancement experiments. Our theorems start from decomposing a generalized MAE loss, which can be upper bounded by the sum of approximation, estimation, and optimization errors for the DNN based vector-to-vector functions. Our previously proposed bound on the representation power of DNNs can be modified to upper bound the approximation error, and a new upper bound based on the empirical Rademacher complexity is derived for the estimation error. Furthermore, the smooth modification of the ReLU function and the assumption of γ-PL conditions under the "over-parametrization" configuration for neural networks can ensure an upper bound on the optimization error. Thus, an aggregated upper bound for MAE can be derived by combining the upper bounds for the three errors.

Our experimental results of image de-noising and speech enhancement show that a deeper DNN with a broader width at the top hidden layer can contribute to better generalization capability in various noisy environments. The estimated MAE based on our related theorems can offer tight upper bounds for the experimental values in practice, which can verify our theorem of upper bounding MAE for DNN based vector-to-vector regression. Besides, our theories show that the MAE value mainly arises from the optimization error for well-regularized DNNs, and an "over-parametrization" for neural networks can ensure small optimization errors.

APPENDIX A

Lemma 2: Let $\hat{\mathcal{L}} \in \mathbb{L}$ denote the loss function for N samples $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ drawn i.i.d. according to a distribution $\mathcal{D}$. For a generalized MAE loss function $\mathcal{L} \in \mathbb{L}$, we have that

$$\mathbb{E}\left[\sup_{f_v \in \mathbb{F}} \left|\mathcal{L}(f_v) - \hat{\mathcal{L}}(f_v)\right|\right] \leq 2\hat{\mathcal{R}}_S(\mathbb{F}). \tag{38}$$

Proof: We utilize the method of symmetrization [66] to bound $\mathbb{E}[\sup_{f_v \in \mathbb{F}} |\mathcal{L}(f_v) - \hat{\mathcal{L}}(f_v)|]$. The symmetrization introduces a ghost dataset $S' = \{\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_N\}$ drawn i.i.d. from $\mathcal{D}$. Let $\hat{\mathcal{L}}' \in \mathbb{L}$ be the empirical risk with respect to the ghost dataset, and we assume $\mathcal{L}(f_v) = \mathbb{E}_{S'}[\hat{\mathcal{L}}'(f_v)]$. Assuming $\mathcal{L}(f_v) \geq \hat{\mathcal{L}}(f_v), \forall f_v \in \mathbb{F}$, we derive that

$$\mathbb{E}_S\left[\sup_{f_v \in \mathbb{F}} \left|\mathcal{L}(f_v) - \hat{\mathcal{L}}(f_v)\right|\right]$$

$$= \mathbb{E}_S\left[\sup_{f_v \in \mathbb{F}} \left(\mathcal{L}(f_v) - \hat{\mathcal{L}}(f_v)\right)\right]$$

$$= \mathbb{E}_S\left[\sup_{f_v \in \mathbb{F}} \left(\mathbb{E}_{S'}[\hat{\mathcal{L}}'(f_v)] - \hat{\mathcal{L}}(f_v)\right)\right]$$

$$\leq \mathbb{E}_S\left[\mathbb{E}_{S'}\left[\sup_{f_v \in \mathbb{F}} \left(\hat{\mathcal{L}}'(f_v) - \hat{\mathcal{L}}(f_v)\right)\right]\right]$$

$$\leq \mathbb{E}_{S,S'}\left[\sup_{f_v \in \mathbb{F}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i(\hat{\mathcal{L}}'(f_v(\mathbf{x}'_i)) - \hat{\mathcal{L}}(f_v(\mathbf{x}_i)))\right]$$

$$\leq 2\hat{\mathcal{R}}_S(\mathbb{F}),$$

where $\sigma_1, \sigma_2, \ldots, \sigma_N$ are Rademacher random variables. Similarly, the assumption of $\mathcal{L}(f_v) \leq \hat{\mathcal{L}}(f_v), \forall f_v \in \mathbb{F}$ also brings the same result. Thus, we finish the proof of Lemma 2. ∎

Lemma 3 (An extension of empirical Rademacher identities): Given any sample set $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, and hypothesis sets $\mathbb{F}_1$, $\mathbb{F}_2$, ..., $\mathbb{F}_q$ of functions $f_v^{(1)} \in \mathbb{F}_1$, $f_v^{(2)} \in \mathbb{F}_2$, ..., $f_v^{(q)} \in \mathbb{F}_q$ mapping from $\mathbb{R}^d$ to $\mathbb{R}^q$, we have that

$$\hat{\mathcal{R}}_S\left(\sum_{i=1}^{q} \mathbb{F}_i\right)$$

$$= \frac{1}{N}\mathbb{E}_\sigma\left[\sup_{f_v^{(1)} \in \mathbb{F}_1, \ldots, f_v^{(q)} \in \mathbb{F}_q} \sum_{i=1}^{N} \sigma_i\left(\sum_{j=1}^{q} f_v^{(j)}(\mathbf{x}_i)\right)\right]$$

$$= \frac{1}{N}\sum_{j=1}^{q} \mathbb{E}_\sigma\left[\sup_{f_v^{(1)} \in \mathbb{F}_1, \ldots, f_v^{(q)} \in \mathbb{F}_q} \sum_{i=1}^{N} \sigma_i f_v^{(j)}(\mathbf{x}_i)\right]$$

$$= \sum_{i=1}^{q} \hat{\mathcal{R}}_S(\mathbb{F}_i).$$

APPENDIX B

Fig. 1 illustrates both training and testing curves of MAE over epochs in our experiments of image de-noising and speech enhancement. The simulations suggest that DNN2 with deeper
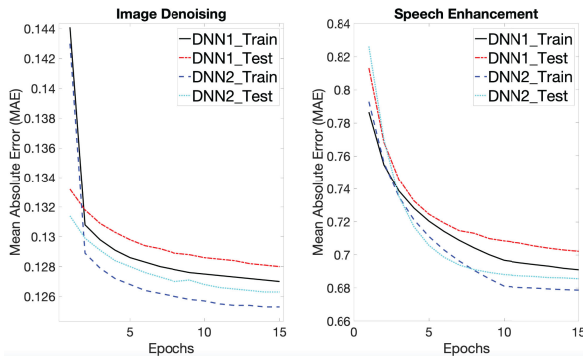
Fig. 1. Training and testing MAE curves over epochs in the two experiments.

architectures consistently achieves lower training and testing MAE values than DNN1 over epochs. When the update of model parameters gets converged, DNN2 finally attains the best performance on the two experiments.

REFERENCES

[1] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[2] J. Qi, H. Hu, Y. Wang, C. H. Yang, S. M. Siniscalchi, and C.-H. Lee, "Tensor-train network based tensor-to-vector regression for multi-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2020, pp. 7504–7508.

[3] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 341–349.

[4] M. Sánchez-Fernández, M. de Prado-Cumplido, J. Arenas-García, and F. Pérez-Cruz, "SVM multiregression for nonlinear channel estimation in multiple-input multiple-output systems," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2298–2307, Aug. 2004.

[5] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[6] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, "A survey on multi-output regression," *Math. Methods Appl. Sci.*, vol. 5, no. 5, pp. 216–233, 2015.

[7] P. Pearson, "Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia," *Philos. Trans. Roy. Soc. London*, vol. 187, pp. 253–318, 1896.

[8] V. Vapnik, S. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation, and signal processing," in *Proc. Advances Neural Inf. Process. Syst.*, 1997, pp. 281–287.

[9] R. M. Milstein, G. Burrow, L. Willkinson, and W. Kessen, "Prediction of screening decisions in a medical school admission process," in *Proc. Advances Neural Inf. Process. Syst.*, 1975, vol. 51, pp. 626–633.

[10] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[13] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[14] J. Qi, J. Du, S. M. Siniscalchi, and C.-H. Lee, "A theory on deep neural network based vector-to-vector regression with an illustration of its expressive power in speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 1932–1943, Dec. 2019.

[15] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 6231–6239.

[16] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," in *Proc. Conf. Learn. Theory*, 2016, pp. 907–940.

[17] S. Liang and R. Srikant, "Why deep neural networks for function approximation?" in *Proc. Int. Conf. Learn. Representation (ICLR)*, 2017.

[18] D. Rolnick and M. Tegmark, "The power of deeper networks for expressing natural functions," in *Proc. Int. Conf. Learn. Representations*, 2018.

[19] A. Kolmogoro, "On the representation of continuous functions of several variables as superpositions of functions of smaller number of variables," in *Proc. Sov. Math. Dokl*, 1956, vol. 108, pp. 179–182.

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

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

[22] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Mach. Learn.*, vol. 14, no. 1, pp. 115–133, 1994.

[23] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.

[24] J. Fan, C. Ma, and Y. Zhong, "A selective overview of deep learning," *Statist. Sci.*, accepted for publication.

[25] B. Neyshabur, R. Tomioka, and N. Srebro, "Norm-based capacity control in neural networks," in *Proc. Conf. Learn. Theory*, 2015, pp. 1376–1401.

[26] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 6240–6249.

[27] N. Golowich, A. Rakhlin, and O. Shamir, "Size-independent sample complexity of neural networks," in *Proc. 31st Conf. Learn. Theory*, 2018, vol. 75, pp. 297–299.

[28] S. Mei, A. Montanari, and P.-M. Nguyen, "A mean field view of the landscape of two-layer neural networks," *Nat. Acad. Sci.*, vol. 115, no. 33, pp. E7665–E7671, 2018.

[29] L. Chizat and F. Bach, "On the global convergence of gradient descent for over-parameterized models using optimal transport," in *Proc. Advances Neural Inf. Process. Syst.*, 2018, pp. 3036–3046.

[30] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, no. Nov, pp. 463–482, 2002.

[31] S. Du and J. Lee, "On the power of over-parametrization in neural networks with quadratic activation," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, 2018, vol. 80, pp. 1329–1338.

[32] B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro, "Towards understanding the role of over-parametrization in generalization of neural networks," in *Proc. Int. Conf. Learn. Representation (ICLR)*, 2019.

[33] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *Proc. 36th Intl. Conf. Mach. Learn. (ICML)*, 2019, vol. 97, pp. 242–252.

[34] Y. Li and Y. Liang, "Learning overparameterized neural networks via stochastic gradient descent on structured data," in *Proc. Advances Neural Inf. Process. Syst.*, 2018, pp. 8157–8166.

[35] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Mach. Learn. Res.*, vol. 2, pp. 499–526, 2002.

[36] C. H. Yang, J. Qi, P.-Y. Chen, X. Ma, and C.-H. Lee, "De-noising audio adversarial examples by U-net speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2020, pp. 3107–3111.

[37] A. Arbi and J. Cao, "Pseudo-almost periodic solution on time-space scales for a novel class of competitive neutral-type neural networks with mixed time-varying delays and leakage delays," *Neural Process Lett*, vol. 46, pp. 719–745, 2017.

[38] A. Arbi, A. A, and J. Cao, "Delta-differentiable weighted pseudo-almost automorphicity on time space scales for a novel class of high-order competitive neural networks with WPAA coefficients and mixed delays," *Neural Process Lett*, vol. 37, pp. 203–232, 2018.

[39] A. Arbi, "Dynamics of BAM neural networks with mixed delays and leakage time-varying delays in the weighted pseudo almost periodic on time-space scales," *Math. Methods Appl. Sci.*, vol. 41, no. 3, pp. 1230–1255, 2018.

[40] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, vol. 31. Berlin, Germany: Springer, 2013.

[41] V. Vapnik, *The Nature of Statistical Learning Theory*. Berlin, Germany: Springer, 2013.

[42] V. Vapnik, "Principles of risk minimization for learning theory," in *Proc. Advances Neural Inf. Process. Syst.*, 1992, pp. 831–838.

[43] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2016, pp. 795–811.

[44] Z. Allen-Zhu, Y. Li, and Y. Liang, "Learning and generalization in over-parameterized neural networks, going beyond two layers," in *Proc. Advances Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 6158–6169.

[45] S. Vaswani, F. Bach, and M. Schmidt, "Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron," in *Proc. Intl. Conf. Artif. Intell. Statist. (AISTATS)*, 2019.

[46] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.

[47] J. Zhu, B. R. Gibson, and T. T. Rogers, "Human rademacher complexity," in *Proc. Advances Neural Inf. Process. Syst.*, 2009, pp. 2322–2330.

[48] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[49] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2018.

[50] M. Paez and T. Glisson, "Minimum mean-squared-error quantization in speech PCM and DPCM systems," *IEEE Trans. Commun.*, vol. 20, no. 2, pp. 225–230, Apr. 1972.

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

[52] W. Hoeffding, "Probability inequalities for sums of bounded random variables," in *The Collected Works of Wassily Hoeffding*. Berlin, Germany: Springer, 1994, pp. 409–426.

[53] B. Hanin, "Universal function approximation by deep neural nets with bounded width and ReLU activations," *Mathematics*, vol. 7, no. 10, p. 992, 2019.

[54] R. Bassily, M. Belkin, and S. Ma, "On exponential convergence of SGD in non-convex over-parametrized learning," 2018, *arXiv:1811.02564*.

[55] Z. Charles and D. Papailiopoulos, "Stability and generalization of learning algorithms that converge to global optima," in *Proc. 35th Intl Conf. Mach. Learn. (ICML)*, 2018, vol. 80, pp. 745–754.

[56] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, no. 1, pp. 79–82, 2005.

[57] Y. Hirose, K. Yamashita, and S. Hijiya, "Back-propagation algorithm which varies the number of hidden units," *Neural Netw.*, vol. 4, no. 1, pp. 61–66, 1991.

[58] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. Speech Synthesis Workshop*, 2016, pp. 146–152.

[59] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2007, pp. 1–8.

[60] J. Qi, D. Wang, Y. Jiang, and R. Liu, "Auditory features based on gammatone filters for robust speech recognition," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2013, pp. 305–308.

[61] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, vol. 1, pp. I/9–I12.

[62] J. Qi, D. Wang, and J. Tejedor Noguerales, "Subspace models for bottleneck features," in *Proc. Interspeech*, 2013, pp. 1746–1750.

[63] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7398–7402.

[64] J. Qi, D. Wang, J. Xu, and J. Tejedor Noguerales, "Bottleneck features based on gammatone frequency cepstral coefficients," in *Proc. Interspeech*, 2013, pp. 1751–1755.

[65] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 749–752.

[66] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*, vol. 47. Cambridge, U.K.: Cambridge Univ. Press, 2018.

Jun Qi (Student Member, IEEE) is a Ph.D. candidate at School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta. Previously, he completed a graduate study in electrical engineering with an MSEE at the University of Washington, Seattle in 2017 and another Master of Engineering at Tsinghua University, Beijing in 2013. Also, he was a Research Intern in Deep Learning Technology Center (DLTC) at Microsoft Research, Redmond in 2017, a research intern at Tencent American AI Lab, Seattle in 2019, and a Graduate Research Intern at Mitsubishi Electric Research Laboratory (MERL), Boston in 2020.

Jun Du (Member, IEEE) received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC) in 2004 and 2009, respectively. From 2004 to 2009, he was with iFlytek Speech Lab of USTC. During the above period, he was an Intern at Microsoft Research Asia (MSRA), Beijing. In 2007, he worked as a Research Assistant for six months at the University of Hong Kong. From July 2009 to June 2010, he was with iFlytek Research. From July 2010 to January 2013, he joined MSRA as an Associate Researcher, working on handwriting recognition, OCR, and speech recognition. Since February 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing (NEL-SLIP) of USTC.

Sabato Marco Siniscalchi (Senior Member, IEEE) is a Professor at the University of Enna and affiliated with the Georgia Institute of Technology. He received his Laurea and Doctorate degrees in computer engineering from the University of Palermo, Italy, in 2001 and 2006, respectively. In 2006, he was a Post Doctoral Fellow at the Georgia Institute of Technology, USA. From 2007 to 2009, he joined the Norwegian University of Science and Technology (NTNU) as a Research Scientist. From 2010 to 2015, he was an Assistant Professor, first, and an Associate Professor, after, at the University of Enna. From 2017 to 2018, he joined as s Senior Speech Researcher the Siri Speech Group, Apple Inc., Cupertino CA, USA. He acted as an Associate Editor in IEEE/ACM Transactions on Audio, Speech and Language Processing (2015-2019). Dr. Siniscalchi is an elected member of the IEEE Speech and Language Technical Committee (2019-2021).

Xiaoli Ma (Fellow, IEEE) is a Professor in School of Electrical and Computer Engineering at Georgia Institute of Technology, Atlanta, GA, USA. She is an IEEE Fellow for her contribution to block transmissions over wireless fading channels. Her research interests are in the areas of signal processing for communications and networks, signal estimation algorithms, coding theory, wireless communication theory, and sensor and ad hoc networks. Ma is a Senior Area Editor for IEEE SIGNAL PROCESSING LETTERS and Elsevier *Digital Signal Processing* and has been an associate editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and Signal Processing Letters. She served as publication chair for the *IEEE GLOBECOM* 2013, local arrangements chair for IEEE GlobeSIP 2014, and general chair for the ACM International Conference on Underwater Networks and Systems 2015 and 2019. Her recent research interests rely on intelligent wireless communication, decentralized networks, and sensor networks.

Chin-Hui Lee (Fellow, IEEE) is a Professor at School of Electrical and Computer Engineering, Georgia Institute of Technology. Before joining academia in 2001, he had 20 years of industrial experience ending in Bell Laboratories, Murray Hill, New Jersey, as a Distinguished Member of Technical Staff and Director of the Dialogue Systems Research Department. Dr. Lee is a Fellow of the IEEE and a Fellow of ISCA. He has published over 400 papers and 30 patents, and was highly cited for his original contributions with an h-index of 66. He received numerous awards, including the Bell Labs President Gold Award in 1998. He won the SPS 2006 Technical Achievement Award for Exceptional Contributions to the Field of Automatic Speech Recognition. In 2012, he gave an ICASSP plenary talk on the future of speech recognition. In the same year, he was awarded the ISCA Medal in scientific achievement for pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition.