Deep Learning for Channel Estimation: Interpretation, Performance, and Comparison

Qiang Hu, Feifei Gao, *Fellow, IEEE*, Hao Zhang, Shi Jin, *Senior Member, IEEE*, and Geoffrey Ye Li, *Fellow, IEEE*

***Abstract*— Deep learning (DL) has emerged as an effective tool for channel estimation in wireless communication systems, especially under some imperfect environments. However, even with such unprecedented success, DL methods are often regarded as black boxes and are lack of explanations on their internal mechanisms, which severely limits their further improvement and extension. In this paper, we present preliminary theoretical analysis on DL based channel estimation for single-input multiple-output (SIMO) systems to understand and interpret its internal mechanisms. As deep neural network (DNN) with rectified linear unit (ReLU) activation function is mathematically equivalent to a piecewise linear function, the corresponding DL estimator can achieve universal approximation to a large family of functions by making efficient use of piecewise linearity. We demonstrate that DL based channel estimation does not restrict to any specific signal model and asymptotically approaches to the minimum mean-squared error (MMSE) estimation in various scenarios without requiring any prior knowledge of channel statistics. Therefore, DL based channel estimation outperforms or is at least comparable with traditional channel estimation, depending on the types of channels. Simulation results confirm the accuracy of the proposed interpretation and demonstrate the effectiveness of DL based channel estimation under both linear and nonlinear signal models.**

***Index Terms*— Explainable deep learning, input space partition, channel estimation, ReLU.**

I. Introduction

DEEP learning (DL) is making profound technological revolution to the concepts, patterns, methods and means of wireless communication systems [1]–[3]. There have been many interesting results for the physical layer (PHY) [4] or network layer of communications [5], including channel estimation [6], [7], channel state information (CSI) feedback compression [8], signal detection [9], and resource management [10], [11], etc. Among all DL applications to wireless communication systems, channel estimation is one of the most widely studied issues. The first attempt has been made in [7] to apply powerful DL methods to learn the characteristics of frequency selective wireless channels and combat the nonlinear distortion and interference for orthogonal frequency division multiplexing (OFDM) systems. In [12], a novel framework incorporates DL methods into massive multiple-input multiple-output (MIMO) systems to address direction-of-arrival (DoA) estimation and channel estimation problems. In [13], DL based channel estimation is extended to doubly selective channels and has numerically demonstrated better performance than the conventional estimators in many scenarios. In [14], the channel matrix is regarded as an image and a DL based image super-resolution and denoising technique is employed to estimate the channel. Furthermore, a sparse complex-valued neural network structure is proposed in [15] to tackle channel estimation in massive MIMO systems. Another branch of research attempts to establish a novel end-to-end deep neural network (DNN) architecture to replace all modules at the transmitter and at the receiver, respectively, instead of strengthening only certain modules [16]–[18].

Despite great success achieved by DL, the DNN embedded wireless communication system is generally considered as a black box for signal transmission/reception. Only numerical and experimental evaluations are available to demonstrate the powerful capability of DL in learning key functional components of wireless systems and there is nearly no analytical interpretation to confirm the advantages and disadvantages of DL methods when applied to communications. It is desired to understand why DL methods achieve astounding performance for a wide range of tasks for further performance improvement and extension to different environments. Moreover, the restrictions of DL methods to wireless communication systems are also very important for better understanding which scenarios are suitable for DL embedded communication systems.

Another important issue is how well newly emerged data-driven DL methods are compared to the traditional expert-designed algorithms in the field of wireless communications [19]. Impairments in PHY communications, such as noise, channel fading, interference, etc, have been thoroughly

Manuscript received May 16, 2020; revised October 11, 2020; accepted November 28, 2020. Date of publication December 9, 2020; date of current version April 9, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102401, in part by the National Natural Science Foundation of China under Grant 61831013 and Grant 61771274, and in part by the Shenzhen Special Projects for the Development of Strategic Emerging Industries under Grant 201806081439290640. The associate editor coordinating the review of this article and approving it for publication was C. Huang. *(Corresponding author: Feifei Gao.)*

Qiang Hu and Hao Zhang are with the Department of Electrical Engineering, Tsinghua University, Beijing 100084, China (e-mail: huq16@mails.tsinghua.edu.cn; haozhang@mail.tsinghua.edu.cn).

Feifei Gao is with the Institute for Artificial Intelligence, Tsinghua University, Beijing 100084, China, also with the State Key Laboratory of Intelligent Technologies and Systems, Tsinghua University, Beijing 100084, China, and also with the Beijing National Research Center for Information Science and Technology, Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: feifeigao@ieee.org).

Shi Jin is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: jinshi@seu.edu.cn).

Geoffrey Ye Li is with the Department of Electrical and Computer Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: geoffrey.li@imperial.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2020.3042074.

Digital Object Identifier 10.1109/TWC.2020.3042074

understood and addressed by well-established signal and coding theories from both practical and theoretical perspectives. It is yet unclear whether the black-box DL methods would be able to outperform the existing white-box approaches. In addition, the traditional ways of signal processing have been overturned by DL methods, in which satisfactory performance is still attainable in the absence of expert knowledge. Little research so far has dealt with how the DL methods learn from data and how the lack of expert knowledge affects the DL embedded communication systems.

In fact, there exits a wealth of literature addressing the complicated inner-workings of DNNs. The very first results have demonstrated the universal approximation of DNNs, that is, any continuous function defined on a compact set can be approximated at any precision using a DNN [20], [21]. As DNNs with rectified linear units (ReLU DNNs) become increasingly popular in recent years, the focus of study has been shifted to analyze the powerful capability of ReLU DNNs at function representation [22]–[24] and ReLU DNNs are also proved to be universal approximators to a large family of functions [25]. However, little evidence exits in previous research to guarantee that DNNs are really capable to rival traditional signal processing methods in communication systems despite the universal approximation. More theoretical support should be provided to verify the effectiveness of DL embedded communication systems.

Recently, more and more research has indicated that DL methods are particularly suited to channel estimation and it has become more common to deploy ReLU DNNs into communication systems. The practical success of ReLU DNNs calls for comprehensive understanding of their behavior on estimating channels to provide guidance and inspiration for the further exploitation on DL based estimation theory. In this paper, we present an initial attempt on interpreting DL for channel estimation in single-input multiple-output (SIMO) systems based on fully-connected ReLU DNNs. Our contributions are listed as follows:

- We analyze and compare the performance of the DL based channel estimation with the conventional methods, i.e., least-squared (LS) and linear minimum mean-squared error (LMMSE) estimators. We demonstrate that the DL estimator built on ReLU DNNs can well approximate the minimum mean-squared error (MMSE) estimator in the asymptotic limit of many training samples.
- We prove that the rate of convergence of the DL estimator to the MMSE estimator scales polynomially fast with the size of training samples. Such a result shows the effectiveness of the DL estimator for channel estimation.
- We demonstrate that the DL estimator experiences serious performance degradation and even fails to provide reliable estimates if the statistics of training data mismatch the deployed environments.

The rest of this paper is organized as follows. The system model and the traditional channel estimation methods are introduced in Section II. The DL based channel estimation is analyzed in Section III. Robustness of the DL based channel estimation to mismatched training data is presented in Section IV. Simulation results are provided in Section V followed by the conclusions in Section VI.

Notations: We use lowercase letters and capital letters in boldface to denote vectors and matrices, respectively. The positive integer set and real number set are denoted by $\mathbb{N}$ and $\mathbb{R}$, respectively. $\mathbf{I}_M$ denotes the $M \times M$ identity matrix. Notation $(\cdot)^T$ represents the transpose of a matrix or a vector, respectively. $\mathbb{E}\{\cdot\}$ denotes the expectation, $\mathrm{tr}\{\cdot\}$ denotes the trace of a matrix, and $\mathrm{vec}(\cdot)$ denotes the vectorization of a matrix. The cardinality of a set is denoted by $|\cdot|$. Notations $\|\cdot\|_2$ and $\|\cdot\|_\infty$ represent the 2-norm and supremum-norm of a vector or a matrix, respectively. Notation $\exp(\cdot)$ denotes a exponential function of e. Notation $\lceil\cdot\rceil$ represents the ceiling of a real number.

II. System Model and Traditional Channel Estimation

In this section, we first introduce the SIMO communication system for channel estimation and then present the traditional channel estimation methods.

A. System Model

Consider a SIMO communication system with d antennas at the base station (BS) and a single antenna at the user side. Assume the uplink channel is with block fading, that is, channel parameters are fixed within a block but vary from one to another. The traditional way of estimating channels at the BS is to use uplink pilot. Let τ be the transmitted pilot symbol with $|\tau|^2 = 1$. The received symbol at the BS can be represented by the following $d \times 1$ vector

$$\mathbf{x} = \tau\mathbf{h} + \mathbf{n}, \tag{1}$$

where $\mathbf{h}$ denotes the $d \times 1$ random channel vector between the user and the BS and $\mathbf{n}$ is the $d \times 1$ white Gaussian noise vector with zero-mean and element-wise variance σ_n^2. We assume that the channel vector $\mathbf{h}$ is with zero mean and covariance matrix $\mathbf{\Xi} = \mathbb{E}\{\mathbf{h}\mathbf{h}^T\}$.[1]

B. Traditional Channel Estimation

The goal of channel estimation is to extract channel vector $\mathbf{h}$ from received signal vector $\mathbf{x}$ as accurately as possible. The traditional estimation methods are based on the signal model in (1).

1) LS Channel Estimator: From (1), the LS estimate of $\mathbf{h}$ can be expressed as [26]

$$\mathbf{h}_{\mathrm{LS}} = \frac{1}{\tau}\mathbf{x} = \mathbf{h} + \frac{1}{\tau}\mathbf{n}, \tag{2}$$

and the corresponding mean-squared error (MSE) is

$$J_{\mathrm{LS}} = \mathbb{E}\{\|\mathbf{h} - \mathbf{h}_{\mathrm{LS}}\|_2^2\} = \frac{d}{1/\sigma_n^2}. \tag{3}$$

As shown in (3), the performance of the LS estimator is inversely proportional to the signal-to-noise ratio (SNR) defined as $1/\sigma_n^2$.

[1] In general, the complex valued signals would decompose into real values before inputting to ReLU DNNs. For convenience, we assume that both of input signals and channels are real values.

2) LMMSE Channel Estimator: The LMMSE estimator exploits the signal model in (1) and channel statistics to obtain the estimation, which can be expressed as [26]

$$\mathbf{h}_{\mathrm{LMMSE}} = \boldsymbol{\Xi}\left(\boldsymbol{\Xi} + \frac{\sigma_n^2}{\tau^2}\mathbf{I}_d\right)^{-1}\mathbf{h}_{\mathrm{LS}}. \tag{4}$$

Then, the MSE of the LMMSE estimator is computed as

$$J_{\mathrm{LMMSE}} = \mathrm{tr}\left\{\boldsymbol{\Xi}\left(\mathbf{I}_d + \frac{1}{\sigma_n^2}\boldsymbol{\Xi}\right)^{-1}\right\} \leq J_{\mathrm{LS}}. \tag{5}$$

3) MMSE Channel Estimator: The MMSE estimator can be expressed as [26]

$$\mathbf{h}_{\mathrm{MMSE}} = \mathbb{E}\{\mathbf{h}|\mathbf{x}\} \tag{6}$$

and is optimal under the criterion of minimizing the MSE. Generally, the MMSE estimator is different from the LMMSE estimator. Only in some special cases, we have $\mathbf{h}_{\mathrm{MMSE}} = \mathbf{h}_{\mathrm{LMMSE}}$ if $\mathbf{x}$ and $\mathbf{h}$ are joint Gaussian distributed for linear models, such as in (1). Therefore, the LS and LMMSE estimators can well address the channel estimation for linear models as in (1). However, the estimation performance of the LS and LMMSE estimators degrades significantly for nonlinear models since both of them are linear.

The LMMSE estimator tends to be more accurate by utilizing channel statistics but is sensitive to the imperfection of channel statistics. On the contrary, the LS estimator is easy to implement due to no prior requirement on channel statistics, but such simplicity is at cost of relatively low accuracy.

Recently, the DL estimator has emerged as a promising alternative to address channel estimation in wireless communication systems. The excellent generalization ability and powerful learning capacity of the DL estimator make it a powerful tool for channel estimation in imperfect and interference corrupted systems.

III. Analysis on DL Based Channel Estimation

Though DL based channel estimation has shown excellent performance in various communication systems, it has seldom been analyzed from a theoretical perspective. In this section, we provide preliminary theoretical analysis on the performance of the DL based channel estimation via statistical learning theory. Specifically, we demonstrate the DL estimator asymptotically approaches to the MMSE estimator as the number of training sample increases.

A. Basic Setting of DL Channel Estimator

Consider a DL estimator $\mathcal{D}$ with a fully-connected ReLU DNN. The input and output of $\mathcal{D}$ are denoted by $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{h} \in \mathbb{R}^d$, respectively. In the subsequent discussion, we will denote

$$\mathcal{Z} = \{(\mathbf{x}_m, \mathbf{h}_m) \,|\, \mathbf{x}_m, \mathbf{h}_m \in \mathbb{R}^d,\, m = 1, \ldots, |\mathcal{Z}|\} \tag{7}$$

as input-output sample set.

The underlying DNN of $\mathcal{D}$ consists of the ReLU activation function, $\varphi(x) = \max\{0, x\}$, $l \in \mathbb{N}$ hidden layers, and the neuron assignment $\mathbf{d} = (d_0, d_1, \ldots, d_l, d_{l+1}) \in \mathbb{N}^{l+2}$ with $d_0 = d_{l+1} = d$. The number of hidden layers l is the depth of $\mathcal{D}$. The width and size of $\mathcal{D}$ are denoted by $\max\{d_1, \ldots, d_l\}$ and $\sum_{i=1}^{l} d_i$, respectively.

Let

$$\Theta = \left\{\boldsymbol{\theta} = \left(\mathrm{vec}(\mathbf{W}_0), \mathbf{b}_0, \ldots, \mathrm{vec}(\mathbf{W}_l), \mathbf{b}_l\right) \in \mathbb{R}^{d_u}\right\} \tag{8}$$

be the set of all parameters of $\mathcal{D}$, where $d_u = \sum_{i=0}^{l} d_{i+1} \times (d_i + 1)$, $\mathbf{W}_i \in \mathbb{R}^{d_{i+1} \times d_i}$ and $\mathbf{b}_i \in \mathbb{R}^{d_{i+1}}$ are the weight matrix and the bias vector of the i-th layer for $i \in \{0, \ldots, l\}$.

For a fixed network structure $\mathbf{d}$, the underlying function that represented by $\mathcal{D}$ can be expressed as

$$\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}) = \mathcal{A}_l \circ \varphi_{d_l} \circ \mathcal{A}_{l-1} \circ \varphi_{d_{l-1}} \circ \cdots \circ \varphi_{d_1} \circ \mathcal{A}_0(\mathbf{x}), \tag{9}$$

where $\mathcal{A}_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i+1}}$ is the affine transformation corresponding to weight $\mathbf{W}_i$ and bias $\mathbf{b}_i$, $\varphi_{d_i} : \mathbb{R}^{d_i} \to \mathbb{R}^{d_i}$ is the entry-wise ReLU activation function, and $\circ$ denotes the function composition. The goal of the DL estimator is to optimize $\boldsymbol{\theta}$ in order to approximate the MMSE estimator for given training sample set $\mathcal{Z}$ and network architecture $\mathbf{d}$.

B. Internal Mechanism of DL Channel Estimator

Since ReLU function is piecewise linear, the neurons in $\mathcal{D}$ consist of only two states: with zero output or replicating input. When $\boldsymbol{\theta}$ is fixed, all the possible activation patterns of neurons in $\mathcal{D}$ can be represented by a set $\mathcal{K} \subseteq \{0, 1\}^{\bar{d}}$, where $\bar{d} = \sum_{i=1}^{l} d_i$ is a total number of neurons in $\mathcal{D}$ and each element in $\mathcal{K}$ is a $\bar{d}$-dimensional vector with its entries either 0 or 1. It is obvious that $|\mathcal{K}|$ is upper bounded by $2^{\bar{d}}$, i.e, $|\mathcal{K}| \leq 2^{\bar{d}}$. Similar to [23], the input space of a ReLU DNN is partitioned into different linear regions according to the corresponding activation patterns so that the ReLU DNN turns into a linear mapping in each region. Denote $\mathcal{X}$ as the input space and $\mathcal{X}_k$ as the input region within $\mathcal{X}$ corresponding to the k-th activation pattern. It is obvious that[2]

$$\mathcal{X}_k \subseteq \mathcal{X},\ k = 1, \ldots, K = |\mathcal{K}|, \quad \mathcal{X} = \cup_{k=1}^{K} \mathcal{X}_k. \tag{10}$$

Let $\tilde{\mathbf{x}}_i = [x_{i,1}, \ldots, x_{i,d_i}]^T$ be the output of the i-th layer with $\tilde{\mathbf{x}}_0 = \mathbf{x}$. For any input $\mathbf{x} \in \mathcal{X}_k$, $\mathcal{A}_i(\tilde{\mathbf{x}}_i)$ in (9) is computed as

$$\mathcal{A}_i(\tilde{\mathbf{x}}_i) = \begin{cases} \mathbf{W}_0\mathbf{x} + \mathbf{b}_0, & i = 0, \\ \tilde{\mathbf{W}}_i\mathcal{A}_{i-1}(\tilde{\mathbf{x}}_{i-1}) + \mathbf{b}_i, & i \geq 1, \end{cases} \tag{11}$$

where $\tilde{\mathbf{W}}_i = \mathbf{W}_i\boldsymbol{\Lambda}_i$ and $\boldsymbol{\Lambda}_i$ is an $\mathbb{R}^{d_i \times d_i}$ diagonal matrix with the diagonal elements either 0 or 1. Note that $\boldsymbol{\Lambda}_0 = \mathbf{I}_d$ and $\tilde{\mathbf{W}}_0 = \mathbf{W}_0\boldsymbol{\Lambda}_0$. Moreover, the diagonal elements of $\boldsymbol{\Lambda}_i$ correspond to the activation pattern of neurons at the i-th layer with their values either 0 or 1. Since all inputs $\mathbf{x} \in \mathcal{X}_k$ have the same activation pattern, the set $\{\boldsymbol{\Lambda}_i\}_{i=0}^{l}$ is fixed. By recursively expanding $\mathcal{A}_i(\tilde{\mathbf{x}}_i)$ layer by layer, we can further express $\mathcal{A}_i(\tilde{\mathbf{x}}_i)$ as

$$\begin{aligned} \mathcal{A}_i(\tilde{\mathbf{x}}_i) &= \prod_{j=0}^{i} \tilde{\mathbf{W}}_j\mathbf{x} + \sum_{j=0}^{i-1}\left(\prod_{p=0}^{j} \tilde{\mathbf{W}}_{i-p}\right)\mathbf{b}_{i-1-j} + \mathbf{b}_i \\ &= \hat{\mathbf{W}}_i\mathbf{x} + \hat{\mathbf{b}}_i, \end{aligned} \tag{12}$$

where $\hat{\mathbf{W}}_i = \prod_{j=0}^{i} \tilde{\mathbf{W}}_j$ is the equivalent weight matrix with respect to (w.r.t) the input $\mathbf{x}$ and

[2]These linear regions are not necessarily disjoint.

$\hat{\mathbf{b}}_i = \sum_{j=0}^{i-1} \left(\prod_{p=0}^{j} \tilde{\mathbf{W}}_{i-p} \right) \mathbf{b}_{i-1-j} + \mathbf{b}_i$ is the sum of the remaining terms. Therefore, $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})$ turns into an affine function for $\mathbf{x} \in \mathcal{X}_k$ and can be expressed as

$$\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{f}_{\mathcal{X}_k}(\mathbf{x}) = \mathbf{W}_{\mathcal{X}_k}\mathbf{x} + \mathbf{b}_{\mathcal{X}_k}, \tag{13}$$

where $\mathbf{W}_{\mathcal{X}_k} = \hat{\mathbf{W}}_l$ and $\mathbf{b}_{\mathcal{X}_k} = \hat{\mathbf{b}}_l$.

Remark 1:

- Generally, most of current state-of-the-art DNNs used by the DL embedded communication systems are based on fully-connected ReLU DNNs. Nearly all these structures involve the use of ReLU activation functions and their network layers, e.g., convolutional layers, can be regarded as variant types of dense layer with weights regularized. Therefore, we choose fully-connected ReLU DNNs as an example to illustrate how the performance of channel estimation is affected by the introduction of DNNs. It is an interesting topic for future research by considering other elaborately designed ReLU DNNs.
- If $\boldsymbol{\theta}$ is fixed, the inputs belonging to the same region have the same form of $\mathbf{f}_{\mathcal{X}_k}(\mathbf{x})$ and correspond to the same activation pattern. From (13), $\mathbf{f}_{\mathcal{X}_k}(\mathbf{x})$ is an affine function for $\mathbf{x} \in \mathcal{X}_k$, and therefore $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})$ is an $\mathbb{R}^d \to \mathbb{R}^d$ piecewise linear function over $\mathcal{X}$.
- The DL estimator has the simplicity and stability of a linear estimator, but also enables remarkable flexibility through the piecewise linear property.
- The DL estimator can model a large family of nonlinear functions by dynamically adjusting the partitioned regions and is more general and flexible compared to the LS and LMMSE estimators. The piecewise linear property of $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})$ is a critical step to interpret DL based channel estimation and will be used in the later analysis.

C. Performance Assessment of DL Channel Estimator

Different from the LS and LMMSE estimators, it is difficult to derive an explicit analytical form of the DL estimate as well as the corresponding MSE. Hence, the performance assessment of the DL estimator and the comparison to the LS and LMMSE estimators are not straightforward. Nevertheless, the DL estimator can approximate to a large family of functions due to its piecewise linear property. We can leverage the universal approximation of the DL estimator to assess its estimation performance and derive its rate of convergence to the MMSE estimator.

Let $\mathbf{f}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^d$ denote a channel estimator for $\mathbf{h}$ and ℓ_2 be the space of $\mathbf{f}(\mathbf{x})$ with finite 2-norm defined as

$$\|\mathbf{f}(\mathbf{x})\|_2 = \left[\sum_{i=1}^{d} \mathbb{E}\left\{ f_i(\mathbf{x})^2 \right\} \right]^{1/2} < +\infty, \tag{14}$$

where $f_i(\mathbf{x})$ is the i-th entry of $\mathbf{f}(\mathbf{x})$.

Define

$$J(\mathbf{f}) = \mathbb{E}\{\|\mathbf{f}(\mathbf{x}) - \mathbf{h}\|_2^2\} \tag{15}$$

as the estimation MSE of $\mathbf{f}(\mathbf{x})$. From the orthogonal principle, we have

$$\begin{aligned} J(\mathbf{f}) &= \mathbb{E}\big\{\|\mathbf{f}(\mathbf{x}) - \mathbf{h}_{\mathrm{MMSE}} + \mathbf{h}_{\mathrm{MMSE}} - \mathbf{h}\|_2^2\big\} \\ &= \mathbb{E}\big\{\|\mathbf{f}(\mathbf{x}) - \mathbf{h}_{\mathrm{MMSE}}\|_2^2\big\} + \mathbb{E}\big\{\|\mathbf{h}_{\mathrm{MMSE}} - \mathbf{h}\|_2^2\big\} \\ &\quad + 2\mathbb{E}\big\{(\mathbf{f}(\mathbf{x}) - \mathbf{h}_{\mathrm{MMSE}})^T(\mathbf{h}_{\mathrm{MMSE}} - \mathbf{h})\big\} \\ &= \mathbb{E}\big\{\|\mathbf{f}(\mathbf{x}) - \mathbf{f}_o(\mathbf{x})\|_2^2\big\} + J_{\mathrm{MMSE}}, \end{aligned} \tag{16}$$

where $\mathbf{f}_o(\mathbf{x}) = \mathbf{h}_{\mathrm{MMSE}}$ denotes the channel estimate of the MMSE estimator and $J(\mathbf{f}_o) = J_{\mathrm{MMSE}}$ is the corresponding MSE.

The first term in the right-hand side (RHS) of (16) is the expectation of the squared 2-norm distance from the use of $\mathbf{f}(\mathbf{x})$ to model $\mathbf{h}_{\mathrm{MMSE}}$ and is non-negative. In this respect, the second term J_{MMSE}, i.e., the MSE of the MMSE estimator, in the RHS of (16) provides a lower bound on $J(\mathbf{f})$, which is independent of $\mathbf{f}(\mathbf{x})$ and is determined by the joint distribution of $\mathbf{x}$ and $\mathbf{h}$. The distance between $J(\mathbf{f})$ and J_{MMSE} then serves as the evaluation criterion for the performance assessment of the estimator $\mathbf{f}(\mathbf{x})$.

Given a ReLU DNN with parameter $\boldsymbol{\theta}$, the input-output relation of the DL estimator can be expressed as a function $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})$ in (13). Then,

$$J(\mathbf{f}_{\boldsymbol{\theta}}) = \mathbb{E}\big\{\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{h}\|_2^2\big\} \tag{17}$$

is the corresponding MSE of the DL estimator with $\mathbf{f}(\mathbf{x})$ in (15) replaced by $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})$.

Let $\Theta_R = \{\boldsymbol{\theta} \,|\, \|\boldsymbol{\theta}\|_\infty \le R, R \ge 1\}$ be the bounded subset of Θ, and we will analyze the performance of the DL estimator within Θ_R. Denote

$$\boldsymbol{\theta}_o = \arg\min_{\boldsymbol{\theta}\in\Theta_R} J(\mathbf{f}_{\boldsymbol{\theta}}),\ J(\mathbf{f}_{\boldsymbol{\theta}_o}) = \min_{\boldsymbol{\theta}\in\Theta_R} J(\mathbf{f}_{\boldsymbol{\theta}}). \tag{18}$$

Hence, $\mathbf{f}_{\boldsymbol{\theta}_o}(\mathbf{x})$ is the optimal DL estimator, and $\boldsymbol{\theta}_o$ is the corresponding parameter of the ReLU DNN. It is obvious that $J(\mathbf{f}_{\boldsymbol{\theta}_o})$ is the minimum MSE over all $\boldsymbol{\theta} \in \Theta_R$.

Similarly, denote

$$\boldsymbol{\theta}_{\mathcal{Z}} = \arg\min_{\boldsymbol{\theta}\in\Theta_R} J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}}),\ J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) = \min_{\boldsymbol{\theta}\in\Theta_R} J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}}), \tag{19}$$

where

$$J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}}) = \frac{1}{|\mathcal{Z}|} \sum_{(\mathbf{x}_m,\mathbf{h}_m)\in\mathcal{Z}} \|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_m) - \mathbf{h}_m\|_2^2 \tag{20}$$

is the least-squared error (LSE) of the DL estimator w.r.t. $\mathcal{Z}$.

In the above, $\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}(\mathbf{x})$ is the optimal DL estimator trained by the dataset $\mathcal{Z}$ using the LSE as in (20). Therefore,

$$\mathbf{h}_{\mathrm{DL}} = \mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}(\mathbf{x}) \tag{21}$$

is the DL channel estimate obtained in practice. The corresponding MSE of $\mathbf{h}_{\mathrm{DL}}$ is $J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})$, which is obviously no less than $J(\mathbf{f}_{\boldsymbol{\theta}_o})$, that is, $J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) - J(\mathbf{f}_{\boldsymbol{\theta}_o}) \ge 0$. Let us then analyze the performance of the DL estimator through quantifying the distance of $J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})$ to J_{MMSE}.

To extend the linear model in (1) to general systems, the following statistical model

$$\mathbf{x} = \mathbf{f}_{\mathrm{u}}(\tau\mathbf{h} + \mathbf{n}) \tag{22}$$

is considered for channel estimation of the DL estimator, where $\mathbf{f}_{\mathrm{u}}(\cdot)$ denotes the unknown distortion imposed on the received signal, e.g., imperfect power amplifier (PA) [27], [28] and quantization error of analog to digital converter (ADC) [29].

First, $J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})$ can break into two different terms as

$$J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) = J(\mathbf{f}_{\boldsymbol{\theta}_o}) + [J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) - J(\mathbf{f}_{\boldsymbol{\theta}_o})]. \tag{23}$$

According to (16), the first term $J(\mathbf{f}_{\boldsymbol{\theta}_o})$ in (23) can be further decomposed into

$$J(\mathbf{f}_{\boldsymbol{\theta}_o}) = \mathbb{E}\{\|\mathbf{f}_{\boldsymbol{\theta}_o}(\mathbf{x}) - \mathbf{f}_o(\mathbf{x})\|_2^2\} + J(\mathbf{f}_o). \tag{24}$$

As $J(\mathbf{f}_o)$ has the lowest MSE, $J(\mathbf{f}_{\boldsymbol{\theta}_o})$ is determined by $\mathbb{E}\{\|\mathbf{f}_{\boldsymbol{\theta}_o}(\mathbf{x}) - \mathbf{f}_o(\mathbf{x})\|_2^2\}$, which is referred to as the approximation error.

The second term, $J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) - J(\mathbf{f}_{\boldsymbol{\theta}_o})$, in the RHS of (23) is non-negative and is determined by $\mathcal{Z}$, which is called the generalization error.

We will analyze the performance of the DL estimator through quantifying the approximation error and generalization error, respectively.

The following theorem demonstrates that the approximation error in (24) of the DL estimator can be narrowed down with any precision by a ReLU DNN of finite depth.

Theorem 1: If $\mathbf{f}_o(\mathbf{x}) \in \ell_2$, then there exits an optimized DL estimator $\mathbf{f}_{\boldsymbol{\theta}_o}(\mathbf{x})$ built on a ReLU DNN of $\boldsymbol{\theta} \in \Theta_R$ with sufficiently large R and at most $\lceil \log_2(d+1) \rceil$ hidden layers such that

$$\mathbb{E}\{\|\mathbf{f}_{\boldsymbol{\theta}_o}(\mathbf{x}) - \mathbf{f}_o(\mathbf{x})\|_2^2\} \leq \varepsilon \tag{25}$$

for any $\varepsilon > 0$.

Proof: From (13), $\mathbf{f}_{\boldsymbol{\theta}_o}(\mathbf{x})$ is equivalent to a $\mathbb{R}^d \to \mathbb{R}^d$ piecewise linear function. According to [25, Theorem 2.1], any $\mathbb{R}^d \to \mathbb{R}$ piecewise linear function can be represented by a DL estimator that is built on a ReLU DNN with no more than $\lceil \log_2(d+1) \rceil$ hidden layers.

On the other hand, any function $\mathbf{f}(\mathbf{x}) \in \ell_2$ has finite 2-norm and can be approximated by a piecewise linear function with arbitrary precision and finite-valued parameter [30]. Let $f_{o,i}(\mathbf{x})$ be the i-th entry of $\mathbf{f}_o(\mathbf{x})$ for $i \in \{1, \ldots, d\}$ and there exists d $\mathbb{R}^d \to \mathbb{R}$ piecewise linear functions to approximate $\{f_{o,1}(\mathbf{x}), \ldots, f_{o,d}(\mathbf{x})\}$ with arbitrary precision if $\mathbf{f}_o(\mathbf{x}) \in \ell_2$. Such a set of d piecewise linear functions can be represented by d ReLU DNNs of $\boldsymbol{\theta} \in \Theta_R$ each with at most $\lceil \log_2(d+1) \rceil$ hidden layers when R is sufficiently large. Then, we can simply put these ReLU DNNs in parallel and combine their outputs to compose a single ReLU DNN.

The depths of these ReLU DNNs may be different and we need to align their depths for the composition. Since the output of any hidden layer of a ReLU DNN can be replicated by adding hidden layers, we can simply add one or multiple of hidden layers for each ReLU DNN to align the depths of these ReLU DNNs. Let $l_{\max}$ be the maximum depth of these ReLU DNNs and then the aligned depth is just given by $l_{\max}$, which is upper bounded by $\lceil \log_2(d+1) \rceil$. Therefore, there exits a DL estimator with parameter $\boldsymbol{\theta}_\varepsilon \in \Theta_R$ with at most $\lceil \log_2(d+1) \rceil$ hidden layers such that

$$\|\mathbf{f}_{\boldsymbol{\theta}_\varepsilon}(\mathbf{x}) - \mathbf{f}_o(\mathbf{x})\|_2^2 \leq \varepsilon \tag{26}$$

for any $\mathbf{x} \in \mathcal{X}$ and $\varepsilon > 0$. From (24), $\mathbf{f}_{\boldsymbol{\theta}_o}(\mathbf{x})$ has the lowest MSE for all $\boldsymbol{\theta} \in \Theta_R$, and we have

$$\mathbb{E}\|\mathbf{f}_{\boldsymbol{\theta}_o}(\mathbf{x}) - \mathbf{f}_o(\mathbf{x})\|_2^2 \leq \mathbb{E}\|\mathbf{f}_{\boldsymbol{\theta}_\varepsilon}(\mathbf{x}) - \mathbf{f}_o(\mathbf{x})\|_2^2 \leq \varepsilon, \tag{27}$$

which completes the proof. ■

Remark 2:

- Theorem 1 shows that the approximation error in (24) can be reduced by increasing the number of linear pieces representable by the DL estimator. Moreover, the number of linear pieces generated by the DL estimator is increased with the network size, i.e., the dimension of the parameter space Θ_R [23]–[25]. Therefore, we can reduce the approximation error by increasing the network size and the DL estimator with deeper network structure typically achieve better performance.
- Theorem 1 also indicates that the DL estimator is powerful at function representation and does not restrict to any type of signal models or channel statistics. If no specific models are known a priori or complicated nonlinear systems are presented, the DL estimator will be a preferred choice for channel estimation.

Next, we will discuss the rate of convergence of the generalization error in (23). Some auxiliary lemmas are provided before the main result.

Lemma 1: Let $\alpha = R\|\mathbf{d}\|_\infty$, $\beta = \alpha/(\alpha - 1)$, and

$$\mu = \max_{i \in \{0,\ldots,4\}} \left\{\mathbb{E}\{(\|\mathbf{x}\|_2 + \beta)^i (\|\mathbf{h}\|_2 - \beta)^{4-i}\}\right\}. \tag{28}$$

Assume that μ are finite. Then, for all $\varepsilon > 0$ and

$$|\mathcal{Z}| \geq 8\mu(\alpha^{l+1} + 1)^4/\varepsilon^2, \tag{29}$$

we have

$$\mathbf{P}\Big(\sup_{\boldsymbol{\theta} \in \Theta_R} |J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}}) - J(\mathbf{f}_{\boldsymbol{\theta}})| > \varepsilon\Big) \leq 4\mathbf{P}\Big(\sup_{\boldsymbol{\theta} \in \Theta_R} |J^\circ_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}})| > \frac{\varepsilon}{4}\Big), \tag{30}$$

where $\mathbf{P}$ denotes the distribution of the training sample in $\mathcal{Z}$ and $J^\circ_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}}) = 1/|\mathcal{Z}| \sum_{m=1}^{|\mathcal{Z}|} \omega_m \|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_m) - \mathbf{h}_m\|_2^2$ with $\{\omega_1, \ldots, \omega_{|\mathcal{Z}|}\}$ a Rademacher sequence.

Proof: If

$$\mathbf{P}(|J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}}) - J(\mathbf{f}_{\boldsymbol{\theta}})| > \frac{\varepsilon}{2}) \leq \frac{1}{2} \tag{31}$$

for all $\boldsymbol{\theta} \in \Theta_R$, then (30) holds according to [31, Lemma 2.3.7].

Let $\sigma^2(\mathbf{f}_{\boldsymbol{\theta}})$ be the variance of $\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{h}\|_2^2$. Chebyshev's inequality [32] assures that

$$\mathbf{P}\big(|J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}}) - J(\mathbf{f}_{\boldsymbol{\theta}})| \geq \frac{\varepsilon}{2}\big) \leq \frac{4\sigma^2(\mathbf{f}_{\boldsymbol{\theta}})}{|\mathcal{Z}|\varepsilon^2} \tag{32}$$

for all $\boldsymbol{\theta} \in \Theta_R$. Specifically, $\sigma^2(\mathbf{f}_{\boldsymbol{\theta}})$ satisfies

$$\begin{aligned}\sigma^2(\mathbf{f}_{\boldsymbol{\theta}}) &= \mathbb{E}\big\{\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{h}\|_2^4\big\} - J(\mathbf{f}_{\boldsymbol{\theta}})^2 \leq \mathbb{E}\big\{\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{h}\|_2^4\big\} \\ &\leq \mathbb{E}\big\{\big(\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})\|_2^2 + 2\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})\|_2\|\mathbf{h}\|_2 + \|\mathbf{h}\|_2^2\big)^2\big\} \\ &= \mathbb{E}\big\{\big(\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})\|_2 + \|\mathbf{h}\|_2\big)^4\big\}. \end{aligned} \tag{33}$$

Assume that the input space of $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})$ follows the partition in (10). Using the triangle inequality yields

$$\left\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})\right\|_2 = \left\|\mathbf{W}_{\mathcal{X}_k}\mathbf{x}+\mathbf{b}_{\mathcal{X}_k}\right\|_2 \le \|\mathbf{W}_{\mathcal{X}_k}\|_2\|\mathbf{x}\|_2 + \|\mathbf{b}_{\mathcal{X}_k}\|_2 \tag{34}$$

for $\mathbf{x}\in\mathcal{X}_k$. Moreover, $\|\mathbf{W}_{\mathcal{X}_k}\|_2$ and $\|\mathbf{b}_{\mathcal{X}_k}\|_2$ in (34) are upper bounded by

$$\begin{aligned}\|\mathbf{W}_{\mathcal{X}_k}\|_2 &= \left\|\prod_{i=0}^{l}\tilde{\mathbf{W}}_i\right\|_2 = \left\|\prod_{i=0}^{l}\mathbf{W}_i\boldsymbol{\Lambda}_i\right\|_2 \\ &\le \prod_{i=0}^{l}\left\|\mathbf{W}_i\boldsymbol{\Lambda}_i\right\|_2 \le \prod_{i=0}^{l}\left\|\mathbf{W}_i\right\|_2 \end{aligned}\tag{35}$$

and

$$\begin{aligned}\|\mathbf{b}_{\mathcal{X}_k}\|_2 &= \left\|\sum_{i=0}^{l-1}\Big(\prod_{j=0}^{i}\tilde{\mathbf{W}}_{l-j}\Big)\mathbf{b}_{l-1-i}+\mathbf{b}_l\right\|_2 \\ &\le \sum_{i=0}^{l-1}\left\|\prod_{j=0}^{i}\tilde{\mathbf{W}}_{l-j}\right\|_2\left\|\mathbf{b}_{l-1-i}\right\|_2+\|\mathbf{b}_l\|_2 \\ &\le \sum_{i=0}^{l-1}\Big(\prod_{j=0}^{i}\|\mathbf{W}_{l-j}\|_2\Big)\left\|\mathbf{b}_{l-1-i}\right\|_2+\|\mathbf{b}_l\|_2, \end{aligned}\tag{36}$$

respectively. Note that $\|\mathbf{W}_i\|_2 \le R\|\mathbf{d}\|_\infty = \alpha$ and $\|\mathbf{b}_i\|_2 \le \alpha$ for $i\in\{0,1,\ldots,l\}$. Substituting these bounds into (35) and (36) yields

$$\|\mathbf{W}_{\mathcal{X}_k}\|_2 \le \alpha^{l+1} \tag{37}$$

and

$$\|\mathbf{b}_{\mathcal{X}_k}\|_2 \le \Big(\sum_{i=0}^{l-1}\alpha^{i+1}\Big)\alpha+\alpha = \frac{\alpha^{l+2}-\alpha}{\alpha-1} \le \beta(\alpha^{l+1}-1), \tag{38}$$

respectively. From (34), (37), and (38), $\|\mathbf{f}(\mathbf{x})\|_2$ is further bounded by

$$\|\mathbf{f}(\mathbf{x})\|_2 \le \alpha^{l+1}(\|\mathbf{x}\|_2+\beta)-\beta. \tag{39}$$

Combining (33) and (39), we have

$$\begin{aligned}\sigma^2(\mathbf{f}) &\le \mathbb{E}\{(\alpha^{l+1}(\|\mathbf{x}\|_2+\beta)+\|\mathbf{h}\|_2-\beta)^4\} \\ &= \sum_{i=0}^{4}\binom{4}{i}\alpha^{i(l+1)}\mathbb{E}\big\{(\|\mathbf{x}\|_2+\beta)^i(\|\mathbf{h}\|_2-\beta)^{4-i}\big\} \\ &\le \mu(\alpha^{l+1}+1)^4. \end{aligned}\tag{40}$$

Replace $\sigma^2(\mathbf{f}_{\boldsymbol{\theta}})$ in (32) by the bound in (40) and we can derive (30) and (31) under the condition that

$$|\mathcal{Z}| \ge 8\mu(\alpha^{l+1}+1)^4/\varepsilon^2, \tag{41}$$

which finishes the proof. ∎

Lemma 2: Assume that both $\frac{1}{|\mathcal{Z}|}\sum_{m=1}^{|\mathcal{Z}|}\|\mathbf{x}_m\|_2^4 \le \delta^4$ and $\frac{1}{|\mathcal{Z}|}\sum_{m=1}^{|\mathcal{Z}|}\|\mathbf{h}_m\|_2^4 \le \delta^4$ for $\delta>0$. For any $\boldsymbol{\theta},\boldsymbol{\lambda}\in\Theta_R$, we have

$$|J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}})-J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\lambda}})| \le 3^{l+1}2\|\mathbf{d}\|_\infty\alpha^{2l+1}(\delta+\beta)^2\|\boldsymbol{\theta}-\boldsymbol{\lambda}\|_\infty. \tag{42}$$

Proof: Note that

$$\begin{aligned}&|J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}})-J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\lambda}})| \\ &\le \frac{1}{|\mathcal{Z}|}\sum_{m=1}^{|\mathcal{Z}|}|(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_m)-\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{x}_m))^T(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_m)+\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{x}_m)-2\mathbf{h}_m)| \\ &\le \frac{1}{|\mathcal{Z}|}\sum_{m=1}^{|\mathcal{Z}|}\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_m)-\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{x}_m)\|_2\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_m)+\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{x}_m)-2\mathbf{h}_m\|_2. \end{aligned}\tag{43}$$

From Lemma 1, $\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})+\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{x})-2\mathbf{h}\|_2$ is upper bounded by

$$\begin{aligned}\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})+\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{x})-2\mathbf{h}\|_2 &\le \|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})\|_2+\|\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{x})\|_2+2\|\mathbf{h}\|_2 \\ &\le 2[\alpha^{l+1}(\|\mathbf{x}\|_2+\beta)+\|\mathbf{h}\|_2-\beta]. \end{aligned}\tag{44}$$

Let $\mathbf{V}_i$ and $\mathbf{p}_i$ be the weight and the bias of the i-th layer corresponding to $\boldsymbol{\lambda}$ for $i\in\{0,\ldots,l\}$. To simplify the notation, we define the partial parameter $\boldsymbol{\theta}_s$ and $\boldsymbol{\lambda}_s$ by

$$\boldsymbol{\theta}_s = (\mathrm{vec}(\mathbf{W}_0),\mathbf{b}_0,\ldots,\mathrm{vec}(\mathbf{W}_{s-1}),\mathbf{b}_{s-1}) \tag{45}$$

and

$$\boldsymbol{\lambda}_s = (\mathrm{vec}(\mathbf{V}_0),\mathbf{p}_0,\ldots,\mathrm{vec}(\mathbf{V}_{s-1}),\mathbf{p}_{s-1}), \tag{46}$$

respectively, the partial network outputs by $\mathbf{f}_{\boldsymbol{\theta}s}(\mathbf{x})$ and $\mathbf{f}_{\boldsymbol{\lambda}s}(\mathbf{x})$, respectively, and the partial error e_s by

$$e_s = \|\mathbf{f}_{\boldsymbol{\theta}s}(\mathbf{x})-\mathbf{f}_{\boldsymbol{\lambda}s}(\mathbf{x})\|_2 \tag{47}$$

for $s\in\{1,\ldots,l+1\}$. Specifically, the first term in (43) can be represented by e_{l+1}.

Using the triangle inequality, we can bound e_{s+1} by

$$\begin{aligned}e_{s+1} &= \|(\mathbf{W}_s\varphi_{d_s}(\mathbf{f}_{\boldsymbol{\theta}_s}(\mathbf{x}))+\mathbf{b}_s)-(\mathbf{V}_s\varphi_{d_s}(\mathbf{f}_{\boldsymbol{\lambda}_s}(\mathbf{x}))+\mathbf{p}_s)\|_2 \\ &\le \|\mathbf{W}_s-\mathbf{V}_s\|_2\|\mathbf{f}_{\boldsymbol{\theta}_s}(\mathbf{x})-\mathbf{f}_{\boldsymbol{\lambda}_s}(\mathbf{x})\|_2 \\ &\quad+\|\mathbf{V}_s\|_2\|\mathbf{f}_{\boldsymbol{\theta}_s}(\mathbf{x})-\mathbf{f}_{\boldsymbol{\lambda}_s}(\mathbf{x})\|_2 \\ &\quad+\|\mathbf{W}_s-\mathbf{V}_s\|_2\|\mathbf{f}_{\boldsymbol{\lambda}_s}(\mathbf{x})\|_2+\|\mathbf{b}_s-\mathbf{p}_s\|_2 \\ &\le \|\mathbf{d}\|_\infty[re_s+Re_s+(y_s+1)r] \\ &\le \|\mathbf{d}\|_\infty[3Re_s+(y_s+1)r]. \end{aligned}\tag{48}$$

for $s\in\{1,\ldots,l\}$, where $r=\|\boldsymbol{\theta}-\boldsymbol{\lambda}\|_\infty$ and $y_s=\alpha^s(\|\mathbf{x}\|_2+\beta)-\beta$ is the upper bound on $\mathbf{f}_{\boldsymbol{\theta}_s}(\mathbf{x})$ and $\mathbf{f}_{\boldsymbol{\lambda}_s}(\mathbf{x})$ from (39).

We now claim that

$$e_{s+1} \le r\|\mathbf{d}\|_\infty\big[(3\alpha)^s(\|\mathbf{x}\|_2+1)+\sum_{i=0}^{s-1}(3\alpha)^i(y_{s-i}+1)\big] \tag{49}$$

for $s\in\{0,\ldots,l\}$ and prove it by induction. The base case $s=0$ holds as

$$\begin{aligned}e_1 &= \|(\mathbf{W}_0\mathbf{x}+\mathbf{b}_0)-(\mathbf{V}_0\mathbf{x}+\mathbf{p}_0)\|_2 \\ &\le r\|\mathbf{d}\|_\infty(\|\mathbf{x}\|_2+1). \end{aligned}\tag{50}$$

For the induction step, assume that (49) is valid for $s \in \{0,\ldots,l-1\}$. It implies by (48) that

$$\begin{aligned} e_{l+1} &\le 3\alpha e_l + r\|\mathbf{d}\|_\infty(y_l+1) = r\|\mathbf{d}\|_\infty\big[(3\alpha)^l(\|\mathbf{x}\|_2+1) \\ &\quad + \sum_{i=0}^{l-2}(3\alpha)^{i+1}(y_{l-1-i}+1) + (y_l+1)\big] \\ &= r\|\mathbf{d}\|_\infty\big[(3\alpha)^l(\|\mathbf{x}\|_2+1) + \sum_{i=0}^{l-1}(3\alpha)^i(y_{l-i}+1)\big]. \end{aligned} \tag{51}$$

Therefore, our claim (49) holds for $s \in \{0,\ldots,l\}$. Using (39), the upper bound on e_{l+1} can be further written as

$$\begin{aligned} e_{l+1} &\le r\|\mathbf{d}\|_\infty[\frac{3^{l+1}-1}{2}\alpha^l\|\mathbf{x}\|_2 + (3\alpha)^l \\ &\quad + \frac{3^l-1}{2}\alpha^l\beta - \frac{(3\alpha)^l-1}{3\alpha-1}(\beta-1)] \\ &\le \frac{3}{2}r\|\mathbf{d}\|_\infty(3\alpha)^l(\|\mathbf{x}\|_2+\beta). \end{aligned} \tag{52}$$

Then, $|J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}}) - J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\lambda}})|$ is upper bounded by

$$\begin{aligned} |J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}}) - J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\lambda}})| &\le \frac{1}{|\mathcal{Z}|}\sum_{m=1}^{|\mathcal{Z}|} 3r\|\mathbf{d}\|_\infty(3\alpha)^l(\|\mathbf{x}_m\|_2+\beta) \\ &\quad [\alpha^{l+1}(\|\mathbf{x}_m\|_2+\beta) + \|\mathbf{h}_m\|_2] \\ &= \frac{3r\|\mathbf{d}\|_\infty(3\alpha)^l}{|\mathcal{Z}|}\sum_{m=1}^{|\mathcal{Z}|}\big[\alpha^{l+1}(\|\mathbf{x}_m\|_2+\beta)^2 \\ &\quad + \|\mathbf{x}_m\|_2\|\mathbf{h}_m\|_2 + \beta\|\mathbf{h}_m\|_2\big] \\ &\le 3r\|\mathbf{d}\|_\infty(3\alpha)^l[\alpha^{l+1}(\delta+\beta)^2 + \delta(\delta+\beta)] \\ &\le 3^{l+1}2r\|\mathbf{d}\|_\infty\alpha^{2l+1}(\delta+\beta)^2, \end{aligned} \tag{53}$$

which finishes the proof. ∎

Define the covering number $C(\varepsilon,\Theta_R)$ as the smallest value of $C \in \mathbb{N}$ for which there exits a collection of functions $\mathbf{f}_1(\mathbf{x}),\ldots,\mathbf{f}_C(\mathbf{x}) \in \mathcal{D}$ with their parameters belonging to Θ_R such that

$$\min_{j\in\{1,\ldots,C\}} |J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}}) - J_{\mathcal{Z}}(\mathbf{f}_j)| \le \varepsilon \tag{54}$$

for $\varepsilon > 0$ and any $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}) \in \mathcal{D}$ with $\boldsymbol{\theta} \in \Theta_R$. The following lemma derives a bound on $C(\varepsilon,\Theta_R)$.

Lemma 3: With settings of Lemma 2, for any $\varepsilon > 0$, it holds that

$$\ln C(\varepsilon,\Theta_R) \le d_u \ln\Big[\frac{3^{l+1}8\alpha^{2(l+1)}(\delta+\beta)^2}{\varepsilon}\Big]. \tag{55}$$

Proof: Choose a collection of parameters $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_C \in \Theta_R$ such that the balls centered at $\boldsymbol{\theta}_j$ with radius

$$r_b = \frac{\varepsilon}{3^{l+1}2\|\mathbf{d}\|_\infty\alpha^{2l+1}(\delta+\beta)^2} \tag{56}$$

cover Θ_R for $j \in \{1,\ldots,C\}$. Then, there exits $j \in \{1,\ldots,C\}$ such that $\|\boldsymbol{\theta}-\boldsymbol{\theta}_j\|_\infty \le r_b$ for any $\boldsymbol{\theta} \in \Theta_R$. Let $C_b = \ln C$ and it is upper bounded by [33]

$$C_b \le d_u \ln(4R/r_b) \tag{57}$$

It implies by Lemma 2 and (56) that

$$\begin{aligned} &|J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}}) - J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}_j})| \\ &\quad \le 3^{l+1}2\|\boldsymbol{\theta}-\boldsymbol{\theta}_j\|_\infty\|\mathbf{d}\|_\infty\alpha^{2l+1}(\delta+\beta)^2 \le \varepsilon. \end{aligned} \tag{58}$$

Then, $\ln C(\varepsilon,\Theta_R)$ is upper bounded by

$$\begin{aligned} \ln C(\varepsilon,\Theta_R) &\le C_b \le d_u\ln(4R/r_b) \\ &= d_u\ln\Big[\frac{3^{l+1}8\alpha^{2(l+1)}(\delta+\beta)^2}{\varepsilon}\Big], \end{aligned} \tag{59}$$

which finished the proof. ∎

Following Lemma 1, Lemma 2, and Lemma 3, the next theorem demonstrates the rate of convergence of the generalization error in (23).

Theorem 2: With the settings of Lemma 1 and Lemma 2, let $\mu_1 = \mathbb{E}\{\|\mathbf{x}\|_2^4\}$, $\mu_2 = \mathbb{E}\{\|\mathbf{h}\|_2^4\}$, $\delta_1 = 8[(\alpha^{l+1}\delta+\beta)^4+\delta^4]$, and $\delta_2 = 3^{l+1}2^7\alpha^{2(l+1)}(\delta+\beta)^2$ for any $\delta^4 > \max\{\mu_1,\mu_2\}$. Denote σ_1 and σ_2 as the variances of $\|\mathbf{x}\|_2^4$ and $\|\mathbf{h}\|_2^4$, respectively. For any $\varepsilon > 0$, it holds that

$$\mathbf{P}([J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) - J(\mathbf{f}_{\boldsymbol{\theta}_o})] > \varepsilon) \le 8\exp\big(-\frac{|\mathcal{Z}|\varepsilon^2}{1024\delta_1}\big) + \frac{8\sigma^2}{|\mathcal{Z}|\delta_3} \tag{60}$$

if $|\mathcal{Z}| \ge 32\mu(\alpha^{l+1}+1)^4/\varepsilon^2$ and $|\mathcal{Z}| \ge (1024\delta_1 d_u \ln\frac{\delta_2}{\varepsilon})/\varepsilon^2$, where $\delta_3 = \min\{(\delta^4-\mu_1)^2,(\delta^4-\mu_2)^2\}$ and $\sigma = \max\{\sigma_1,\sigma_2\}$.

Proof: From (18) and (19), $J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) - J(\mathbf{f}_{\boldsymbol{\theta}_o})$ is bounded by

$$\begin{aligned} 0 &\le J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) - J(\mathbf{f}_{\boldsymbol{\theta}_o}) \\ &= [J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) - J(\mathbf{f}_{\boldsymbol{\theta}_o})] - [J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) - J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}_o})] \\ &\quad + [J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) - J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}_o})] \\ &\le [J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) - J(\mathbf{f}_{\boldsymbol{\theta}_o})] - [J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) - J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}_o})] \\ &\le 2\sup_{\boldsymbol{\theta}\in\Theta_R}|J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}}) - J(\mathbf{f}_{\boldsymbol{\theta}})|. \end{aligned} \tag{61}$$

According to Lemma 1, the last inequality in (61) satisfies

$$\begin{aligned} &\mathbf{P}(J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) - J(\mathbf{f}_{\boldsymbol{\theta}_o}) > \varepsilon) \\ &\quad \le \mathbf{P}\big(\sup_{\boldsymbol{\theta}\in\Theta_R}|J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}}) - J(\mathbf{f}_{\boldsymbol{\theta}})| > \frac{\varepsilon}{2}\big) \\ &\quad \le 4\mathbf{P}\big(\sup_{\boldsymbol{\theta}\in\Theta_R}|J^\circ_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}})| > \frac{\varepsilon}{8}\big), \end{aligned} \tag{62}$$

if $|\mathcal{Z}| \ge 32\mu(\alpha^{l+1}+1)^4/\varepsilon^2$.

Assume that $\mathcal{Z}$ is fixed with $\frac{1}{|\mathcal{Z}|}\sum_{m=1}^{|\mathcal{Z}|}\|\mathbf{x}_m\|_2^4 \le \delta^4$ and $\frac{1}{|\mathcal{Z}|}\sum_{m=1}^{|\mathcal{Z}|}\|\mathbf{h}_m\|_2^4 \le \delta^4$ for $\delta > 0$. Choose a collection of functions $\mathbf{f}_1(\mathbf{x}),\ldots,\mathbf{f}_C(\mathbf{x}) \in \mathcal{D}$, where $C = C(\varepsilon/16,\Theta_R)$, such that

$$\min_{j\in\{1,\ldots,C\}} |J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}}) - J_{\mathcal{Z}}(\mathbf{f}_j)| \le \frac{\varepsilon}{16} \tag{63}$$

for any $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}) \in \mathcal{D}$ with $\boldsymbol{\theta} \in \Theta_R$. Let $\mathbf{f}^*(\mathbf{x})$ represent $\mathbf{f}_j(\mathbf{x})$ at which the minimum value in (63) is achieved. Since

$$|J^\circ_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}})| = \big|\frac{1}{|\mathcal{Z}|}\sum_{m=1}^{|\mathcal{Z}|}\omega_m J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}})\big| \le \big|J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}})\big|, \tag{64}$$

we have

$$\begin{aligned}
&\mathbf{P}\big(\sup_{\boldsymbol{\theta}\in\Theta_R}|J^{\circ}_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}})|>\frac{\varepsilon}{8}|\mathcal{Z}\big)\\
&\quad\le \mathbf{P}\big(\sup_{\boldsymbol{\theta}\in\Theta_R}\big[|J^{\circ}_{\mathcal{Z}}(\mathbf{f}^*)|+|J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}})-J_{\mathcal{Z}}(\mathbf{f}^*)|\big]>\frac{\varepsilon}{8}|\mathcal{Z}\big)\\
&\quad\le \mathbf{P}\big(\max_{j\in\{1,\ldots,C\}}|J^{\circ}_{\mathcal{Z}}(\mathbf{f}_j)|>\frac{\varepsilon}{16}|\mathcal{Z}\big)\\
&\quad\le \sum_{j=1}^{C}\mathbf{P}\big(|J^{\circ}_{\mathcal{Z}}(\mathbf{f}_j)|>\frac{\varepsilon}{16}|\mathcal{Z}\big). \qquad (65)
\end{aligned}$$

Hoeffding's Inequality [31] gives the following bound

$$\begin{aligned}
&\mathbf{P}\big(|J^{\circ}_{\mathcal{Z}}(\mathbf{f}_j)|>\frac{\varepsilon}{16}|\mathcal{Z}\big)\\
&\quad=\mathbf{P}\big(|\sum_{m=1}^{|\mathcal{Z}|}\omega_m\|\mathbf{f}_j(\mathbf{x}_m)-\mathbf{h}_m\|_2^2|>\frac{|\mathcal{Z}|}{16}\varepsilon\,|\,\mathcal{Z}\big)\\
&\quad\le 2\exp\big[-2(\frac{|\mathcal{Z}|}{16}\varepsilon)^2/\sum_{m=1}^{|\mathcal{Z}|}(2\|\mathbf{f}_j(\mathbf{x}_m)-\mathbf{h}_m\|_2^2)^2\big] \qquad (66)
\end{aligned}$$

for each $\mathbf{f}_j(\mathbf{x})$. From (39), $\sum_{m=1}^{|\mathcal{Z}|}(\|\mathbf{f}_j(\mathbf{x}_m)-\mathbf{h}_m\|_2^2)^2$ is upper bounded by

$$\begin{aligned}
&\sum_{m=1}^{|\mathcal{Z}|}(\|\mathbf{f}_j(\mathbf{x}_m)-\mathbf{h}_m\|_2^2)^2\\
&\quad\le\sum_{m=1}^{|\mathcal{Z}|}(\|\mathbf{f}_j(\mathbf{x}_m)\|_2+\|\mathbf{h}_m\|_2)^4\\
&\quad\le 8\sum_{m=1}^{|\mathcal{Z}|}(\|\mathbf{f}_j(\mathbf{x}_m)\|_2^4+\|\mathbf{h}_m\|_2^4)\\
&\quad\le 8|\mathcal{Z}|[(\alpha^{l+1}\delta+\beta)^4+\delta^4]=|\mathcal{Z}|\delta_1. \qquad (67)
\end{aligned}$$

Replace the last term in (66) with (67) and substituting (66) into (65) yields

$$\mathbf{P}\big(\sup_{\boldsymbol{\theta}\in\Theta_R}|J^{\circ}_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}})|>\frac{\varepsilon}{8}|\mathcal{Z}\big)\le 2\exp\big(\ln C-\frac{|\mathcal{Z}|\varepsilon^2}{512\delta_1}\big). \qquad (68)$$

According to Lemma 3, it holds that

$$\begin{aligned}
\ln C&=\ln C(\varepsilon/16,\Theta_R)\\
&\le d_u\ln\Big[\frac{3^{l+1}2^7\alpha^{2(l+1)}(\delta+\beta)^2}{\varepsilon}\Big]\le d_u\ln\frac{\delta_2}{\varepsilon}. \qquad (69)
\end{aligned}$$

If

$$|\mathcal{Z}|\ge(1024\delta_1 d_u\ln\frac{\delta_2}{\varepsilon})/\varepsilon^2, \qquad (70)$$

then $\ln C\le|\mathcal{Z}|\varepsilon^2/1024\delta_1$ and

$$\mathbf{P}\big(\sup_{\boldsymbol{\theta}\in\Theta_R}|J^{\circ}_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}})|>\frac{\varepsilon}{8}|\mathcal{Z}\big)\le 2\exp\big(-\frac{|\mathcal{Z}|\varepsilon^2}{1024\delta_1}\big). \qquad (71)$$

Integrating out $\mathbf{P}(\sup_{\boldsymbol{\theta}\in\Theta_R}|J^{\circ}_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}})|>\frac{\varepsilon}{8}|\mathcal{Z})$ over $\mathcal{Z}$ produces

$$\mathbf{P}\big(\sup_{\boldsymbol{\theta}\in\Theta_R}|J^{\circ}_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}})|>\frac{\varepsilon}{8}\big)\le 2\exp\big(-\frac{|\mathcal{Z}|\varepsilon^2}{1024\delta_1}\big)+\mathbf{P}_{\mathcal{Z}}, \qquad (72)$$

where

$$\mathbf{P}_{\mathcal{Z}}=\mathbf{P}\big(\frac{1}{|\mathcal{Z}|}\sum_{m=1}^{|\mathcal{Z}|}\|\mathbf{x}_m\|_2^4\ge\delta^4\big)+\mathbf{P}\big(\frac{1}{|\mathcal{Z}|}\sum_{m=1}^{|\mathcal{Z}|}\|\mathbf{h}_m\|_2^4\ge\delta^4\big). \qquad (73)$$

If $\delta^4>\max\{\mu_1,\mu_2\}$, using Chebyshev's inequality [32] yields

$$\begin{aligned}
\mathbf{P}_{\mathcal{Z}}&\le\mathbf{P}\big(|\frac{1}{|\mathcal{Z}|}\sum_{m=1}^{|\mathcal{Z}|}\|\mathbf{x}_m\|_2^4-\mu_1|\ge(\delta^4-\mu_1)\big)\\
&\quad+\mathbf{P}\big(|\frac{1}{|\mathcal{Z}|}\sum_{m=1}^{|\mathcal{Z}|}\|\mathbf{h}_m\|_2^4-\mu_2|\ge(\delta^4-\mu_2)\big)\\
&\le\frac{\sigma_1^2}{|\mathcal{Z}|(\delta^4-\mu_1)^2}+\frac{\sigma_2^2}{|\mathcal{Z}|(\delta^4-\mu_2)^2}\le\frac{2\sigma^2}{|\mathcal{Z}|\delta_3}. \qquad (74)
\end{aligned}$$

Combining (62) and (72), we arrive at

$$\mathbf{P}([J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})-J(\mathbf{f}_{\boldsymbol{\theta}_o})]>\varepsilon)\le 8\exp(-\frac{|\mathcal{Z}|\varepsilon^2}{1024\delta_1})+\frac{8\sigma^2}{|\mathcal{Z}|\delta_3}, \qquad (75)$$

which finishes the proof. ■

Together with Theorem 1, the following corollary presents our main conclusion on the performance of the DL estimator.

Corollary 1: With the settings of Theorem 1 and Theorem 2, there exits a DL estimator powered by a ReLU DNN of $\boldsymbol{\theta}\in\Theta_R$ with at most $\lceil\log_2(d+1)\rceil$ hidden layers and sufficiently large R such that

$$\lim_{|\mathcal{Z}|\to+\infty}\mathbf{P}([J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})-J(\mathbf{f}_o)]>\varepsilon)=0 \qquad (76)$$

for any $\varepsilon>0$.

Proof: According to (23) and (24), $J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})-J(\mathbf{f}_o)$ is decomposed into

$$J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})-J(\mathbf{f}_o)=\mathbb{E}\{\|\mathbf{f}_{\boldsymbol{\theta}_o}(\mathbf{x})-\mathbf{f}_o(\mathbf{x})\|_2^2\}+J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})-J(\mathbf{f}_{\boldsymbol{\theta}_o}). \qquad (77)$$

From Theorem 1, there exits an optimized DL estimator $\mathbf{f}_{\boldsymbol{\theta}_o}(\mathbf{x})$ with sufficiently large R at most $\lceil\log_2(d+1)\rceil$ hidden layers such that $\mathbb{E}\{\|\mathbf{f}_{\boldsymbol{\theta}_o}(\mathbf{x})-\mathbf{f}_o(\mathbf{x})\|_2^2\}\le\varepsilon$ for any $\varepsilon>0$.

From Theorem 2, we have $J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})-J(\mathbf{f}_{\boldsymbol{\theta}_o})\xrightarrow{\mathbf{P}}0$, where the notation $\xrightarrow{\mathbf{P}}$ denotes the convergence in probability, that is

$$\lim_{|\mathcal{Z}|\to+\infty}\mathbf{P}\big([J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})-J(\mathbf{f}_{\boldsymbol{\theta}_o})]>\varepsilon\big)=0 \qquad (78)$$

for any $\varepsilon>0$. Combining Theorems 1 and Theorem 2, we have

$$\lim_{|\mathcal{Z}|\to+\infty}\mathbf{P}\big([J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})-J(\mathbf{f}_o)]>\varepsilon\big)=0 \qquad (79)$$

for any $\varepsilon>0$, which completes the proof. ■

Fig. 1 illustrates the relationship between $\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}(\mathbf{x})$, $\mathbf{f}_{\boldsymbol{\theta}_o}(\mathbf{x})$, and $\mathbf{f}_o(\mathbf{x})$ to better understand Theorem 1 and Theorem 2. These two theorems demonstrate that the estimate of the DL estimator can arbitrarily well approximate to the estimate of the MMSE estimator, i.e., $\mathbf{f}_o(\mathbf{x})$ or $\mathbf{h}_{\mathrm{MMSE}}$, as $|\mathcal{Z}|$ gets large and the underlying ReLU DNN is suitably configured. We then derive the main result on the performance of the DL estimator based on Corollary 1 as

$$J(\mathbf{f}_o)\approx J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}). \qquad (80)$$

Specifically, the approximation error of the DL estimator in Theorem 1 can be eliminated in the linear systems. Suppose that $\mathbf{h}$ is a $d\times 1$ zero mean Gaussian vector and $\mathbf{h}_{\mathrm{LMMSE}}$

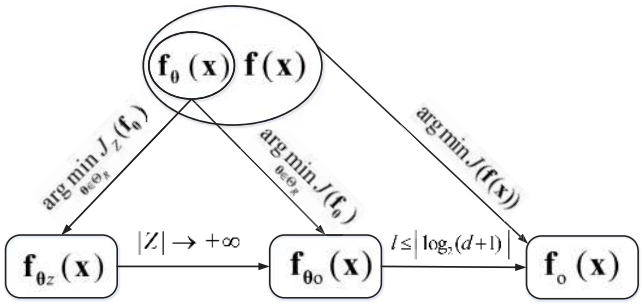


Fig. 1. The relationship between $\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}(\mathbf{x})$, $\mathbf{f}_{\boldsymbol{\theta}_o}(\mathbf{x})$, and $\mathbf{f}_o(\mathbf{x})$.

is simply equivalent to $\mathbf{h}_{\mathrm{MMSE}}$, i.e., $\mathbf{f}_o(\mathbf{x}) = \mathbf{h}_{\mathrm{LMMSE}}$ and $J_{\mathrm{MMSE}} = J_{\mathrm{LMMSE}}$. Let $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^d$ denote the affine transformation with weight $\mathbf{W} \in \mathbb{R}^{d\times d}$ and bias $\mathbf{b} \in \mathbb{R}^d$. Due to the fact that

$$\mathcal{A} = (\mathbf{I}_d \circ \varphi_d \circ \mathcal{A}) + (-\mathbf{I}_d \circ \varphi_d \circ (-\mathcal{A})), \tag{81}$$

the RHS of (81) is equivalent to a 2-layer ReLU DNN of size $2d$ [25, Lemma D.4.]. Such a representation is extensible to a wide class of the DL estimators with more than two layers, since the output of any hidden layer of a ReLU DNN can be repeated by adding one or multiple of hidden layers with the identity transformation as

$$\begin{aligned}\mathcal{A} = (&\mathbf{I}_d \circ \cdots \circ \varphi_d \circ \mathbf{I}_d \circ \varphi_d \circ \mathcal{A}) \\ &+(-\mathbf{I}_d \circ \cdots \circ \varphi_d \circ \mathbf{I}_d \circ \varphi_d \circ (-\mathcal{A})).\end{aligned} \tag{82}$$

The width of the ReLU DNN can also be arbitrarily large provided that its size is bigger than $2d$. Therefore, any affine transformation is representable by a suitably configured ReLU DNN.

Let $\mathcal{A}(\mathbf{x}) = \mathbf{h}_{\mathrm{LMMSE}}$ and there exits a DL estimator to represent $\mathbf{h}_{\mathrm{LMMSE}}$ from (81). As a result, the approximation error of the DL estimator in the linear systems is equal to zero. Moreover, Theorem 2 demonstrates that $J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) \xrightarrow{\mathbf{P}} J_{\mathrm{LMMSE}}$ as $|\mathcal{Z}|$ gets sufficiently large. Then, we reach the following conclusion for the linear systems as

$$J_{\mathrm{LMMSE}} \approx J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) \le J_{\mathrm{LS}}. \tag{83}$$

Remark 3:

- Corollary 1 and (83) demonstrate that the DL estimator is able to build up a stable and precise model to estimate $\mathbf{h}$ by using its universal approximation. Therefore, the DL estimator has a great potential to combat nonlinear distortion and some other unknown detrimental effects in real world communication systems, where the performance of the LS and LMMSE estimators degrades significantly.
- Theorem 2 implies that $J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})$ converges to $J(\mathbf{f}_{\boldsymbol{\theta}_o})$ in probability at a rate polynomially fast with $|\mathcal{Z}|$ if the underlying network structure is fixed. Such a result assures the efficiency and accuracy of the DL estimator when applied to channel estimation problems.
- Theorem 2 also demonstrates that the generalization error is increased with the network size if $|\mathcal{Z}|$ is fixed, but the approximation error can be reduced by enlarging the network size as indicated by Theorem 1. Hence, there exits a tradeoff between the generalization error and the approximation error as the network size of the DL estimator varies.
- Owing to no assumption about underlying signal model, the DL estimator has to take sufficiently large training data to train an effective estimator from scratch which is relatively inefficient compared to the LS or LMMSE estimators. In fact, we can retrain a learned DL estimator that is originally trained at similar scenarios to accelerate the training process as what the transfer learning has done in image processing [34].
- The channel estimate of the DL estimator derived through numerical optimization is only effective for a small range of the input space that contains training samples, though $\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}(\mathbf{x})$ is defined at a global scope. This phenomenon will be discussed in Section IV.

IV. Robustness to Mismatched Information

The optimality of the LMMSE and DL estimators depends on the perfect knowledge of channel statistics and matching training data, respectively, while it is a typical problem that the channel covariance matrix $\mathbf{G}$ is not perfectly known or the statistics of training data do not match the deployed environments. In this section, we analyze channel estimation with inaccurate channel statistics and mismatched training data for the linear system model in (1) and show how these imperfections affect the performance of the LMMSE and DL estimators.

A. LMMSE Estimator

Denote the channel covariance matrix used by the LMMSE estimator as $\boldsymbol{\Xi}_1 = \boldsymbol{\Xi} + \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is the $d \times d$ Hermitian random error matrix independent of $\boldsymbol{\Xi}$. Replacing $\boldsymbol{\Xi}$ by $\boldsymbol{\Xi}_1$ in (4), the LMMSE estimator under inaccurate channel statistic can be expressed as

$$\mathbf{h}_{\mathrm{LM-ER}} = \tau\boldsymbol{\Xi}_1(\boldsymbol{\Xi}_1 + \sigma_n^2\mathbf{I}_d)^{-1}\mathbf{x}, \tag{84}$$

and the corresponding MSE is given by

$$J_{\mathrm{LM-ER}} = \mathrm{tr}\left\{\left(\boldsymbol{\Xi}_1^{-1} + \frac{1}{\sigma_n^2}\mathbf{I}_d\right)^{-1} - \boldsymbol{\Pi}\boldsymbol{\Omega}\boldsymbol{\Pi}^T\right\}, \tag{85}$$

where $\boldsymbol{\Pi} = \mathbf{I}_d - \boldsymbol{\Xi}_1(\boldsymbol{\Xi}_1 + \sigma_n^2\mathbf{I}_d)^{-1}$.

It is difficult to figure out how $\boldsymbol{\Omega}$ affects the estimation accuracy of the LMMSE estimator directly from (85). However, we can take the uncorrelated channel as an example to demonstrate the influence of $\boldsymbol{\Omega}$ on $J_{\mathrm{LM-ER}}$ in general since the channels between different received antennas are asymptotically uncorrelated when d gets large [35].

Suppose that the covariance matrix of $\mathbf{h}$ is diagonal with $\boldsymbol{\Xi} = \sigma_c^2\mathbf{I}_d$, where σ_c^2 is element-wise variance. Moreover, assume that $\boldsymbol{\Omega}$ can be decomposed into

$$\boldsymbol{\Omega} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T, \tag{86}$$

where $\mathbf{U}$ is the $d \times d$ eigenvector matrix and $\boldsymbol{\Sigma}$ is the $d \times d$ eigenvalue matrix. Substituting $\boldsymbol{\Xi} = \sigma_c^2\mathbf{I}_d$ and (86) into (84),

we can rewrite $J_{\mathrm{LM-ER}}$ as

$$J_{\mathrm{LM-ER}} = J_{\mathrm{LMMSE}} + \sum_{i=1}^{d} \frac{\sigma_{e,i}^4 \sigma_n^4 d}{(\sigma_c^2 + \sigma_{e,i}^2 + \sigma_n^2)^2(\sigma_c^2 + \sigma_n^2)}, \tag{87}$$

where $\sigma_{e,i}^2$ is the i-th diagonal element of $\mathbf{\Sigma}$.

Reversely, if $\mathbf{\Xi} = \mathbf{\Xi}_1 + \mathbf{\Omega}$ and $\mathbf{\Xi} = \sigma_c^2 \mathbf{I}_d$, then we can still obtain the same form of $J_{\mathrm{LM-ER}}$ in (87).

Remark 4:

- According to (87), $J_{\mathrm{LM-ER}}$ is always larger than J_{LMMSE} and is increased with $\sigma_{e,i}^2$. The performance of the LMMSE estimator is mainly determined by the accuracy of $\mathbf{\Xi}$.
- The DL estimator needs to know neither the exact signal model nor the information of $\mathbf{\Xi}$ to estimate $\mathbf{h}$. Whether $\mathbf{\Xi}$ is accurate or not does not affect the accuracy of the DL estimator. Hence, the DL estimator will outperform the LMMSE estimator if $\sigma_{e,i}^2$ exceeds certain threshold, i.e., $J_{\mathrm{LM-ER}} \geq J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})$.

B. DL Estimator

The DL estimator is data-driven with its performance mainly determined by how well the training data matches the working environment and mismatched training data will lead to significant performance degradation. Furthermore, different from the LS and LMMSE estimators defined w.r.t. the whole input space, the learned DL estimator can only make valid channel estimates for the inputs that are restricted to the regions where training samples are not empty and will behave randomly outside these regions, as discussed in Remark 3. The restricted effective input range puts severe limit on the performance of the DL estimator.

Let

$$\mathcal{Z}_k = \{m \,|\, \mathbf{x}_m \in \mathcal{X}_k, m = 1, \ldots, |\mathcal{Z}|\} \tag{88}$$

be the set of index of samples that fall into $\mathcal{X}_k$. Note that $\sum_{i=1}^{K} |\mathcal{Z}_i| = |\mathcal{Z}|$. Using (13), we rewrite $J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})$ as

$$J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) = \frac{1}{|\mathcal{Z}|} \sum_{k=1}^{K} \sum_{m \in \mathcal{Z}_k} \mathrm{tr}\big\{(\mathbf{h}_m - \mathbf{W}_{\mathcal{X}_k}\mathbf{x}_m - \mathbf{b}_{\mathcal{X}_k}) (\mathbf{h}_m - \mathbf{W}_{\mathcal{X}_k}\mathbf{x}_m - \mathbf{b}_{\mathcal{X}_k})^T\big\}. \tag{89}$$

An important issue is that only a small number of partitioned regions within $\mathcal{X}$, where $\mathbf{x}$ falls into with high probabilities, contain training samples. For the regions without training samples, i.e., $|\mathcal{Z}_k| = 0$, the DL estimator is unable to optimize its estimate through $J_{\mathcal{Z}}(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}})$, as shown in (89), and will simply output a random channel estimate if $\mathbf{x}$ is located at these regions. In general, this limitation has a little impact on the performance of the DL estimator when training data has accurate statistics since the probability that $\mathbf{x}$ falls into the regions without training samples is negligible. However, if the statistics of training data do not match real channels, such a probability can not be ignored and the limitation on the effective input range will lead to serious issues.

Denote by $\mathbf{h}_{\mathrm{DL-ER}}$ and $J_{\mathrm{DL-ER}}$ the estimate and the MSE of the DL estimator, respectively, when training data mismatches real channel. The MSE of the DL estimator under mismatched training data is discussed in the following two separate cases.

1) Case I: Assume that $\mathbf{h}_{\mathrm{er}}$ distributes in a broader range than $\mathbf{h}$, where the variance of $\mathbf{h}_{\mathrm{er}}$ is larger than that of $\mathbf{h}$. The corresponding statistical models of the training data are described as

$$\mathbf{h}_{\mathrm{er}} = \mathbf{h} + \boldsymbol{\zeta}, \tag{90}$$

and

$$\mathbf{x}_{\mathrm{er}} = \tau \mathbf{h}_{\mathrm{er}} + \mathbf{n}, \tag{91}$$

where $\mathbf{h}$ is Gaussian distributed and $\boldsymbol{\zeta}$ denotes the $d \times 1$ zero mean random error vector that is independent of $\mathbf{h}$ with covariance matrix $\mathbf{\Omega}_\zeta = \mathbb{E}\{\boldsymbol{\zeta}\boldsymbol{\zeta}^T\}$.

In Case 1, the probability that $\mathbf{x}$ falls into the regions without training samples is still close to zero since $\mathbf{x}_{\mathrm{er}}$ is more broadly distributed than $\mathbf{x}$. From Corollary 1, the target estimator that the DL estimator approaches to as $|\mathcal{Z}|$ gets large is the MMSE estimator w.r.t. $\mathbf{h}_{\mathrm{er}}$, and the corresponding channel estimate is given by

$$\mathbf{h}_{\mathrm{MM-ER}} = \mathbf{C}_{\mathbf{h}_{\mathrm{er}}\mathbf{x}_{\mathrm{er}}} \mathbf{C}_{\mathbf{x}_{\mathrm{er}}\mathbf{x}_{\mathrm{er}}}^{-1} \mathbf{x}, \tag{92}$$

where $\mathbf{C}_{\mathbf{h}_{\mathrm{er}}\mathbf{x}_{\mathrm{er}}}$ is the cross-covariance of $\mathbf{h}_{\mathrm{er}}$ and $\mathbf{x}_{\mathrm{er}}$ and $\mathbf{C}_{\mathbf{x}_{\mathrm{er}}\mathbf{x}_{\mathrm{er}}}$ is the covariance of $\mathbf{x}_{\mathrm{er}}$. If the DL estimator is properly configured and $|\mathcal{Z}|$ is sufficiently large, then we have

$$\mathbf{h}_{\mathrm{DL-ER}} \approx \mathbf{h}_{\mathrm{MM-ER}} \tag{93}$$

according to Corollary 1. From (16), the corresponding MSE is

$$J_{\mathrm{DL-ER}} \approx J_{\mathrm{LMMSE}} + \|(\mathbf{C}_{\mathbf{h}_{\mathrm{er}}\mathbf{x}_{\mathrm{er}}} \mathbf{C}_{\mathbf{x}_{\mathrm{er}}\mathbf{x}_{\mathrm{er}}}^{-1} - \mathbf{C}_{\mathbf{h}\mathbf{x}} \mathbf{C}_{\mathbf{x}\mathbf{x}}^{-1})\mathbf{x}\|_2^2, \tag{94}$$

where $\mathbf{C}_{\mathbf{x}\mathbf{x}}$ is the covariance of $\mathbf{x}$.

Similar to $J_{\mathrm{LM-ER}}$ in (87), how $\boldsymbol{\zeta}$ affects $J_{\mathrm{DL-ER}}$ is difficult to justify from (94). To provide some insight into the influence of the mismatched training data on the DL estimator, we assume that $\mathbf{\Xi} = \sigma_c^2 \mathbf{I}_d$. Then, the covariance matrix $\mathbf{\Omega}_\zeta$ can be decomposed into

$$\mathbf{\Omega}_\zeta = \mathbf{U}_\zeta \mathbf{\Sigma}_\zeta \mathbf{U}_\zeta^T, \tag{95}$$

where $\mathbf{U}_\zeta$ is the $d \times d$ eigenvector matrix and $\mathbf{\Sigma}_\zeta$ is the $d \times d$ eigenvalue matrix. Substituting $\mathbf{\Xi} = \sigma_c^2 \mathbf{I}_d$ and (95) into (94) yields

$$J_{\mathrm{DL-ER}} \approx J_{\mathrm{LMMSE}} + \sum_{i=1}^{d} \frac{\sigma_{\zeta,i}^4 \sigma_n^4}{(\sigma_c^2 + \sigma_{\zeta,i}^2 + \sigma_n^2)^2(\sigma_c^2 + \sigma_n^2)}, \tag{96}$$

where $\sigma_{\zeta,i}^2$ is the i-th diagonal element of $\mathbf{\Sigma}_\zeta$ and quantifies the mismatch degree between the training data and the real systems. The obtained $J_{\mathrm{DL-ER}}$ in (96) is similar to $J_{\mathrm{LM-ER}}$ in (87) and also increases with $\sigma_{\zeta,i}^2$.

2) Case II: We consider that the input-output pair of training data is generated from the following statistical model

$$\mathbf{h} = \mathbf{h}_{\rm er} + \boldsymbol{\zeta}, \tag{97}$$

and

$$\mathbf{x}_{\rm er} = \tau \mathbf{h}_{\rm er} + \mathbf{n}. \tag{98}$$

In Case II, $\mathbf{x}$ distributes in a broader range than $\mathbf{x}_{\rm er}$, and the probability that $\mathbf{x}$ falls at regions without training samples is much higher than Case I. From (89), the DL estimator is not optimized for the whole input space, and its effective input range is dependent on the training data distribution. The channel estimates of the DL estimator corresponding to the inputs at empty regions are totally random and unacceptable if the discrepancy between $\mathbf{h}$ and $\mathbf{h}_{\rm er}$ is very large. In this case, the DL estimator basically fails to provide a reliable channel estimate, and its performance degrades severely.

Remark 5:

- In Case I, the probabilities that the inputs are located at the regions without training samples are negligible. The DL estimator can well approximate $\mathbf{h}_{\rm MM-ER}$ and provide a stable channel estimate. Hence, the limited effective input range has a little impact on the performance of the DL estimator.
- In Case II, the probabilities that the inputs are located at the regions without training samples can not be neglected. The limitation on the effective input range of the DL estimator gets really serious, and the DL estimator is unable to provide a valid channel estimate when the input is located outside the regions with training samples. The LMMSE estimator, however, is designed over the whole input space based on the expert knowledge, and therefore the error introduced by the discrepancy between $\boldsymbol{\Xi}$ and $\boldsymbol{\Xi}_1$ is controllable no matter how $\boldsymbol{\Omega}$ varies. In this case, the traditional LMMSE estimator is more robust to the imperfect data than the DL estimator.
- An important issue for the DL estimator is to incorporate traditional signal processing techniques to enhance its robustness to imperfect information rather than purely relying on training data [36]. Hence, it is a major topic on how to take advantage of model based approaches to improve the performance of the DL estimator .

V. Simulation Results

In this section, computer simulation is conducted to provide further evidence and insights into the performance assessment of various estimators, which also verifies the advantages and disadvantages of the DL channel estimation.

A. Linear Systems

Fig. 2 compares the MSEs of the LS, LMMSE, and DL estimators versus SNR under linear signal model (1). The channel, $\mathbf{h}$, is assumed to be Gaussian with zero mean and element-wise unit variance. The sizes of training and test sets are $20,000$ and $5,000$, respectively. The underlying network of the DL estimator has 4 layers and equal numbers of neurons at each hidden layer, i.e., equal widths. Denote $\tilde{d}$ as the width of

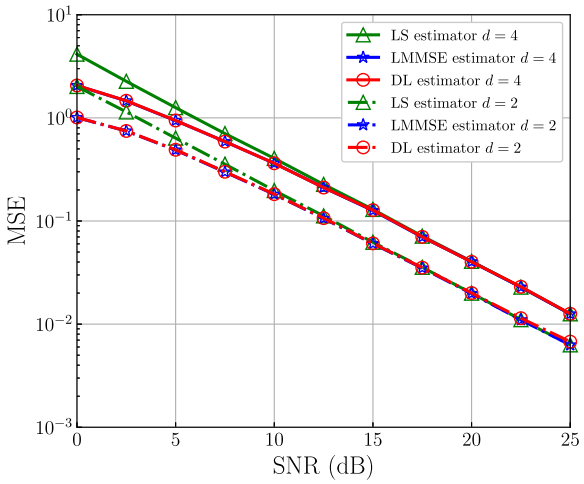


Fig. 2. The MSE performance of the LS, LMMSE, and DL estimators versus SNR under linear signal model.

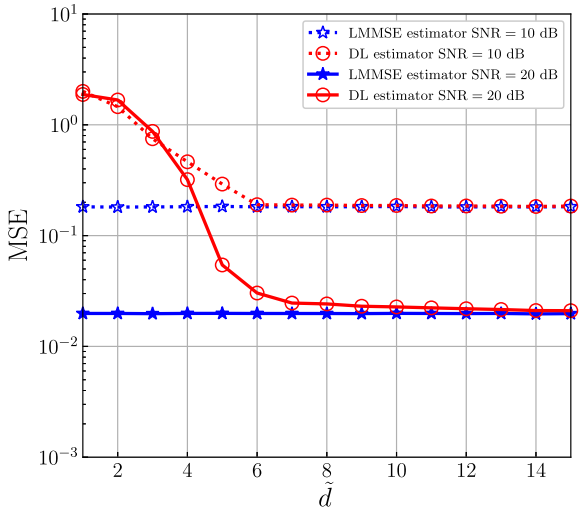


Fig. 3. The MSE performance of the DL estimator versus $\tilde{d}$ under linear signal model.

hidden layer and $\tilde{d}$ is set to be 40. From Fig. 2, the MSEs of the LMMSE and the DL estimators are almost overlapped. Since the LMMSE estimator is equivalent to the MMSE estimator in this case, the DL estimator can well approximate $\mathbf{h}_{\rm MMSE}$, which confirms that $J(\mathbf{f}_{\boldsymbol{\theta}_{\mathcal{Z}}}) \approx J_{\rm LMMSE}$ in the linear systems. Moreover, both the DL and LMMSE estimators outperform the LS estimator in Fig. 2 as noted by (83).

Fig. 3 shows the MSEs of the DL estimator versus the width of ReLU DNN, $\tilde{d}$, under linear signal model (1) with fixed SNRs, sample size, and $d = 2$. The MSEs of the LMMSE estimators derived under the same SNRs are used as the benchmark. The approximation error becomes the main factor that affects the MSE of the DL estimator in this case since $|\mathcal{Z}|$ is sufficiently large. When $\tilde{d}$ is small, the dimension of the parameter space Θ is very low and the approximation error becomes relatively high. As a result, the MSEs of the DL estimator are significantly larger than the MSEs of the LMMSE estimator. As $\tilde{d}$ increases, the parameter space Θ is enlarging and the approximation error decreases with it.

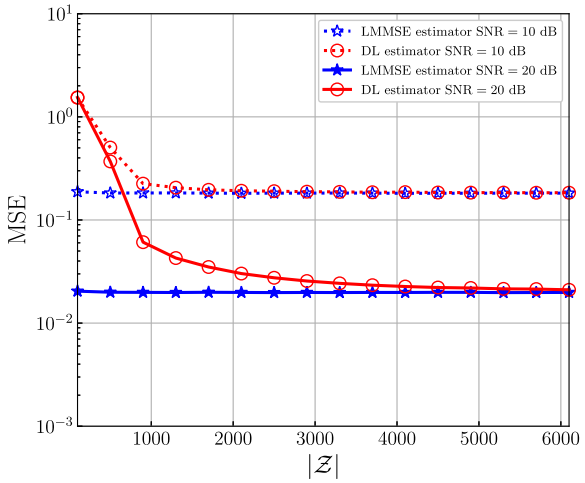


Fig. 4. The MSE performance of the DL estimator versus $|\mathcal{Z}|$ under linear signal model.

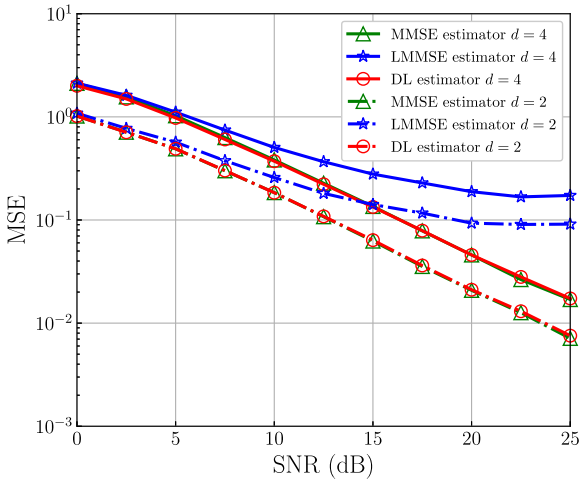


Fig. 5. The MSE performance of the MMSE, LMMSE, and DL estimators versus SNR under nonlinear signal model.

All MSEs of the DL estimator at different SNRs approach to the MSEs of the LMMSE estimator. Such a result evidences the conclusions in Theorem 1.

Fig. 4 shows the MSEs of the DL estimator versus the size of training samples, $|\mathcal{Z}|$, under linear signal model (1) with fixed SNRs, network structure, and $d = 2$. As in Fig. 4, the LMMSE estimator is used as the benchmark. In this case, the generalization error determines the performance of the DL estimator. When $|\mathcal{Z}|$ is small, the generalization error is very high and the MSEs of the DL estimator are significantly larger than the MSEs of the LMMSE estimator. As $|\mathcal{Z}|$ increases, the generalization error decreases and all MSEs of the DL estimator at different SNRs asymptotically approach to the MSEs of the LMMSE estimator. Such a result evidences the conclusions in Theorem 2.

B. Nonlinear Systems

In this subsection, we evaluate the performance of the MMSE, LMMSE and DL estimators under a nonlinear signal model. Let $\mathbf{x}_{\text{in}} = \mathbf{h}\tau + \mathbf{n}$ and the following nonlinear model

$$x_i = x_{\text{in},i}\left(1 + \left(\frac{x_{\text{in},i}}{x_{\text{sat}}}\right)^{2\omega}\right)^{-\frac{1}{2\omega}} \tag{99}$$

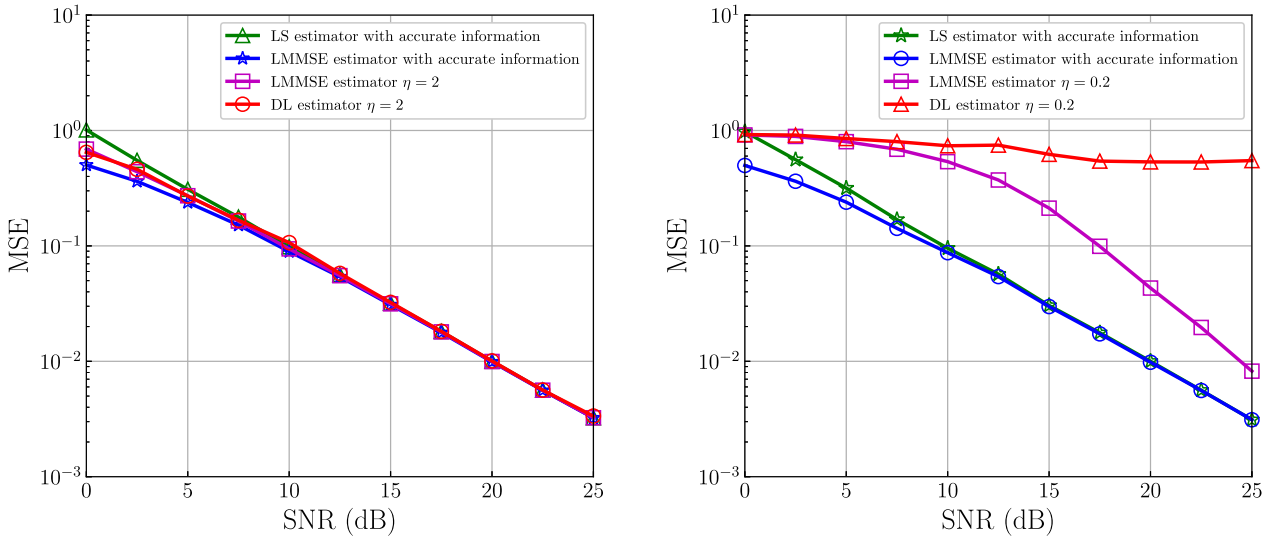
is adopted, where x_i and $x_{\text{in},i}$ are the i-th elements of $\mathbf{x}$ and $\mathbf{x}_{\text{in}}$, respectively, for $i \in \{1, \ldots, d\}$, x_{sat} is the saturation level, and ω is the smoothness factor. The other settings are the same as in Section V-A. The model in (99) is typically used by nonlinear signal detection caused by imperfection of PA and is commonly known as Rapp model [37]. Here, we apply such a model to illustrate channel estimation for nonlinear systems.

Fig. 5 shows the MSEs of the MMSE, LMMSE, and DL estimators versus SNR under nonlinear model in (99), where the saturation level, x_{sat}, is fixed as 1.5 and the smoothness factor ω is set be 1. Since no analytical form of $\mathbf{h}_{\text{MMSE}}$ for nonlinear model (99) is available, we use Monte Carlo simulation to estimate $\mathbf{h}_{\text{MMSE}}$ in Fig. 5 and the number of trials is set as 2×10^7. The performance of the MMSE, LMMSE, and DL estimators is close to each other at low SNRs as the noises dominate the overall MSEs. As the SNR increases, the approximation errors to the MMSE estimator will contribute a larger percentage of the MSEs. According to Theorem 1, the approximation error of the DL estimator is significantly lower than that of the LMMSE estimator for nonlinear systems. As a result, the MSE of the DL estimator is very close to that of the MMSE estimator and becomes significantly better than that of the LMMSE estimator for high SNRs.

C. Robustness to Mismatched Information

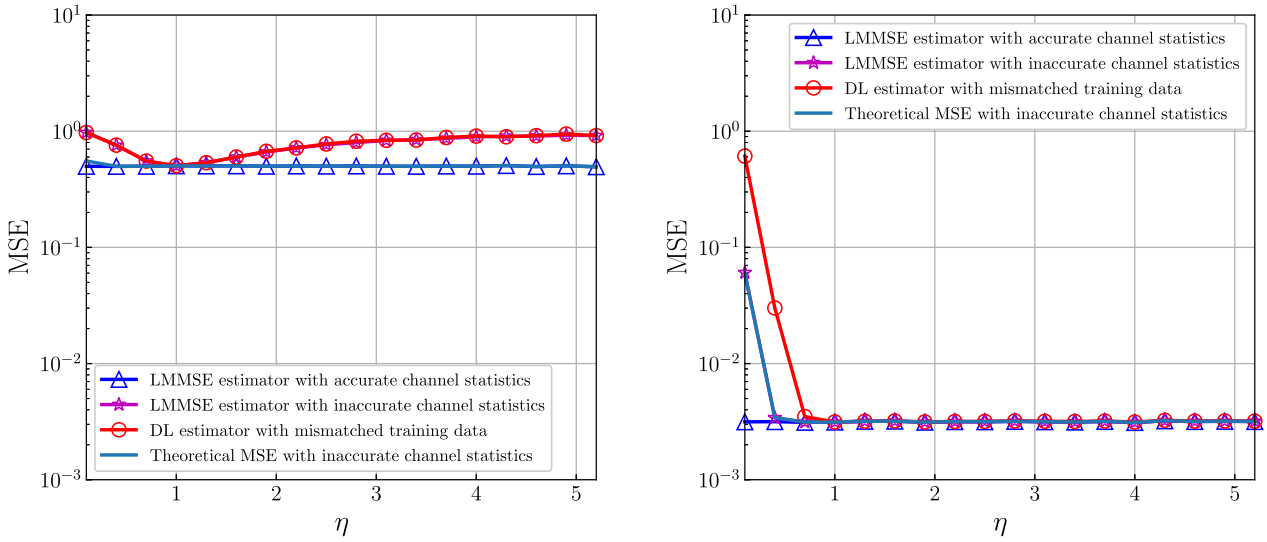
We then compare the MSEs of channel estimation using the LMMSE and the DL estimators under inaccurate statistics of channel and mismatched training data in the linear systems. Assume that $\mathbf{\Xi}_1 = \sigma_{c,1}^2\mathbf{I}_d$ and $\mathbf{\Xi}_{\text{er}} = \sigma_{\text{er}}^2\mathbf{I}_d$, where $\sigma_{c,1}^2$ and σ_{er}^2 are the element-wise variances. Moreover, we define the scaling coefficient η as the ratio $\sigma_{c,1}^2/\sigma^2$ or $\sigma_{\text{er}}^2/\sigma^2$. The other settings are the same as Section V-A. When $\eta > 1$, i.e., Case I in Section IV-B, the performance of the DL estimator is only affected by the degree of the mismatch for training data with real channel statistics. When $\eta < 1$, i.e., Case II in Section IV-B, the DL estimator may malfunction and outputs random estimates due to the restricted effective input range.

Fig. 6(a) illustrates the MSEs of the LMMSE and the DL estimators versus SNR under the linear signal model in (1) with $d = 1$ and $\eta = 2$ that corresponds to Case I. The MSEs of the LMMSE estimator with accurate channel statistics and the LS estimator are served as the benchmarks. In Fig. 6(a), both the LMMSE estimator with inaccurate channel statistics and the DL estimator with mismatched training data perform poorer than the LMMSE estimator with accurate channel statistics but still better than the LS estimator. Furthermore, under the same η, the MSEs of the LMMSE with inaccurate statistics and the DL estimator with mismatched training data are overlapped, which confirms (87) and (96). Specifically, in high SNRs, the MSEs of these estimators are almost the same and the errors of channel statistics have little impact on the overall estimation performance.



(a) Case I ($\eta = 2 > 1$) (b) Case II ($\eta = 0.2 < 1$)

Fig. 6. The MSE performance of the LMMSE and the DL estimators versus SNR under inaccurate statistics of channel and mismatched training data.



(a) SNR = 0 dB (b) SNR = 25 dB

Fig. 7. The MSE performance of the LMMSE and the DL estimators versus η under inaccurate statistics of channel and mismatched training data.

Fig. 6(b) illustrates the MSEs of the LMMSE and the DL estimators versus SNR under linear signal model in (1) with $d = 1$ and $\eta = 0.2$ that corresponds to Case II. When $\eta < 1$, the MSE of the LMMSE estimator with inaccurate channel statistics is significantly larger than those of the LS and LMMSE estimators with accurate channel statistics. The performance of the LMMSE estimator with inaccurate channel statistics degrades more severely than Case I in Fig. 6(a). The performance of the DL estimator is even worse since its MSE is totally random and uncorrelated to the SNR. Such phenomenon verifies the analysis in Section IV-B when the variance of training data is lower than that of true channel, i.e., $\eta < 1$. Therefore, the mismatch of training data with the true environment is a serious problem if $\eta < 1$ and should be carefully considered when applying DL methods to wireless communication systems.

Fig. 7(a) visualizes the MSEs of the LMMSE estimator with inaccurate channel statistics and the DL estimator with mismatched training data versus η under linear signal model (1) with $d = 1$ and SNR $= 0$ dB. We adopt the MSE of the LMMSE estimator with inaccurate channel statistics in (87) as the theoretical MSE. In Fig. 7(a), the MSEs of the LMMSE estimator with inaccurate channel statistics and the DL estimator with mismatched training data are overlapped with the theoretical MSE and are slightly higher than the MSE of the LMMSE estimator with accurate channel statistics when $\eta > 1$. Such a result verifies the correctness of (87) and (96), as (96) is equivalent to (87) under the same η.

When $\eta = 1$, there is no error. When $\eta < 1$, the MSEs of the LMMSE estimator with inaccurate channel statistics and the DL estimator with mismatched training data are larger than that of the LMMSE estimator with accurate channel statistics. As illustrated in Fig 6(b), the MSE of the LMMSE estimator with inaccurate channel statistics is comparable to that of the DL estimator with mismatched training data at low SNRs. Therefore, the gap between the MSEs of the LMMSE estimator with inaccurate channel statistics and the DL estimator with mismatched training data is not significant in Fig. 7(a).

Fig. 7(b) shows the MSEs of the LMMSE estimator with inaccurate channel statistics and the DL estimator with mismatched training data versus η when $d = 1$ and $\mathrm{SNR} = 25$ dB. When $\eta \geq 1$, the MSEs of the LMMSE estimator with inaccurate channel statistics and the DL with mismatched training data are almost equal to that of the LMMSE estimator. When $\eta < 1$, the MSE of the DL estimator with mismatched training data is significantly larger than that of the LMMSE estimator with inaccurate channel statistics and the theoretical MSE. The reason for this phenomenon is that the estimate of the DL estimator with mismatched training data gets more random as η decreases and is nearly uncorrelated to the SNR when $\eta < 1$, as shown in Fig. 6(b), while the MSE of the LMMSE with inaccurate channel statistics still decreases with the SNR. Therefore, the gap between the MSEs of the LMMSE estimator with inaccurate channel statistics and the DL estimator with mismatched training data becomes much more significant at high SNRs. Such a result verifies the analysis in Section IV-B again. Moreover, the MSE of the LMMSE estimator with inaccurate channel statistics matches the theoretical MSE and is slightly higher than the MSE of the LMMSE estimator with accurate channel statistics across the entire range of η, which shows the robustness of the LMMSE estimator to inaccurate channel statistics.

VI. Conclusion

In this paper, we have made the first attempt on interpreting DL based channel estimation under linear, nonlinear, and inaccurate channel statistics using a multiple antenna system as an example. We have explained that the DL estimator equipped with a ReLU DNN is mathematically equivalent to a piecewise linear function and can attain universal approximation to the MMSE estimator under suitably configured structure and large training samples. Extensive simulation results have confirmed the performance of the DL estimator and showed that the DL estimator is close to the LMMSE estimator under linear systems but significantly outperforms it when the signal model is nonlinear. However, the DL estimator is sensitive to the quality of training data and its performance would significantly degrade if the data in real environments distributes broader than the training data. The benefits of the DL estimator have to weigh against its costs when applied to the channel estimation in real wireless communication systems. We should strike a balance between DL based channel estimation and traditional channel estimation. An important issue of future analysis is to incorporate traditional signal processing techniques into the DL estimator to alleviate the influence of imperfect training data and improve the robustness.

References

[1] T. Wang, C.-K. Wen, H. Wang, F. Gao, T. Jiang, and S. Jin, "Deep learning for wireless physical layer: Opportunities and challenges," *China Commun.*, vol. 14, no. 11, pp. 92–111, Nov. 2017.

[2] X. Gao, S. Jin, C.-K. Wen, and G. Y. Li, "ComNet: Combination of deep learning and expert knowledge in OFDM receivers," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2627–2630, Dec. 2018.

[3] X. Yang, M. Matthaiou, J. Yang, C.-K. Wen, F. Gao, and S. Jin, "Hardware-constrained millimeter-wave systems for 5G: Challenges, opportunities, and solutions," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 44–50, Jan. 2019.

[4] Z. Qin, H. Ye, G. Y. Li, and B.-H.-F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019.

[5] L. Liang, H. Ye, and G. Y. Li, "Toward intelligent vehicular networks: A machine learning framework," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 124–135, Feb. 2019.

[6] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based channel estimation for beamspace mmWave massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 852–855, Oct. 2018.

[7] H. Ye, G. Y. Li, and B.-H.-F. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 114–117, Feb. 2018.

[8] T. Wang, C.-K. Wen, S. Jin, and G. Y. Li, "Deep learning-based CSI feedback approach for time-varying massive MIMO channels," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 416–419, Apr. 2019.

[9] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cognit. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Dec. 2017.

[10] H. Ye, G. Y. Li, and B.-H.-F. Juang, "Deep reinforcement learning based resource allocation for V2 V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.

[11] L. Liang, H. Ye, G. Yu, and G. Y. Li, "Deep-learning-based wireless resource allocation with application to vehicular networks," *Proc. IEEE*, vol. 108, no. 2, pp. 341–356, Feb. 2020.

[12] H. Huang, J. Yang, H. Huang, Y. Song, and G. Gui, "Deep learning for super-resolution channel estimation and DOA estimation based massive MIMO system," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8549–8560, Sep. 2018.

[13] Y. Yang, F. Gao, X. Ma, and S. Zhang, "Deep learning-based channel estimation for doubly selective fading channels," *IEEE Access*, vol. 7, pp. 36579–36589, Mar. 2019.

[14] M. Soltani, V. Pourahmadi, A. Mirzaei, and H. Sheikhzadeh, "Deep learning-based channel estimation," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 652–655, Apr. 2019.

[15] Y. Yang, F. Gao, G. Y. Li, and M. Jian, "Deep learning-based downlink channel prediction for FDD massive MIMO system," *IEEE Commun. Lett.*, vol. 23, no. 11, pp. 1994–1998, Nov. 2019.

[16] S. Dorner, S. Cammerer, J. Hoydis, and S. T. Brink, "Deep learning based communication over the air," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 132–143, Feb. 2018.

[17] H. Ye, G. Y. Li, B.-H.-F. Juang, and K. Sivanesan, "Channel agnostic End-to-End learning based communication systems with conditional GAN," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–5.

[18] H. Ye, L. Liang, G. Y. Li, and B.-H.-F. Juang, "Deep learning-based End-to-End wireless communication systems with conditional GANs as unknown channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3133–3143, May 2020.

[19] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 748–751, Oct. 2018.

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

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

[22] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Art. Intell. Stat. (AISTATS)*, Apr. 2011, pp. 315–323.

[23] G. F. Montúfar, R. Pascanu, K. Cho, and Y. Bengio, "On the number of linear regions of deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeuraIPS)*, Dec. 2014, pp. 2924–2932.

[24] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. S. Dickstein, "On the expressive power of deep neural networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 2847–2854.

[25] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, "Understanding deep neural networks with rectified linear units," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–17.
[26] S. M. Kay, *Fundamentals of Statistical Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.
[27] E. Costa, M. Midrio, and S. Pupolin, "Impact of amplifier nonlinearities on OFDM transmission system performance," *IEEE Commun. Lett.*, vol. 3, no. 2, pp. 37–39, Feb. 1999.
[28] E. Costa and S. Pupolin, "M-QAM-OFDM system performance in the presence of a nonlinear amplifier and phase noise," *IEEE Trans. Commun.*, vol. 50, no. 3, pp. 462–472, Mar. 2002.
[29] L. Xu, X. Lu, S. Jin, F. Gao, and Y. Zhu, "On the uplink achievable rate of massive MIMO system with low-resolution ADC and RF impairments," *IEEE Commun. Lett.*, vol. 23, no. 3, pp. 502–505, Mar. 2019.
[30] H. L. Royden and P. M. Fitzpatrick, *Real Analysis*. Upper Saddle River, NJ, USA: Prentice-Hall, 2010.
[31] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*. Cham, Switzerland: Springer, 1996.
[32] W. Mendenhall, R. J. Beaver, and B. M. Beaver, *Introduction to Probability and Statistics*. Boston, MA, USA: Cengage, 2012.
[33] S. van de Geer, "Estimating a regression function," *Ann. Statist.*, vol. 18, no. 2, pp. 907–924, Jun. 1990.
[34] S. Jialin Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[35] F. Rusek, D. Persson, B. Kiong Lau, E. G. Larsson, T. L. Marzetta, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.
[36] H. He, S. Jin, C.-K. Wen, F. Gao, G. Y. Li, and Z. Xu, "Model-driven deep learning for physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 77–83, Oct. 2019.
[37] J. Joung, C. K. Ho, K. Adachi, and S. Sun, "A survey on power-amplifier-centric techniques for Spectrum- and energy-efficient wireless communications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 315–333, 1st Quart., 2015.

Qiang Hu received the B.S. degree in applied physics and the M.S. degree in telecommunication engineering from the Beijing University of Posts and Communications, China, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree with Tsinghua University. His research interests include machine learning theory, statistical signal processing, and convex optimization.

Feifei Gao (Fellow, IEEE) received the B.Eng. degree from Xi'an Jiaotong University, Xi'an, China, in 2002, the M.Sc. degree from McMaster University, Hamilton, ON, Canada, in 2004, and the Ph.D. degree from the National University of Singapore, Singapore, in 2007. He was a Research Fellow with the Institute for Infocomm Research, A*STAR, Singapore in 2008, and an Assistant Professor with the School of Engineering and Science, Jacobs University, Bremen, Germany, from 2009 to 2010. In 2011, he joined the Department of Automation, Tsinghua University, Beijing, China, where he is currently an Associate Professor. His research areas include communication theory, signal processing for communications, array signal processing, and convex optimizations, with an emphasis on MIMO techniques, multi-carrier communications, cooperative communication, and cognitive radio networks. He has authored or coauthored more than 150 refereed IEEE journal articles and more than 150 IEEE conference proceeding articles. He has served as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE SIGNAL PROCESSING LETTERS, the IEEE COMMUNICATIONS LETTERS, the IEEE WIRELESS COMMUNICATIONS LETTERS, and *China Communications*, and a Senior Editor for the IEEE SIGNAL PROCESSING LETTERS and the IEEE COMMUNICATIONS LETTERS. He has also served as the Symposium Co-Chair for the 2014 IEEE Global Communications Conference, the 2014 IEEE Vehicular Technology Conference Fall (VTC), the 2015 IEEE Conference on Communications (ICC), 2018 IEEE VTC, and 2019 IEEE ICC, and a technical committee member for many other IEEE conferences.

Hao Zhang received the B.S. and M.S. degrees in applied mathematics and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 1995, 1997, and 2001, respectively. Since 2001, he has been with the Department of Electronic Engineering, Tsinghua University, where he is currently an Associate Professor. His research interests include high-resolution spectral analysis, array processing, and advanced statistical and intelligent techniques applied to signal processing.

Shi Jin (Senior Member, IEEE) received the B.S. degree in communications engineering from the Guilin University of Electronic Technology, Guilin, China, in 1996, the M.S. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2003, and the Ph.D. degree in information and communications engineering from the Southeast University, Nanjing, in 2007. From 2007 to 2009, he was a Research Fellow with the Adastral Park Research Campus, University College London, London, U.K. He is currently with the Faculty of the National Mobile Communications Research Laboratory, Southeast University. His research interests include space–time wireless communications, random matrix theory, and information theory. He served as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE COMMUNICATIONS LETTERS, and *IET Communications*. He and his coauthors received the 2010 Young Author Best Paper Award and the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communication theory from the IEEE Signal Processing Society.

Geoffrey Ye Li (Fellow, IEEE) was with the Georgia Institute of Technology, GA, USA, as a Professor for 20 years and with AT&T Bell Laboratories Research, NJ, USA, as a Principle Technical Staff Member for five years. In 2020, he moved to the Imperial College London, U.K., where he is currently the Chair Professor. His general research interests include statistical signal processing and machine learning for wireless communications. He has authored or coauthored over 500 journals and conference articles in the related areas. He holds over 40 granted patents. His publications have been cited over 40 000 times and he has been recognized as a Highly Cited Researcher, by Thomson Reuters, almost every year.

He was elected as an IEEE Fellow for his contributions to signal processing for wireless communications in 2005. He received several prestigious awards, including the Donald G. Fink Overview Paper Award in 2017 from the IEEE Signal Processing Society, the James Evans Avant Garde Award in 2013 and the Jack Neubauer Memorial Award in 2014 from the IEEE Vehicular Technology Society, and the Stephen O. Rice Prize Paper Award in 2013, the Award for Advances in Communication in 2017, and the Edwin Howard Armstrong Achievement Award in 2019 from the IEEE Communications Society. He also received the 2015 Distinguished ECE Faculty Achievement Award from the Georgia Institute of Technology.

Dr. Ye Li has been involved in editorial activities for over 20 technical journals, including the founding Editor-in-Chief of the IEEE JSAC Special Series on ML in Communications and Networking. He has organized and chaired many international conferences, including the Technical Program Vice Chair of the IEEE ICC'03, the General Co-Chair of the IEEE GlobalSIP'14, the IEEE VTC'19 (Fall), and the IEEE SPAWC'20.