Generalization Performance of Empirical Risk Minimization on Over-Parameterized Deep ReLU Nets

Shao-Bo Lin, Yao Wang, and Ding-Xuan Zhou

Abstract—In this paper, we study the generalization performance of global minima of empirical risk minimization (ERM) on over-parameterized deep ReLU nets. Using a novel deepening scheme for deep ReLU nets, we rigorously prove that there exist perfect global minima achieving optimal generalization error rates for numerous types of data under mild conditions. Since over-parameterization of deep ReLU nets is crucial to guarantee that the global minima of ERM can be realized by the widely used stochastic gradient descent (SGD) algorithm, our results present a potential way to fill the gap between optimization and generalization of deep learning.

Index Terms—Deep learning, empirical risk minimization, global minima, over-parameterization.

I. INTRODUCTION

Deep learning [1] that conducts feature extraction and statistical modelling on a unified deep neural network (deep net) framework has attracted enormous research activities in the past decade. It has made significant breakthroughs in numerous applications including computer vision [2], speech recognition [3] and go games [4]. The great success of deep learning brings several challenges in understanding its running mechanism, including the landscape analysis [5], convergence analysis [6], benign over-fitting verification [7], and generalization optimality demonstration [8].

In this paper, we focus on benign over-fitting verification and generalization optimality demonstration for over-parameterized deep nets (OPDNN). As shown in Figure 1, generalization performances of specific deep learning algorithms are determined by the approximation errors that reflect the suitability of the hypothesis space, i.e. deep nets in our study, sample errors which refer to the qualities of the adopted optimization strategies, and optimization errors that quantitatively describe the efficiency of designed optimization algorithms. A deep learning algorithm of high quality is achieved by balancing approximation error, sample error and optimization error, simultaneously. Unfortunately, there lacks unified generalization error analysis for deep learning algorithms in the sense that efforts on statistical learning theory [9], [10], [11] only payed attentions to the approximation error and sample error to pursue theoretical advantages of deep learning models, while progress on non-convex optimization only focused on optimization error [6], [12], [13] for some iterative algorithms such as gradient descent (GD) and stochastic gradient descent (SGD) to describe the differences between models and algorithms.

A. Problem Setting

It seems natural to derive perfect generalization error rates for deep learning algorithms by combining the existing results in [6], [9], [10], [11], [12], and [13]. The problem is, however, that good generalization for a deep learning model is derived based on the well known bias-variance trade-off principle [14], requiring the number of free parameters to be smaller than the size of data (under-parameterized setting), while provable optimization algorithm requires huge free parameters that are much more than the size of data (over-parameterized setting). This inconsistency between generalization for model selection and optimization for algorithm design makes the running mechanism of deep learning be still a mystery.

Roughly speaking, there are two scenarios to settle the aforementioned inconsistency. One is to pursue the convergence guarantee for optimization algorithms on the size of data (under-parameterized setting), while provable optimization algorithm requires huge free parameters that are much more than the size of data (over-parameterized setting). This inconsistency between generalization for model selection and optimization for algorithm design makes the running mechanism of deep learning be still a mystery.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
under-parameterized deep ReLU nets (UPDNN) in [9], [10], and [11] to guarantee perfect generalization properties of deep nets. Two main obstacles for this scenario are different structures to guarantee optimal generalization performances for different types of data [9], [10], [11] and complex landscapes [15] for UPDNN. In this way, it is quite difficult to design a provable and unified iterative algorithm to solve different UPDNN models. The other is to present generalization error for ERM-OPDNN. Due to the power of width

In the landscape analysis [15] and global convergence [6], [13], generalization error rates for gradient-type algorithms have already been derived in [16] and [17] under some non-trivial assumptions on the data. These interesting results provide springboards to deduce optimal generalization error for deep learning algorithms. However, as shown in Figure 2, the step-stones of the optimization-based analysis approach are local convergence of gradient-type algorithms for specific initialization points such as the random orthogonal initialization [18] and MSRA initialization [19], and sample error analysis in the neighbourhood around them under some data separation or separable by random features assumptions [16], [20]. The pros of this optimization-based analysis approach is the provable convergence property of algorithms but the cons is the additional (somewhat strong) data assumption, making the optimal generalization error rates unclear. Furthermore, local convergence and given initialization schemes force the algorithm to find estimates near such initialization points, leaving a critical gap to judge whether there are perfect global minima (global minima possessing optimal generalization error rates) around them.

Noting the above dilemma of optimization-based analysis, we drive a different direction as shown in Figure 3. Our road-map begins with perfect generalization error analysis for ERM-OPDNN and then pursues the local convergence of optimization algorithms. In the following, we aim to find a suitable initialization point near the global minima of ERM-OPDNN with perfect generalization performance. Under this circumstance, the optimal generalization error analysis can be easily deduced. Such an approximation-based road-map inevitably arises three important problems:

**Problem 1:** Are there global minima of ERM-OPDNN achieving optimal generalization error rates?

**Problem 2:** How to develop an iterative algorithm possessing local convergence for any initialization points?

**Problem 3:** How to find an initialization point near the global minima in Problem 1?

Compared with Problems 1 and 3, Problem 2 is not so difficult since the local convergences of non-convex optimizer have already been established in [21], [22], [23], and [24], though Problem 2 requires more extensive results on the local convergence than those presented in [6] and [13] in the sense that local convergence should hold for any initialization points rather than specific initialization schemes. The answer to Problem 3, in addition, depends heavily on that to Problem 1, since the local convergence only guarantees the convergence of deep learning algorithms near the initialization points that must be selected close to the perfect global minima of ERM-OPDNN. Under this circumstance, it is urgent to present an affirmative answer to Problem 1, which is the purpose of this paper.

**B. Related Work**

The inconsistency between generalization and optimization stimulates to rethink the classical bias-variance trade-off in modern machine learning practice [25], since evidences in [7] illustrated that there are OPDNNs generalizing well despite they achieved extremely small training errors. In particular, [26] constructed an exact interpolation of training data that achieves optimal generalization error rates based on Nadaraya-Watson kernel estimates; [27] derived a sufficient condition for the data under which the global minimum of over-parameterized linear regression possesses excellent generalization performance; [28] studied the generalization performance of kernel-based least norm interpolation and presented generalization error estimates under some restrictions on the data distribution. Similar results can also be found in [29], [30], [31], [32], and [33] and references therein. All these exciting results present novel insights in developing learning algorithms to avoid the traditional bias-variance dilemma. It should be highlighted that the core of these results is to figure out restrictions on data distributions so that all global minima for over-parameterized linear models are good learners, which does not hold for deep learning, since it is easy to provide a counterexample for benign over-fitting of deep ReLU nets, even for noise-free data (see Proposition 1 below).

Since the above existing theoretical analysis for benign over-fitting is only available to convex linear models while ERM-OPDNN involves highly nonconvex nonlinear models, novel approaches such as the neural tangent kernel technique [34] and local convergence analysis [16] were developed. Utilizing the local convergence analysis, it was proved in [35] that gradient methods succeed in deriving benign over-fitted estimates under some separation conditions. Due to the local convergence, starting with Guassian random initialization or MSRA initialization, provable generalization error rates have been successfully deduced in [16], [17], [36], [37], and [38] under some conditions such as the separation, random feature separation and data non-degeneration assumptions,
showing that implementing SGD/GD to solve ERM-OPDNN is able to derive estimates with small sample errors. As discussed above, these optimization-based analysis results are beneficial to demonstrate the effectiveness for optimization algorithms in solving ERM-OPDNN, but lack optimality verifications to show the power of depth. Our research target on Problem 1, in the flavor of approximation-based analysis as shown in Figure 3, focuses on deep learning models rather than optimization algorithms. We place more emphasis on optimality verifications of global minima of ERM-OPDNN. Due to the highly nonconvex nature, these global minima inevitably contain infinite many elements whose generalization performance are totally different.

C. Our Contributions

Our main contributions can be concluded as follows.

• Methodology novelties: Different from the widely adopted optimization-based road-map, we propose an approximation-based road-map that aims to embody the superiority of deep learning, and takes the initial attempt to provide the affirmative answer to Problem 1. Technically, we develop a novel network deepening approach based on the localized approximation property [11], [39] and product-gate property [40], [41] of deep nets. The network deepening approach succeeds in constructing OPDNN (student network) via deepening and widening an arbitrary UPDNN (teacher network) so that the obtained student network exactly interpolates the training data and possesses the same generalization capability as the teacher network.

• Theoretical assessments: Our theoretical analysis starts with negative results to show that not all global minima of ERM-OPDNN possess good generalization performances, even when the data are clean. We then prove that for deep ReLU nets with more than two hidden layers, there always exist perfect global minima of ERM-OPDNN that achieve optimal generalization error rates for both clean and noisy data, provided the number of free parameters achieves a certain level. This finding partly demonstrates the reason of the benign over-fitting phenomenon of deep learning and shows that ERM-OPDNN derives an estimator of high quality for numerous types of data.

• Numerical verifications: We conduct several of numerical experiments to verify theoretical assertions in terms of finding global minima of ERM-OPDNN with zero training error and smaller testing error than shallow nets. Furthermore, we compared the proposed deep learning approach with kernel interpolation, kernel ridge regression and linear regression to show that for low dimensional data, deep learning succeeds in finding perfect global minima of ERM-OPDNN while other approaches fail. Though our numerical results are only given in terms of existence, i.e., choosing a good result for several trials, we believe that this existence verifications can be overcome by selecting suitable initialization points, which is the future focus of our study.

The rest of this paper is organized as follows. In the next section, we mathematically formulate our studied problem. In Section III, we provide theoretical guarantee for the existence of perfect global minima of ERM-OPDNN. In Section IV, we conduct numerical experiments to verify our theoretical assertions. We prove our results in the last section.

II. MATHEMATICAL FORMULATIONS

In this section, we give several definitions concerning OPDNN and mathematically formulate Problem 1. Let \( L \in \mathbb{N} \) be the depth of a deep net, \( d_0 = d \) and \( d_L \in \mathbb{N} \) be the width of the \( \ell \)-th hidden layer for \( \ell = 1, \ldots, L \). Denote by \( f_\ell(x) := W_\ell x + b_\ell \) for \( d_0 \times d_{\ell-1} \) weight matrix \( W_\ell \) and bias vector \( b_\ell \in \mathbb{R}^{d_\ell} \) the affine operator from \( \mathbb{R}^{d_{\ell-1}} \) to \( \mathbb{R}^{d_\ell} \). For the ReLU function \( \sigma(t) := \max\{t, 0\} \), write \( \sigma(x) = (\sigma(x^{(1)}), \ldots, \sigma(x^{(d)})^T \) for \( x = (x^{(1)}, \ldots, x^{(d)})^T \). Then, an \( L \)-layer deep ReLU net is given by

\[
N_{d_0, \ldots, d_L}(x) = a \cdot \sigma \circ f_L \circ \sigma \circ f_{L-1} \circ \cdots \circ \sigma \circ f_1(x),
\]

where \( a \in \mathbb{R}^{d_L} \). The structure of \( N_{d_0, \ldots, d_L} \) is determined by the weight matrices \( W_\ell \) and bias vectors \( b_\ell, \ell = 1, \ldots, L \). In particular, full weight matrices correspond to deep fully connected nets (DFCN) [40]; sparse weight matrices are associated with deep sparsely connected nets (DSCN) [41]; and Toeplitz-type weight matrices are related to deep convolutional neural networks (DCNN) [42].

For DFCN, we have the number of training parameters

\[
n_L = d_L + \sum_{\ell=1}^L (d_{\ell-1}d_\ell + d_\ell). \tag{2}
\]

We mainly focus on analyzing the benign over-fitting phenomenon for DFCN. Denote by \( N_{d_0, \ldots, d_L} \) the set of all DFCNs of the form (1). Define the width of DFCN to be \( d_{\text{max}} := \max\{d, d_1, \ldots, d_L\} \).

We conduct our study in the standard least-square regression formulation [14], in which the inputs \( \Lambda := \{x_i\}_{i=1}^m \) are i.i.d. drawn according to an unknown distribution \( \rho_X \) on an input space \( \mathcal{X} \) and outputs

\[
y_i = f^*(x_i) + \varepsilon_i, \quad i = 1, \ldots, m, \tag{3}
\]

with \( \{\varepsilon_i\}_{i=1}^m \) being independent random variables that are independent of \( \{x_i\}_{i=1}^m \) and satisfy \( |\varepsilon_i| \leq M, \mathbb{E}[\varepsilon_i] = 0, i = 1, \ldots, m \) for some \( M > 0 \). Without loss of generality, we assume \( \mathcal{X} = [0, 1]^d \) throughout the paper. Write \( D = \{(x_i, y_i)\}_{i=1}^m \) as the data set. The aim is to derive an estimate \( f_0 \) based on \( D \) to well approximate \( f^* \).

If the number of free parameters \( n \) in a deep net is more than the size of data \( m \), we then say this deep net to be over-parameterized. In this case, \( n \to \infty \) when \( m \) increases to infinity. We denote \( n \sim m \) if there exists some \( C > 0 \) such that \( \frac{1}{C}m \leq n \leq Cm \) as \( m \to \infty \). The notations \( n \sim m^s \) with \( s > 1 \) and \( n \sim m^s \) with \( u, v > 0 \) are defined in the same way. To describe the extent of over-parameterization, we present the following definition.

Definition 1: Let \( N_{d_0, \ldots, d_L} \) possess \( n \) free parameters satisfying \( n \geq m \). If \( n \sim m \), we say that \( N_{d_0, \ldots, d_L} \) is an equivalently over-parameterized deep nets. If there is some \( s > 1 \) such that \( n \sim m^s \), we say that \( N_{d_0, \ldots, d_L} \) is algebraically over-parameterized deep nets of order \( s \). If there exist some \( u, v > 0 \) such that \( n \sim m^{us} \), we then say that
over-parameterized deep nets of order $(u,v)$. If for any $u,v > 0$ independent of $m,n$, there always holds $n > e^{uv}$ as $m \to \infty$, $\mathcal{N}_{d_1,...,d_l}$ is said to be superexponentially over-parameterized.

We are interested in deriving the estimate via the following empirical risk minimization (ERM), especially on over-parameterized deep nets:

$$f_D := \arg\min_{f \in \mathcal{N}_{d_1,...,d_l}} \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2.$$  

Convergence of gradient descent type algorithms to solve (4) has been verified for exponentially over-parameterized deep nets in [6]. Generalization and convergence of gradient descent algorithms for (4) have been demonstrated in [13] for algebraically over-parameterized deep nets of order $s \geq 10$ under several conditions on data distributions. In this paper, we consider generalization performance of global minima of (4) on algebraically over-parameterized deep nets of order at least 2. Since the optimization problem defined in (4) is nonconvex and the deep nets are over-parameterized, its solutions, i.e. global minima, are not unique (see Lemma 1 below). Denote by $\mathcal{P}_{d_1,...,d_l,m}$ the set of all global minima of (4), i.e.,

$$\mathcal{P}_{d_1,...,d_l,m} := \{f : f \text{ is a solution to (4)}\}.$$  

We aim at deriving optimal generalization error rates of elements in $\mathcal{P}_{d_1,...,d_l,m}$.

Write

$$\mathcal{M}(\Theta, \Xi) := \{(f^*, \rho_X) : f^* \in \Theta, \rho_X \in \Xi\}$$

with $\Theta$ a set of functions defined on $X$ and $\Xi$ a set of Borel probability measures on $X$. Let $\Gamma_D$ be the class of all functions that are derived based on the data set $D$. Define

$$e(\Theta, \Xi) := \sup_{(f^*, \rho_X) \in \mathcal{M}(\Theta, \Xi)} \inf_{f \in \Gamma_D} \mathbb{E}\left[\|f^* - f\|^2_{\mathcal{L}^2}\right].$$  

Since $e(\Theta, \Xi)$ concerns infimum over $\Gamma_D$, it measures theoretically optimal generalization error bounds learning schemes based on $D$ can achieve for $f^* \in \Theta$ and $\rho_X \in \Xi$. In this way, $e(\Theta, \Xi)$ provides a baseline to judge the generalization performance of specific learning algorithms. For any $f_D \in \mathcal{P}_{d_1,...,d_l,m}$, define

$$\mathcal{V}_{\Theta, \Xi}(f_D) := \sup_{\rho_x \in \mathcal{M}(\Theta, \Xi)} \mathbb{E}\left[\|f^* - f_D\|^2_{\mathcal{L}^2}\right].$$

as the generalization error. We then present the following definition of (rate,almost-) optimal estimators by comparing $\mathcal{V}_{\Theta, \Xi}(f_D)$ with $e(\Theta, \Xi)$.

**Definition 2:** Let $\mathcal{M}(\Theta, \Xi)$ be the set of all distributions satisfying $f^* \in \Theta$ and $\rho_X \in \Xi$. If for an arbitrary $\varepsilon > 0$, there holds for sufficiently large $m$,

$$e(\Theta, \Xi) \leq \mathcal{V}_{\Theta, \Xi}(f_D) \leq (1 + \varepsilon)e(\Theta, \Xi),$$  

$f_D$ is said to be an optimal estimator for $\mathcal{M}(\Theta, \Xi)$. Denote the set of all optimal estimators by $\mathcal{Opt}(\Theta, \Xi)$. If

$$\mathcal{V}_{\Theta, \Xi}(f_D) \sim e(\Theta, \Xi),$$  

$f_D$ is said to be a rate-optimal estimator for $\mathcal{M}(\Theta, \Xi)$. Denote by $\mathcal{Ropt}(\Theta, \Xi)$ the set of all rate-optimal estimators. If

$$e(\Theta, \Xi) \leq \mathcal{V}_{\Theta, \Xi}(f_D) \leq \tilde{c}e(\Theta, \Xi)\log^m(m)$$  

for some $\tilde{c}, \varepsilon > 0$, $f_D$ is said to be an almost-optimal estimator for $\mathcal{M}(\Theta, \Xi)$ and we denote by $\mathcal{O}_\mathcal{R}opt(\Theta, \Xi)$ the set of all almost-optimal estimators with coefficient $\tilde{c}$ and exponent $\varepsilon$.

Based on the above definitions, Problem 1 boils down to check elements in $\mathcal{Opt}(\Theta, \Xi) \cap \mathcal{P}_{d_1,...,d_l,m}$, $\mathcal{Ropt}(\Theta, \Xi) \cap \mathcal{P}_{d_1,...,d_l,m}$ or $\mathcal{O}_\mathcal{R}opt(\Theta, \Xi) \cap \mathcal{P}_{d_1,...,d_l,m}$ for some $\tilde{c}, \varepsilon \geq 0$. To guarantee zero training error, some restrictions on the width of DFCN should be given. The following lemma that can be easily derived from [43, Theorem 5.1] presents an example for such a restriction.

**Lemma 1:** If $d_1 \geq m$, then for any $L \geq 1$ and $d_2, \ldots, d_L \geq 2$, there are infinitely many functions $f \in \mathcal{P}_{d_1,...,d_l,m}$ and for any $f \in \mathcal{P}_{d_1,...,d_l,m}$, there holds $f(x_i) = y_i, i = 1, \ldots, m$.  

**Lemma 1** implies that if $d_1 \geq m$, then there are always infinitely many elements in $\mathcal{P}_{d_1,...,d_l,m}$ and any global minimum $f$ exactly interpolates the given data. We then mathematically formulate Problem 1 as follows:

**Problem 4:** Let $L, d_1, \ldots, d_L \in \mathbb{N}$. Given $\mathcal{M}(\Theta, \Xi)$, if $d_1 \geq m$ and $d_L \geq 2$ for $2 \leq l \leq L$, does

$$\mathcal{Opt}(\Theta, \Xi) \cap \mathcal{P}_{d_1,...,d_l,m} \neq \emptyset$$  

always hold for any $\Theta$ and $\Xi$?

Problem 4 concerns whether benign over-fitting always happens for OPDNN. An affirmative answer to Problem 4 plays a crucial role in understanding advantages of OPDNN. Unfortunately, the answer is frequently negative, even for the well adopted smooth functions [40], [41], [44]. Let $c_0 > 0$ and $r = s + \mu$ with $s \in \mathbb{N}_0 := \{0\} \cup \mathbb{N}$ and $0 < \mu \leq 1$. We say a function $f : \mathcal{A} \subseteq \mathbb{R}^d \to \mathbb{R}$ is $(r,c_0)$-smooth if $f$ is $s$-times differentiable and for every $a_j \in \mathbb{N}_0$, $j = 1, \ldots, d$ with $a_1 + \cdots + a_d = s$, its $s$-th partial derivative satisfies the Lipschitz condition

$$\left|\frac{\partial^s f}{\partial x_1^{a_1} \cdots \partial x_d^{a_d}}(x) - \frac{\partial^s f}{\partial x_1^{a_1} \cdots \partial x_d^{a_d}}(x')\right| \leq c_0\|x - x'\|_2^\mu,$$

for any $x, x' \in \mathcal{A}$.  

**Denote by $\text{Lip}_{\mathcal{A}}^{(r,c_0)}$ the set of all $(r,c_0)$-smooth functions defined on $\mathcal{A}$. The following lemma can be easily derived from [45, Theorem 2].**

**Lemma 2:** If $d_1 \geq m$, then for any $L \geq 1$ and $d_2, \ldots, d_L \geq 2$ and any $h \in \mathcal{P}_{d_1,...,d_l,m}$, there exists a probability measure $\rho^*_X$ on $X$ such that

$$\sup_{f^* \in \text{Lip}_{\mathcal{A}}^{(r,c_0)}} \mathbb{E}\left[\|f^* - h\|^2_{\mathcal{L}^2}\right] \geq 1/6.$$  

**Lemma 2** not only implies that $\mathcal{Opt}(\Theta, \Xi) \cap \mathcal{P}_{d_1,...,d_l,m} = \emptyset$, but also shows that all global minima of ERM are not universally consistent, if we do not impose any restrictions on $\Theta$ or $\Xi$. It demonstrates the necessity of imposing non-trivial assumptions on data as [16], [36], [37], and [38] did in deriving generalization error rates for OPDNN-based algorithms and more importantly, leads to the following problem:

$^1$The conclusion is made on the function level, i.e., $\sigma(2t)$ and $2\sigma(t)$, are regarded as the same function.
Problem 5: Let \( L, d_1, \ldots, d_L \in \mathbb{N} \). Under which conditions on \( M(\Theta, \Xi) \) and \( d_1, \ldots, d_L \) satisfying \( d_l \geq m \) and \( d_l \geq 2 \) for \( 2 \leq l \leq L \) there holds
\[
\text{Opt}(\Theta, \Xi) \cap \Psi_{d_1, \ldots, d_L, m} \neq \emptyset.
\]
(13)

To relax the difficulty, \( \text{Opt}(\Theta, \Xi) \cap \Psi_{d_1, \ldots, d_L, m} \) in Problem 5 can be replaced by \( \text{R Opt}(\Theta, \Xi) \cap \Psi_{d_1, \ldots, d_L, m} \) or \( \text{Opt}^\gamma(\Theta, \Xi) \cap \Psi_{d_1, \ldots, d_L, m} \) for some \( \varepsilon, \gamma > 0 \). It should be highlighted that Problem 5 has not been considered in the literature. On one hand, interesting work in [16], [36], [37], and [38] did not take (rate, almost-) optimal estimator for \( M \).

A. OPDNN Versus UPDNN

In this subsection, we aim to demonstrate the outperformance of OPDNN over UPDNN. Proposition 1 provides extremely bad examples for global minima of (4) in generalization, even for noise-free data. It seems that OPDNNs are always worse than UPDNNs, which is standard from the viewpoint of the bias-variance trade-off principle [14]. However, in our following theorem, we will show that if the number of free parameters increases to a certain extent, then there are also infinitely many global minima of (4) possessing excellent generalization performance. Throughout the paper, the notation \( a \geq b \) for \( a, b > 0 \) means that there is some constant \( C \) depending only on \( r, d \) and \( c_0 \) such that \( a \geq Cb \).

Theorem 6: Let \( r, c_0 > 0 \) and \( N \in \mathbb{N} \). If \( f^* \in \text{Lip}_{r,0}(\mathbb{R}^d) \) satisfies (14), \( N \geq q_\Lambda^{-d} \) holds almost surely, \( L \geq \log N, d_1 \geq N \) and \( d_L \geq \log N \) for \( 2 \leq L \leq L \), then there are infinitely many \( h^* \in \Psi_{d_1, \ldots, d_L, m} \), such that
\[
\|h^* - f^*\|_{L^p(\mathbb{R}^d)} \leq CN^{-r/d},
\]
(16)
where \( C > 0 \) is a constant depending only on \( r, d \) and \( c_0 \), and
\[
q_\Lambda = \frac{1}{2} \min_{1 \leq i < j} \|x_i - x_j\|_2
\]
(17)
denotes the separation radius of \( \Lambda \).

A consensus on deep nets approximation is that it can break the "curse of dimensionality", which was verified in the interesting work [46], [47] in terms of deriving dimension-independent approximation rates. However, it should be pointed out that to achieve such dimension-independent approximation rates, strict restrictions have to be imposed on target functions, which become stronger as the dimension \( d \) grows, just as [48, P68] observed. In this way, though the approximation error of deep nets is independent of the dimension, the applicable target functions become more and more stringent as \( d \) grows. Our result presented in Theorem 6 drives a different direction to show that for well-studied smooth functions, there is a deep net that interpolates the training data without degrading the approximation rate. We highlight that the approximation rate depends on the a-priori knowledge of the target functions. In particular, if we impose strict restrictions such as \( f^* \in \text{Lip}_{r,0}(\mathbb{R}^d) \) with \( r \geq d/2 \), which are standard in kernel learning [49], then the approximation rate can be at least of order \( n^{-1/2} \), which is also dimension-independent.

Noting that there are totally \( O(N \log N) \) parameters in \( h^* \) in (16), the derived error rate is almost optimal in the sense that up to a logarithmic factor, the derived upper bound is of the same order of lower bound [40], [50], i.e., \( C_1(N \log N)^{-r/d} \) for some \( C_1 > 0 \) independent of \( N \). This means that for \( N \geq q_\Lambda^{-d} \) and \( L \geq \log N \), we can get almost optimal deep nets via finding suitable global minima of (4). Theorem 6 also shows that in the over-parameterized setting, where all global minima exactly interpolate the data, the interpolation restriction does not always affect the approximation performance of deep nets. Theorem 6 actually presents a sufficient condition for the number of free parameters of deep ReLU nets, \( N \geq q_\Lambda^{-d} \), to guarantee the existence of perfect global minima when we are faced with noise-free data.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
It should be highlighted that $q_\Lambda$ is half of the smallest distance between any two distinct points in $\Lambda$ and naturally satisfies $q_\Lambda \leq m^{-1/d}$. It can be numerically determined, provided the data set is given. If $\Lambda$ is distributed quasi-uniformly [51], i.e., $q_\Lambda \sim m^{-1/d}$, the studied deep nets is equivalent over-parameterized. For general $\Lambda$, the restriction $N \geq q_\Lambda^d$ results in algebraically and even exponentially OPDNN. If $\{x_i\}_{i=1}^m$ are i.i.d. drawn according to some distribution, the lower bound of $q_\Lambda$ is easy to be derived in theory [14], [28], [32]. In particular, for $M'$, $\sigma > 0$, write $\Lambda_{M',\sigma}$ as the set of mean zero random vectors with variance $\sigma^2$ and uniform bound $M'$. If $\Lambda = \{x_i\}_{i=1}^m$ is a set of random vectors i.i.d. drawn from $\Xi_{M',\sigma}$, then it was proved in [28] and [32] that $q_\Lambda \geq \frac{c \sigma}{\sqrt{2M'}}$, holds with confidence at least $1-2m^2 \exp\left(\frac{-\sigma^2}{4M'}\right)$, implying the deep nets to be algebraically over-parameterized with high probability. Therefore, for i.i.d. random inputs and high dimensional data, $N \geq q_\Lambda$ is not so strict.

In the following, we present a corollary to demonstrate the power of over-parameterization. According to Lemma 2, there exists a distribution $p_X$ such that all global minima of ERM-OPDNN perform extremely bad and the problem of Lemma 2 in [45] shows that $p_{\Lambda}$ is a special distribution that is even not continuous with respect to the Lebesgue measure. Therefore, it is reasonable and necessary to impose mild conditions of $\rho_X$ to exclude such distributions to guarantee the good performance of elements in $\Psi_{d_1,\ldots,d_m}$. The restriction we adopted in this paper is the following well known distortion assumption of $\rho_X$ [11], [52]. Let $p \geq 2$ and $J_p$ be the identity mapping $L^p(X) \rightarrow L^2_{\rho_X}$. Define $D_{p\rho_X} := \|J_p\|$, where $\| \cdot \|$ denotes the operator norm. Then $D_{p\rho_X}$ is called the distortion of $\rho_X$ (with respect to the Lebesgue measure), which measures how much $\rho_X$ distorts the Lebesgue measure. In our analysis, we assume $D_{p\rho_X} < \infty$, which holds for the uniform distribution for all $p \geq 2$ obviously. Denote by $\Xi_p$ the set of all $\rho_X$ satisfying $D_{p\rho_X} < \infty$. Combining all above assertions, we can derive the following corollary directly.

**Corollary 1**: Let $p \geq 2$, $r, c_0 > 0$ and $N \in \mathbb{N}$. If $N \geq q_N^d$ holds almost surely, $L \geq \log N$, $d_\ell \geq N$ and $d_\ell \geq \log N$ for $\ell = 2, \ldots, L$, then for any there exist infinitely many $h \in \Psi_{d_1,\ldots,d_m}$ such that

$$\frac{(N/\log^3 N)^{r/d}}{\mathcal{N}_{Lip^{(c_0)}_{\rho}}(h)} \leq CD_{p\rho_X} N^{r/d},$$

(18)

where $\bar{c}, \nu$ are constants depending only on $r, d$.

Noticing that there are totally $O(N \log^2 N)$ free parameters in the deep nets in Corollary 1, we can derive the lower bound of (18) directly via [50]. The upper bound can be deduced from Theorem 6 and the definition of $D_{p\rho_X}$ directly. Corollary 1 shows that for the noise-free setting, the generalization performance of OPDNN depends only on the size of network, provided $N \geq q_N^d$ holds almost surely. This actually implies that OPDNN performs much better than UPDNN in the same setting, showing the power of over-parameterization and presenting the reason why OPDNN performs well in some real-world applications. It should be mentioned that similar results also hold for sufficiently small noise by using the same approach in this paper.

---

**TABLE I**

| Classes | Refer. | Rate | Width | Depth |
|---------|--------|------|-------|-------|
| $Lip^{(c_0)}_{\rho}$ | [10] | $m^{-\delta}$ | $m^{-\delta}$ | $\log m$ |
| $\Psi_{\varepsilon,\kappa,\Delta}$ | [9] | $m^{-\delta}$ | $m^{-\delta}$ | $\log m$ |
| $Lip^{(c_0)}_{\rho}$ | [11] | $m^{-\delta}$ | $m^{-\delta}$ | $\log m$ |

**B. Optimality of ERM-OPDNN**

In the previous subsection, we analyzed the generalization performance of ERM-OPDNN for tackling noise-free data and demonstrated advantages of OPDNN over UPDNN. This subsection focuses on handling noisy data and presents the answer to Problem 5. At first, we introduce the following definition of UPDNN-achievable class.

**Definition 3**: Let $L^*, n^*, d_1^*, \ldots, d_L^* \in \mathbb{N}$ satisfying (2) and $n^* < m$. A class of functions $\Theta \subseteq \mathcal{L}_{p_0}$ is said to be $(L^*, n^*)$-UPDNN-achievable, $(L^*, n^*)$-rate UPDNN-achievable and $(L^*, n^*)$-almost UPDNN-achievable, if $\mathcal{O}(p_0, \Xi) \cap \Psi_{d_1^*,\ldots,d_L^*} \neq \emptyset$, $\mathcal{R}_\mathcal{O}(p_0, \Xi) \cap \Psi_{d_1^*,\ldots,d_L^*} \neq \emptyset$ and $\mathcal{O}(p_0, \Xi) \cap \Psi_{d_1^*,\ldots,d_L^*} \neq \emptyset$ respectively, where $\xi$ and $\nu$ are constants independent of $m, n^*, d_1^*, \ldots, d_L^*$.

UPDNN-achievable classes have been well studied in [9], [10], and [11]. Besides the smoothness class, we introduce other two popular classes. The first one is the generalized additive models that are widely used in statistics and machine learning [9], [14]. For $r, x_0, c_0 > 0$, we say that $f$ admits a generalized additive model if $h \in L^p_{\rho_X}$ and $f \in L_{\rho_X}$. Write $\Psi_{d_1^*,\ldots,d_L^*}$ as the set of all functions admitting a generalized additive model. The other is the spatially sparse functions which has been shown in [11], [53], and [54] to be an important data feature for image and signal processing. Partition $\mathbb{N}$ by $(N^\ell)^d$ sub-cubes $\{A_j\}_{j=1}^{(N^\ell)^d}$ of side length $(N^\ell)^{-1}$ and with centers $\{\zeta_j\}_{j=1}^{(N^\ell)^d}$. For $u \in \mathbb{N}$ with $u \leq (N^\ell)^d$, define

$$\Sigma_u := \{j \in \{1, \ldots, (N^\ell)^d\}, \ell \leq u \leq u + N^\ell\}.$$

If the support of $f \in L^p_{\rho_X}$ is contained in $S := \bigcup_{j \in \Sigma_u} A_j$ for a subset $\Sigma_u$ of $\{1, \ldots, (N^\ell)^d\}$ of cardinality at most $u$, then we say that $f$ is $u$-sparse in $(N^\ell)^d$ partitions. Denote by $\mathcal{L}^{(c_0)}_{\rho}$ the set of all $f \in L^p_{\rho_X}$ which are $\ell$-sparse in $(N^\ell)^d$ partitions. Table I provides details of three almost UPDNN-achievable classes for some $\xi, \nu$ independent of $m, d_1, \ldots, d_L$.

Based on these helps, we present our next main result to show the existence of perfect global minima of ERM-OPDNN.

**Theorem 7**: Let $p \in [2, \infty)$, $L^*, n^* \in \mathbb{N}$. If $\Theta$ is $(L^*, n^*)$-rate UPDNN-achievable or $(L^*, n^*)$-almost UPDNN-achievable, then for any $L \geq C'(L^* + \log m)$ and $d_\ell \geq C''m$, $\ell = 1, \ldots, m$, there holds

$$|\mathcal{R}_\mathcal{O}(p_0, \Xi) \cap \Psi_{d_1^*,\ldots,d_L^*}| = \infty,$$

(19)

respectively, where $|A|$ denotes the cardinality of the set $A$, $C'$, $C''$, $\xi$ and $\nu$ are constants independent of $m, L$, $L^*, n^*$, $d_1, \ldots, d_L$.

---

2Though [10], [11] focused on DSCN, their results can be derived for DFCN easily by deepening the network.

---

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
The proof of Theorem 7 depends on a novel deepening scheme developed in Theorem 8 below. In Theorem 7, we are only interested in rate-optimal and almost optimal UPDNN-achievable classes, due to the fact that even for UPDNN whether there exists an optimal deep learning algorithm remains open. In fact, according to Definition 2, optimal algorithms not only require optimal generalization error rates, but also need smallest constants, beyond the shape of existing statistical tools like concentration inequalities [55] and localization schemes [14]. Moreover, (19) shows that for rate UPDNN-achievable (or almost UPDNN-achievable) classes, there are infinitely many elements in $\mathcal{R}opt(\Theta, \Xi)$ or $\mathcal{O}pt(\Theta, \Xi)$, showing that ERM-UPDNN succeeds in deriving rate- (or almost)-optimal estimates.

From UPDNN to OPDNN, a direct consequence of Lemma 1 is that the training error of elements in $\Psi_{d_1, \ldots, d_l, m}$ achieves zero, i.e., for any $h \in \Psi_{d_1, \ldots, d_l, m}$, there holds

$$h(x_i) = y_i, \quad \forall i = 1, \ldots, m.$$  (20)

This together with Theorem 7 shows that there exists benign over-fitted estimates under mild conditions on the distributions, provided the deep nets are algebraically over-parameterized of degree at least 2. Combining Theorem 7 with Table I, it is easy to derive the following corollary directly.

**Corollary 2**: Let $p \geq 2$, and $\gamma, c_0, c'_0 > 0$, $N \in \mathbb{N}$, $u \leq N^d$ and $m$ satisfy $m^{\frac{\gamma}{\log m}} \geq \frac{\gamma}{N^d(u \log^2 u)^\frac{1}{2}}$. If $L \geq \log m$, $d_1 \geq m$ for any $\ell = 1, \ldots, m$, then there are $h_1, h_2, h_3 \in \Psi_{d_1, \ldots, d_l, m}$ such that

$$C_1 m^{\frac{2}{\gamma}} \leq \sup_{f \in L^{\gamma}, f \in \Xi} \mathbb{E}[\|h_1 - f\|^2_{L^{\gamma}}] \leq C_2 \left(\frac{m}{\log m}\right)^{\frac{2}{\gamma}},$$

$$C_3 \left(m^{\frac{2}{\gamma}} + m^{-\frac{2\gamma}\log^2 m}\right) \leq \sup_{f \in W_{\gamma}, f \in \Xi} \mathbb{E}[\|h_2 - f\|^2_{L^{\gamma}}],$$

$$\leq C_4 \left(m^{\frac{2}{\gamma}} + m^{-\frac{2\gamma}\log^2 m}\right) \log^3 m,$$  (21)

and

$$C_5 m^{\frac{2}{\gamma}} \left(\frac{u}{N^{d'}}\right)^\frac{2}{\gamma} \leq \sup_{p \in M(L^{d'}, \Xi)} \mathbb{E}[\|h_3 - f\|^2_{L^{\gamma}}],$$

$$\leq C_6 \left(\frac{m}{\log m}\right)^{\frac{2}{\gamma}} \left(\frac{u}{N^{d'}}\right)^\frac{2}{\gamma},$$  (22)

where $C_1, \ldots, C_6$ are constants independent of $m, L, d_1, \ldots, d_l$.

Besides the power of over-parameterization in approximating noise-free functions exhibited in Theorem 6, Corollary 2 presents another advantage of OPDNN in terms of its versatility, i.e., adapting to different learning tasks with a unified network structure. Theorem 7 and Corollary 2 theoretically answer Problem 5 to demonstrate the selection of non-trivial $\Theta$ and $\Xi$ so that (13) holds, and therefore succeed in providing the affirmative to Problem 1.

A quite related work for the flavor of Theorem 7 and Corollary 2 is given in [56], where the generalization performance of interpolation methods based on histograms was discussed. The main arguments of [56] and our result are similar: there are global minima of ERM-OPDNN that can avoid overfitting. The main differences are as follows: 1) It is well known that an approximator or learning based on histograms suffers from the well known saturation problem in the sense that the approximation or learning rate cannot be improved further once the regularity (or smoothness) of the regression function achieves certain level [57]. Our results avoid this saturation phenomenon via deepening the networks. 2) We provide detailed structures of deep ReLU nets and derive the quantitative requirement of the number of free parameters to guarantee the existence of global minima of ERM-OPDNN, which is different from [56]. 3) More importantly, we devote to answering Problem 1 via showing the optimal generalization error rates and the power of depth of some global minima of ERM-OPDNN.

### C. Discussions

In this subsection, we present discussions on our results.

- **UPDNN versus OPDNN**: According to the bias and variance trade-off principle, UPDNN succeeds in embodying the power of depth in numerous learning tasks [9], [10], [11]. The main advantage of UPDNNs is the uniform generalization error estimates in terms that all global minima of ERM-UPDNN possess perfect generalization performance. Two main drawbacks, however, are the a-priori dependent network structures and absence of provable optimization algorithms. Due to the power of width in optimization [15], ERM-OPDNNs have clearer landscapes in the sense that there are not local minima in solving (4) provided the width of the last hidden layer is more than $m$. Our result in Theorem 7 as well as its proof shows that there are also provable schemes to find the global minima of ERM-OPDNN with perfect generalization performances. Furthermore, there is a unified OPDNN, running ERM on which is capable of finding estimates possessing perfect generalization performance for different learning tasks. All these show advantages of OPDNN over UPDNN. However, it should be highlighted that the generalization error analysis for ERM-OPDNN presented in Theorem 6 and Theorem 7 are not unified, i.e., there are also global minima of ERM-OPDNN performing extremely badly. We conclude the comparisons of UPDNN and OPDNN in Table I.

- **Optimization-based analysis versus approximation-based analysis**: As shown in Figures 2 and 3, the purpose of both road-maps is to derive optimal generalization error rates for deep learning algorithms to make clear of the running mechanisms of deep learning and then present the theoretical insights on why and when deep learning performs so well in practice. However, existing results on both road-maps are not complete. Indeed, studies [16], [38] following the optimization-based road-map benefit in algorithm convergence but require strict assumptions on data and high-extent over-parameterization.

---

**Table II**

| OPDNN | UPDNN |
|-------|-------|
| Optimal Generalization | ✓ | ✓ |
| Solvability | ✓ | ✗ |
| Unified structure | ✓ | ✗ |
| Unified theory | ✗ | ✓ |
As a result, it is quite difficult to judge whether such a road-map is capable of presenting theoretical verifications of perfect generalization performances of deep learning. This is the main reason why we focus on the approximation-based road-map, though the study is at the infancy. As discussed above, Theorem 6 and Theorem 7 demonstrate the existence of perfect global minima of ERM-OPDNN in generalization, showing that the approximation-based road-map provides a theoretically possible way to reveal the veil of deep learning. However, the task is still arduous since our result only implies the existence and is algorithm-independent. Table III presents a detailed comparison between these two road-maps.

| Approximation-based | Optimization-based |
|---------------------|-------------------|
| Key points          | Local convergence |
| Challenges          | Initialization point selection |
| Assumptions         | Weak conditions on functions |
| Pros                | Optimal generalization guarantee |
| Cons                | Algorithm independent |

- Local convergence and initialization selection: Though the existence issue of approximation-based road-map in Figure 3 has been verified in Theorem 7, it still remains open whether there are implementable optimization algorithms to achieve these optimal generalization err rates. The key point, exhibited in Table III, is the local convergence of algorithms. Different from [6] and [13] that derived local convergence of gradient-type algorithms around specific initialization points, the requirements for local convergence in the approximation-based road-map must be stricter since it is unclear whether small neighbourhoods of these initialization points contain the perfect global minima presented in Theorem 7. Fortunately, the local convergence of algorithms for deep nets have triggered enormous research activities in recent years, two typical results of which can be found in [22] and [24]. In particular, [22] proposed a block coordinate descent (BCD) algorithm for both UPDNN and OPDNN and provided its local convergence, while [24] showed that SGD possessed local convergence under a local Lojasiewicz condition. It can be found in [24] that the local Lojasiewicz condition actually presents some restrictions on the initialization points. These pioneer results provide guidance on designing efficient iterative algorithms with local convergence guarantee. However, it is unclear whether the local convergence guarantee is achievable for initialization points around the perfect global minima established in Theorem 7. The answer is an important consideration for practical applications for OPDNN and depends not only on the algorithms themselves, but also the selection of initialization points. We will continue our study of the local convergence of some efficient iterative algorithms provided when initialization point schemes are provided.

The selection of initialization points is crucial for the approximation-based road-map, just as Table III purports to show. Several initialization approaches have been posed to guarantee the global (or local) convergence of gradient-type algorithms, but it is quite difficult to judge whether the corresponding initialization points lie near a perfect global minima of ERM-OPDNN. Under this circumstance, a preferable method is warm-starting, i.e., using some initialization points trained by data. The random sketching scheme proposed in [58] and constructive scheme developed in [54] may be good choices.

IV. NUMERICAL EXAMPLES

In this section, we conduct numerical simulations to support our theoretical assertions on the existence of benign overfitting of running ERM on over-parameterized deep ReLU nets. There are mainly four purposes of our simulations. In the first simulation, we aim to show the relation between the generalization performance of global minima of (4) and the number of parameters (or width) of deep ReLU nets. In the second one, we devote to verifying the over-fitting resistance of (4) via showing the relation between the generalization error and the number of algorithmic iterations (epochs). In the third one, we show the existence of good and bad global minima of (4). Finally, we compare our learned global minima with some widely used learning schemes to show the learning performance of (4) on over-parameterized deep ReLU nets.

For these purposes, we adopt fully connected ReLU neural networks with \( L \) hidden layers and \( k \) neurons paved on each layer. In all simulations, we set \( L \in \{1,2,4\} \) and \( k \in \{1,\ldots,2000\} \). We use the well known Adam optimization algorithm on deep ReLU nets with step-size being constantly 0.001 and initialization being the default PyTorch values. Without especial declaration, the training is stopped after 50,000 iterations.

We report our results on numerical simulations on two real-world datasets. The first one is a Wine Quality dataset from UCI database. The Wine Quality dataset is related to red and white variants of the Portuguese “Vinho Verde” wine with 1599 red and 4898 white examples. We select white wine for experiments. There are 12 attributes in the data set: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality (score between 0 and 10). Therefore, it can be viewed as a regression task on input data of 11 dimension. Regarding the preferences, each sample was evaluated by a minimum of three sensory assessors (using masked tastes), which graded the wine in a scale ranging from 0 (very bad) to 10 (excellent). We sample 2/3 data points as our training set and 1/3 for testing.

The second dataset is MNIST. MNIST dataset is widely used in classification tasks. Here we follow [28] to create a regression task using MNIST. MNIST inculdes 70000 samples in total. Each sample includes a 28*28 dimensional feature and a target representing a digit ranging from 0 to 9. We randomly pick 291 samples with targeting digits equaling to 0 or 1, and separate 221 samples for training and 70 for testing. We label digit 0 as -1 and digit 1 as 1. Then we get a dataset with a 28 \( \times 28 = 784 \) dimensional feature and a label -1 or 1. The regression task is then built.

Simulation 1: In this simulation, we study the relation between the RMSE (rooted mean squared error) of test error and widths of deep ReLU nets with \( L = 1,2,4 \). Our results are recorded via 5 independent single trials. The solid line is the mean value from these trials, and the shaded part indicates the deviation. The numerical results are reported in Figure 4.
Figure 4 presents the perfect global minima by exhibiting the relation between training and testing RMSE and the number of free parameters. From Figure 4, we obtain the following four observations: 1) The left figure shows that neural networks with more hidden layers are easier to produce exact interpolations of the training data. This coincides with the common consensus since more hidden layers with the same width involves much more free parameters; 2) For each depth, it can be found in the right figure that the testing curves exhibit an approximate doubling descent phenomenon as declared in [25] for linear models. It should be highlighted that such a phenomenon does not always exist for deep ReLU nets training and we only choose a good trend from several trails to demonstrate the existence of perfect global minima; 3) As the width (or capacity of the hypothesis space) increases, it can be found in the right figure that the testing error does not increase, exhibiting a totally different phenomenon from the classical bias-variance trade-off. This shows that for over-parameterized deep ReLU nets, there exist good global minima of (4), provided the depth is appropriately selected; 4) It can be found that deeper ReLU nets perform better in generalization, which demonstrates the power of depth in tackling the Wine Quality data. All these verify our theoretical assertions in Section III and show that there exist perfect global minima of (4) to realize the power of deep ReLU nets.

**Simulation 2:** In this simulation, we devote to numerically studying the role of iterations (epochs) in (4) in both under-parameterized and over-parameterized settings. We run ERM on DFCNs with depth 4 and width 2, 40, 2000 on Wine dataset and MNIST dataset. Since the number of training data is 3265(221 in MNIST dataset, resp.), it is easy to check that deep ReLU nets with depth 4 and widths 2 and 40 are under-parameterized, while those with depth 4 and width 2000 is over-parameterized. The numerical results are reported in Figure 5.

There are also three interesting findings exhibited in Figure 6: 1) For under-parameterized ReLU nets, it is almost impossible to produce a global minimum acting as an exact interpolation of the data. However, for over-parameterized deep ReLU nets, running Adam with sufficiently many epoches attains a training error to be zero. Furthermore, after a specific value, the number of iterations does not affect the training error. This means that Adam converges to a global minimum of (4) on over-parameterized deep ReLU nets; 2) The testing error for under-parameterized ReLU nets, exhibited in the right figure, behaves according to the classical bias-variance trade-off principle in the sense that the error firstly decreases with respect to the epoch and then increases after a specific value of epoches. Therefore, early-stopping is necessary to guarantee the good performance in this setting; 3) The testing error for over-parameterized ReLU nets is always non-increasing with respect to the epoch. This shows the over-fitting resistance of deep ReLU nets training and also verifies the existence of the perfect global minima of (4) on
over-parameterized deep ReLU nets. It should be highlighted the numerical result presented in Figure 6 is also a single trial selected from numerous results, since we are concerned with the existence of perfect global minima. In fact, there are also numerous examples for bad global minima of (4).

Simulation 3: In this simulation, we show that although there exist perfect global minima in over-parameterized settings, bad global minima can also be found sometimes. We test the performance of deep ReLU nets with depth 4 and width 2000 on Wine dataset and MNIST dataset. It take different
numbers of steps to converge to a good training performance on these two datasets. We trigger several runs with different learning rates and net parameter initializations, and pick good and bad global minima from two trials respectively. We report the numerical results in Figure 6.

From Figure 6, we find that different global minima of Figure 6 perform totally differently in generalization, though the training loss both comes to 0. In particular, the testing errors of bad global minima can be much larger than those of good global minima. It should be mentioned that the bad interpolants in the above simulations are also derived from Adam. Therefore, their orders of testing errors are comparable with those of good interpolants. We highlight that this is due to the implementation of the ADAM algorithm rather than the model (4). As far as the model is concerned, it can be shown in our next simulation that the orders of bad interpolants are also larger than those of good ones.

**Simulation 4:** In this simulation, we compare (4) on over-parameterized deep ReLU nets with some standard learning algorithms including ridge regression (Ridge), support vector regression (SVR), kernel interpolation (KIR) and kernel ridge regression (KRR) to show that the numerical phenomenon exhibited in previous figures is not built upon sacrificing the generalization performance. For the sake of fairness, we test various models under the same condition of our best efforts via tuning hyper-parameters. In particular, we implement the referenced methods by using the standard scikit-learn package. In the experiment with the wine dataset specifically, we use ridge regression with regularization parameter being 1. In KRR, we use the Gaussian kernel with width being 20 and regularization parameter being 0.0002. KIR uses the Gaussian kernel with width being 5 and regularization parameter being 0. SVR keeps the default sklearn parameter being. 0.0002. All these findings verify our theoretical assertions that there are extremely bad global minima of (4). As far as the model is concerned, it can be shown in our next simulation that the orders of bad interpolants are also larger than those of good ones.

There are four interesting observation in Table IV: 1) Learning schemes such as SVR, KRR and Ridge perform stably, since for both high-dimensional applications and low-dimensional simulations. The main reason is that a regularization term is introduced to balance the bias and variance for these schemes. As a result, the training error of these schemes is always non-zero; 2) Kernel interpolation performs well in high dimensional applications but fails to generalize well in low dimensional simulations. The main reason is that if d is large, then the separation radius qA is large [28], [32], which in turn implies that the condition number of the kernel matrix is relatively small, making the kernel interpolation perform well. However, if d is small, the condition number of the kernel matrix is usually extremely large, making the prediction instable; 3) There exist deep ReLU nets exactly interpolating the training data, leading to zero training error, but possessing an excellent generalization capability in yield small testing error, implying that the obtained estimator is a benign over-fitter for the data. Furthermore, it is shown in the table that the testing error of over-parameterized deep ReLU nets is the smallest, demonstrating the power of depth as declared in our theoretical assertions in Section III; 4) There also exist deep ReLU nets interpolating the data but performing extremely badly in generalization, for both high-dimensional applications and low dimensional simulations. All these findings verify our theoretical assertions that there are good global minima for ERM on over-parameterized deep ReLU nets but not all global minima are good.

| **TABLE IV.** COMPARISON WITH OTHER REGRESSION METHODS |
|------------------|------------------|------------------|
| **Wine Quality Data** |     |     |
| Methods          | Train RMSE | Test RMSE |
| Ridge            | 0.534      | 0.735      |
| kernel interpolation | 0.000     | 13.031     |
| KRR              | 0.648      | 0.706      |
| SVR              | 0.628      | 0.696      |
| 4-hidden layer DFCN (good case) | 0.000 | 0.668       |
| CN (bad case)    | 0.000      | 5.931      |
| **MNIST Data**   |     |     |
| Methods          | Train RMSE | Test RMSE |
| Ridge            | 0.056      | 0.304      |
| kernel interpolation | 0.000     | 0.135      |
| KRR              | 0.031      | 0.140      |
| SVR              | 0.073      | 0.154      |
| 4-hidden layer DFCN (good case) | 0.000 | 0.097       |
| CN (bad case)    | 0.000      | 1.000      |

V. PROOFS

In this section, we aim at proving our results stated in Section III. The main novelty of our proof is a deepening scheme that produces an over-parameterized deep ReLU net (student network) based on a specific under-parameterized one (teacher network) so that the student network exactly interpolates the training data and possesses almost the same generalization performance as the teacher network.

### A. Deepening Scheme for ReLU Nets

Given a teacher network g, the deepening scheme devotes to deepening and widening it to produce a student network f
that exactly interpolates the given data \( D \) and possesses the same generalization performance as \( g \). The following theorem presents the deepening scheme in our analysis.

**Theorem 8:** Let \( g_{n,L,U} \) be any deep ReLU nets with \( L \) layers, \( n \) free parameters and width not larger than \( U \in \mathbb{N} \) satisfying \( \|g_{n,L,U}\|_{\mathcal{L}^p(\mathbb{R}^d)} \leq C^* \) for some \( C^* > 0 \). If \( \rho_N \in \Xi_p \) with \( p \in [2, \infty) \), then for any \( \varepsilon > 0 \), there exist infinitely many DFCNs \( f_{D,n,L,U,\varepsilon} \) of depth \( O(L + \log \varepsilon^{-1}) \) and width \( O(m + U + \log \varepsilon^{-1}) \) such that

\[
\|f_{D,n,L,U,\varepsilon}(x_i) - y_i\| \leq \varepsilon, \quad \forall i = 1, \ldots, m, \quad (24)
\]

and

\[
\|f_{D,n,L,U,\varepsilon} - g_{n,L,U}\|_{\mathcal{L}^p(\mathbb{R}^d)} \leq \varepsilon, \quad (25)
\]

where \( \tilde{C} \) is a constant depending only on \( d \).

The deepening scheme developed in Theorem 8 implies that all deep ReLU nets that have been verified to possess good generalization performances in the under-parameterized setting [9], [10], [32] can be deepened to corresponding deep ReLU nets in the over-parameterized setting such that the deepened networks exactly interpolate the given data and possess good generalization error rates.

The main tools for the proof of Theorem 8 are the localized approximation property of deep ReLU nets developed in [11] and the product gate property of deep ReLU nets proved in [40]. Let us introduce the first tool as follows. For \( a, b \in \mathbb{R} \) with \( a < b \), define a trapezoid-shaped function \( T_{\tau,a,b}(t) \) with a parameter \( 0 < \tau \leq 1 \) as

\[
T_{\tau,a,b}(t) := \frac{1}{\tau} \left\{ (b-a+\tau) - (b-a) - (b-a) + (b-a-\tau) \right\}. \quad (26)
\]

We consider

\[
N_{a,b,\tau}(x) := \sigma \left( \sum_{j=1}^{d} T_{\tau,a,b}(x^j) - (d-1) \right). \quad (27)
\]

The following lemma proved in [11] presents the localized approximation property of \( N_{a,b,\tau} \).

**Lemma 3:** Let \( a < b, 0 < \tau \leq 1 \) and \( N_{a,b,\tau} \) be defined by (27). Then we have \( 0 \leq N_{a,b,\tau}(x) \leq 1 \) for all \( x \in \mathbb{R}^d \) and

\[
N_{a,b,\tau}(x) = \begin{cases} 0, & \text{if } x \notin [a, b + \tau]^d, \\ 1, & \text{if } x \in [a, b]^d. \end{cases} \quad (28)
\]

The second tool, as shown in the following lemma, presents the product-gate property of deep ReLU nets [40].

**Lemma 4:** For any \( \ell \in \{2, 3, \ldots, \} \) and \( \nu \in (0, 1) \), there exists a DFCN with ReLU activation functions \( \tilde{x}_{\ell,\nu} : \mathbb{R}^\ell \to \mathbb{R} \) with \( O\left( \ell \log \frac{1}{\nu} \right) \) depth, \( O\left( \ell \log \frac{1}{\nu} \right) \) width, and free parameters bounded by \( O\left( \ell^p \nu^{-p} \right) \) for some \( \beta > 0 \) such that

\[
|u_1 u_2 \cdots u_\ell - \tilde{x}_{\ell,\nu}(u_1, \ldots, u_\ell)| \leq \nu, \quad \forall u_1, \ldots, u_\ell \in [-1, 1] \quad (32)
\]

and

\[
\tilde{x}_{\ell,\nu}(u_1, \ldots, u_\ell) = 0, \quad \text{if } u_j = 0 \quad \text{for some } j = 1, \ldots, \ell. \quad (33)
\]

With the above tools, we can prove Theorem 8 as follows.

**Proof of Theorem 8:** Let \( N_{\tau} = N_{-\tau,\tau/2} \) be given in Lemma 3 and \( \tilde{x}_{2,\nu} : \mathbb{R}^2 \to \mathbb{R} \) in Lemma 4 with \( \ell = 2 \). Then it follows from (28) that

\[
N_{\tau}(x - x_i) = \begin{cases} 0, & \text{if } x \notin [x_i + [-3\tau/2, 3\tau/2]^d, \\ 1, & \text{if } x \in [x_i + [-\tau, \tau]^d. \end{cases} \quad (29)
\]

Since \( \|g_{n,L,U}\|_{\mathcal{L}^p(\mathbb{R}^d)} \leq C^* \), we can define a function \( N_{\tau,v,D,U} \) on \( \mathbb{R}^d \) by

\[
N_{\tau,v,D,U}(x) := \sum_{i=1}^{m} y_i N_{\tau}(x - x_i) + C^* \tilde{x}_{2,\nu}\left( \frac{g_{n,L,U}(x)}{C^*}, 1 - \sum_{i=1}^{m} N_{\tau}(x - x_i) \right). \quad (30)
\]

Moreover, for \( i \neq j \) and any \( x \in \mathbb{R}^d \),

\[
\|x_i - x - (x_j - x)\|_2 = \|x_i - x_j\|_2 \leq \frac{2q_{\nu}}{\sqrt{d}} > 3\tau
\]

implies \( N_{\tau}(x - x_j) = 0 \). Hence \( 1 - \sum_{i=1}^{m} N_{\tau}(x - x_i) = 0 \). Therefore, Lemma 4 yields \( \tilde{x}_{2,\nu}\left( \frac{g_{n,L,U}(x)}{C^*}, 1 - \sum_{i=1}^{m} N_{\tau}(x - x_i) \right) = 0 \). This implies

\[
N_{\tau,v,D,U}(x_j) = y_j, \quad j = 1, \ldots, m. \quad (32)
\]

Define further a function \( h_D \) on \( \mathbb{R}^d \) by

\[
h_D(x) := \sum_{i=1}^{m} y_i N_{\tau}(x - x_i) + g_{n,L,U}(x) \left( 1 - \sum_{i=1}^{m} N_{\tau}(x - x_i) \right). \quad (33)
\]

It follows from Lemma 4 that

\[
|h_D(x) - N_{\tau,v,D,U}(x)| \leq \nu, \quad \forall x \in \mathbb{R}^d. \quad (33)
\]

If \( x_i - x \notin [-3\tau/2, 3\tau/2]^d \) for all \( i = 1, \ldots, m \), then it follows from (29) that \( \sum_{i=1}^{m} N_{\tau}(x - x_i) = 0 \), which implies \( h_D(x) = g_{n,L,U}(x) \). Hence,

\[
\|g_{n,L,U} - h_D\|_{\mathcal{L}^p(\mathbb{R}^d)} = \int_{\mathbb{R}^d} \|g_{n,L,U}(x) - h_D(x)\|^p dx \leq \sum_{i=1}^{m} \int_{[-3\tau/2, x_i + 3\tau/2]^d} |g_{n,L,U}(x) - h_D(x)|^p dx \leq m(3\tau^d)^{2p} C^*^p. \quad (34)
\]

This implies

\[
\|g_{n,L,U} - h_D\|_{\mathcal{L}^p(\mathbb{R}^d)} \leq 2C^* 3^d m^{1/p} \nu^{d/p}. \quad (34)
\]

The above estimate together with (33) yields

\[
\|g_{n,L,U} - N_{\tau,v,D,U}\|_{\mathcal{L}^p(\mathbb{R}^d)} \leq \|h_D - N_{\tau,v,D,U}\|_{\mathcal{L}^p(\mathbb{R}^d)} + \|g_{n,L,U} - h_D\|_{\mathcal{L}^p(\mathbb{R}^d)} \leq 2^d \nu + 2C^* 3^d m^{1/p} \nu^{d/p}. \quad (34)
\]

Set \( \nu = \varepsilon \) and \( \tau \leq \min\{2q_{\nu}/(3 \sqrt{d}), m^{-1/d} \nu^{d/p}\} \). We obtain

\[
\|g_{n,L,U} - N_{\tau,v,D,U}\|_{\mathcal{L}^p(\mathbb{R}^d)} \leq C^* \varepsilon, \quad (34)
\]
where $C' := 2^{d/p} + 2C^* 3^{d/p}$. Denote by $N'(t) = \sigma(t) - \sigma(-t) = t$.

Recalling (30), we can define

$$f_{D,n,d,L,U,q} := \sum_{i=1}^{m} \phi_i N'(\cdots \cdot N([x_i - x_i]_{\mathbb{Z}}))$$

$$+ C^* \sum_{j=1}^{m} \phi_j \left( \frac{\|g_{D,n,L,U}(x)\|_{C^*}}{C^*} \right)^{1} - \sum_{i=1}^{m} \phi_i N([x_i - x_i]_{\mathbb{Z}})$$

with $\tau$ and $v$ as above so that the two items on the righthand side of $f_{D,n,d,L,U,q}$ have the same depth. Then $f_{D,n,d,L,U,q}$ is a DFCN of depth $O(L + \log e^{-1})$ and width $O(m + U + \log e^{-1})$. Noting further $\rho_x \in \Xi_p$, we then have $\|f\|_{L^2(\mathbb{Z}^d)} \leq D_{p,x} \|f\|_{L^2(\mathbb{Z}^d)}$. This together with (34) yields

$$\|g_{D,n,L,U} - f_{D,n,d,L,U,q}\|_{L^2(\mathbb{Z}^d)} \leq C' e^{v}.$$

Recalling that there are infinitely many $\tau$ satisfying $\tau \leq \min\{2a_1/(3 \sqrt{d}), m^{-1/d} \rho / \|i\|_{d}\}$, then there are infinitely many such $f_{D,n,d,L,U,q}$. Theorem 8 is then proved by scaling. \hfill \Box

### B. Existing Theoretical Results for UPDNN

Let us begin with the widely used class of smooth regression functions. The classical results in [14] showed that

$$C'_1 m^{-\frac{2d}{2d}} \leq e(\inf_{f \in \mathcal{L}(\mathbb{R}^d), \|\rho\|_{\mathbb{R}^d}}) \leq C'_2 m^{-\frac{2d}{2d}}, \quad p \geq 2,$$

where $C'_1$, $C'_2$ are constants independent of $m$ and $e(\inf_{f \in \mathcal{L}(\mathbb{R}^d), \|\rho\|_{\mathbb{R}^d}})$ is defined by (6). It demonstrates the optimal generalization error rates for $f^* \in \mathcal{L}(\mathbb{R}^d)$ and $\rho_x \in \Xi_p$ that a good learning algorithm should achieve. Furthermore, it can be found in [9] and [10] that there is some network structure $\Phi_{n,d}$ such that all global minima of ERM-UPDNN can reach almost optimal error rates in the sense that if $L \sim \log m$, $d_1 \sim m^{\frac{2d}{2d}}$ and $d_2, \ldots, d_L \sim \log m$, then

$$C'_1 m^{-\frac{2d}{2d}} \leq \sup_{f \in \mathcal{L}(\mathbb{R}^d), \|\rho\|_{\mathbb{R}^d}} e\left(\inf_{f \in \mathcal{L}(\mathbb{R}^d)} - \|f\|_{L^2(\mathbb{Z}^d)}^2\right) \leq C'_2 \left(\frac{m}{\log m}\right)^{-\frac{2d}{2d}},$$

where $C'_1$ is a constant depending only on $r, c_0, d, p$ and $\inf_{f \in \mathcal{L}(\mathbb{R}^d)}$ is an arbitrary global minimum of (4) with depth and width specified as above.

The above interesting result shows that ERM-UPDNN performs not worse than any other learning schemes, but it does not demonstrate the power of depth since shallow learning also reaches these bounds [14]. We then present their power in learning additive models. If $L \sim \log m$, $d_1 \sim m^{\frac{2d}{2d}}$ and $d_2, \ldots, d_L \sim \log m$, it can be found in [9] and [10] that for any $p \geq 2$, there holds

$$C'_3 \left(\frac{m}{\log m}\right)^{-\frac{2d}{2d}} \leq e(\mathcal{W}_{r,c_0,d_0}) \leq C'_4 \left(\frac{m}{\log m}\right)^{-\frac{2d}{2d}},$$

where $C'_3$, $C'_4$ are constants independent of $m$ and $\inf_{f \in \mathcal{L}(\mathbb{R}^d)}$ is an arbitrary global minimum of (4). Noticing that shallow learning is difficult to achieve the above generalization error rates; even for a special case of the generalized additive model with $f(x_1) = (x_1)^2$, it has been proved in [59] that shallow nets with any activation functions cannot achieve the aforementioned generalization error rates.

Finally, we introduce the power of depth in capturing the spatially sparse features to embody the role of massive data [11]. Let $L \sim \log m$, $d_1, d_2 \sim (m^{\frac{2d}{2d}})^{1/(2d+2)}$, $m_{1}, \ldots, m_{2} \sim \log m$. If $m$ is large enough to satisfy $\frac{m}{\log m} \geq \frac{C'_4}{\|f\|_{\mathcal{L}(\mathbb{R}^d)}}$, then it can be easily deduced from [11] and [54] that there exists a deep net $f_{\text{under}}$ with the aforementioned structure such that

$$C'_5 m^{-\frac{2d}{2d}} \leq e(\inf_{f \in \mathcal{L}(\mathbb{R}^d)}) \leq C'_6 \left(\frac{m}{\log m}\right)^{-\frac{2d}{2d}},$$

where $C'_5$, $C'_6$, $C'_7$ are constants independent of $m$ and $f_{\text{under}} \in \mathcal{L}(\mathbb{R}^d)$ is an arbitrary global minimum of ERM on the DFCN. In (38), $\left(\frac{m}{\log m}\right)^{-\frac{2d}{2d}}$ reflects the smoothness of $f^*$ and $\left(\frac{m}{\log m}\right)^{-\frac{2d}{2d}}$ embodies the spatial sparseness of $f^*$.

Given a sparsity level $u$ and the number of partitions $N^* d$, the size of data should satisfy $\frac{m}{\log m} \geq C'_4 \chi_{\text{under}}$ to embody the spatial sparsity feature of $f^*$. In particular, if the number of samples is smaller than the sparsity level $u$, it is impossible to develop learning schemes to realize the support of $f^*$. Recalling the localized approximation property of deep ReLU nets [11], [39], with the help of massive data, (38) shows that deep ReLU nets are capable of capturing the spatial sparseness, which is beyond the capability of shallow nets due to its lack of localized approximation [39]. We refer the readers to [11] for more details about the above assertions. If $p = 2$, it can be found in (38) the estimates achieve almost optimal generalization error rates.

### C. Proofs

In this part, we prove our main results by using the proposed deepening scheme (for results concerning noisy data) and a functional analysis approach developed in [60] (for results concerning noise-free data).

Firstly, we prove Proposition 1 based on Lemma 3.

**Proof of Proposition 1:** If $y_i = 0$, $i = 1, \ldots, m$, we can set $f_{D,n}(x) = 0$. Then our conclusion naturally holds. Otherwise, we define

$$N_{r,d}(x) := \sum_{i=1}^{m} \phi_i N_{\tau,r,d/2}(x - x_i),$$

If $\tau < \frac{2a_1}{3 \sqrt{d}}$, then it follows from (30) that $N_{r,d}(x_i) = y_i$. Since $\|f^*\|_{L^2(\mathbb{Z}^d)} \geq c$, a direct computation yields $\|f^* - N_{r,d}\|_{L^2(\mathbb{Z}^d)} \geq \|f^*\|_{L^2(\mathbb{Z}^d)} \geq c - \|N_{r,d}\|_{L^2(\mathbb{Z}^d)}$.

But (29) together with (14) and $N_{r,d}(x - x_i) \leq 1$ for any $x \in \mathbb{Z}^d$ yields

$$\|N_{r,d}\|_{L^p(\mathbb{Z}^d)} \leq \sum_{i=1}^{m} \left(\int_{\mathbb{Z}^d} |f^*(x_i)|N_{r,d}(x - x_i)^p \, dx\right)^{1/p}.$$
Lemma 5: Let $v \in (0, 1)$ and $s, N \in \mathbb{N}_0$. If $v = N^{-r/d}$ and $f \in \text{Lip}_v^{(r,c_0)}$ with $0 < r \leq s + 1$, then there holds
\[
\min_{h \in \Phi_{N,v}} \|f - h\|_{L_v^{(r,c_0)}} \leq C'_1 c_0 N^{-r/d},
\]
where $C'_1$ is a constant depending only on $d$ and $r$.

To prove Theorem 6, we need the following lemma proved in [60]. It should be mentioned that for DFCN with larger depth and width, the above assertions obviously hold. We use the following functional analysis tool that presents a close relation interpolation and approximation to minimize the depth.

Lemma 6: Let $U$ be a (possibly complex) Banach space, $\mathcal{V}$ a subspace of $\mathcal{U}$, and $W^* = \text{span}\{\mathcal{U}^*, \mathcal{V}^*\}$ a finite-dimensional subspace of $\mathcal{U}^*$, the dual of $\mathcal{U}$. If for every $w^* \in W^*$ and some $\gamma > 1$, $\gamma$ independent of $w^*$,
\[
\|w^*\|_{\mathcal{U}^*} \leq \gamma \|w^*\|_{\mathcal{V}^*},
\]
then for any $u \in \mathcal{U}$ there exists $v \in \mathcal{V}$ such that $v$ interpolates $u$ on $W^*$; that is, $w^*(u) = w^*(v)$ for all $w^* \in W^*$. In addition, $v$ approximates $u$ in the sense that $\|u - v\|_{\mathcal{U}} \leq (1 + 2\gamma) \text{dist}_{\mathcal{U}}(u, \mathcal{V})$.

To use the above lemma, we need to construct a special function to facilitate the proof. Our construction is motivated by [60]. For any $w^* = \sum_{j=1}^m c_j \delta_{x_j} \in W^*$, define
\[
g_w(x) = \sum_{j=1}^m \text{sgn}(c_j) \left(1 - \frac{\|x - x_j\|_{q_{\Lambda_j}}}{q_{\Lambda_j}}\right),
\]
where $\delta_{x_j}$ is the point evaluation operator and $\text{sgn}(t)$ is the sign function satisfying $\text{sgn}(t) = 1$ for $t \geq 0$ and $\text{sgn}(t) = 0$ for $t < 0$. Then it is easy to see that $g_w$ is a continuous function. In the following, we present three important properties of $g_w$.

Lemma 7: Let $W^* = \text{span}\{\delta_{x_j} : i = 1, \ldots, m\}$. Then for any $w^* \in W^*$, there holds (i) $\|g_w\|_{L_v^{(r,c_0)}} = 1$, (ii) $w^*(g_w) = \|w^*\|$, (iii) $g_w \in \text{Lip}_v^{(1,q_{\Lambda_j})}$.

Proof: Denote $A_j = B(x_j, q_{\Lambda_j}) \cap \mathbb{R}^d$, where $B(x_j, q_{\Lambda_j})$ is the ball with center $x_j$ and radius $q_{\Lambda_j}$. Then it follows from the definition of $q_{\Lambda_j}$ that $A_j \cap \Lambda_k = \emptyset$, where $A_j = A_j \setminus \partial A_j$ and $\partial A_j$ denotes the boundary of $A_j$. Without loss of generality, we assume $\mathbb{R}^d \setminus \bigcup_{j=1}^m A_j \neq \emptyset$. From (46), we have $g_w(x) = 0$ for $x \in \mathbb{R}^d \setminus \bigcup_{j=1}^m A_j \neq \emptyset$. If there exist some $j \in \{1, \ldots, m\}$ such that $x \in A_j$, then
\[
g_w(x) = \text{sgn}(c_j) \left(1 - \frac{\|x - x_j\|_{q_{\Lambda_j}}}{q_{\Lambda_j}}\right).
\]
So
\[
|g_w(x)| = 1 - \frac{\|x - x_j\|_{q_{\Lambda_j}}}{q_{\Lambda_j}} \leq |g_w(x_j)| = 1.
\]
Thus, $|g_w(x)| \leq 1$ for all $x \in \mathbb{R}^d$. Since $|g_w(x_j)| = 1$, $j = 1, \ldots, m$, we get $\|g_w\|_{L_v^{(r,c_0)}} = 1$, which verifies (i). For $w^* \in W^*$, we have
\[
w^*(g_w) = \sum_{j=1}^m c_j g_w(x_j) = \sum_{j=1}^m c_j g_w(x_j).
\]
Thus (ii) holds. The remainder is to prove that $g_w$ satisfies (iii). We divide the proof into four cases.
If $x, x' \in A_j$ for some $j \in \{1, \ldots, m\}$, then it follows from (46) that
\[
\frac{|g_{w'}(x) - g_{w'}(x')|}{q_{A}} = \frac{\text{sgn}(c_j) \left(1 - \frac{|x - x'_1|}{q_{A}}\right) - \text{sgn}(c_j) \left(1 - \frac{|x' - x'_2|}{q_{A}}\right)}{q_{A}}
\leq \frac{|x - x'_1|}{q_{A}} + \frac{|x' - x'_2|}{q_{A}} 
\leq \frac{|x - x'|}{q_{A}}.
\]

If $x, x' \in \mathbb{I} \setminus \bigcup_{j=1}^{m} A_j$, then the definition of $g_{w}$ yields $g_{w}(x) = g_{w}(x') = 0$, which implies $|g_{w}(x) - g_{w}(x')| \leq \frac{|x - x'|}{q_{A}}$. If $x \in A_j, x' \in A_k$ for $j \neq k$, then it is easy to see that for any $z \in \partial B(x_j, q_{A}), j = 1, \ldots, m$, there holds $g_{w}(z) = 0$. Let $z_j, z_k$ be the intersections of the line segment $xx'$ and $\partial B(x_j, q_{A})$, and the line segment $xx'$ and $\partial B(x_k, q_{A})$, respectively. Then, we have $|x - x'|/q_{A} \geq |x - z_j|/q_{A} + |x' - z_k|/q_{A}$. Since $x, z_j \in A_j, x', z_k \in A_k$ and $g_{w}(z_j) = g_{w}(z_k) = 0$, we have
\[
\frac{|g_{w}(x) - g_{w}(x')|}{q_{A}} \leq \frac{|x - z_j|}{q_{A}} + \frac{|x' - z_k|}{q_{A}} \leq \frac{|x - x'|}{q_{A}}.
\]

Combining all the above cases verifies $g_{w} \in Li_{p_{1}}^{(q_{A})}$. This completes the proof of Lemma 7.

With these, we are in a position to prove Theorem 6.

**Proof of Theorem 6:** Let $\mathcal{U} = C(\mathcal{I})$, the space of continuous functions defined on $\mathbb{I}^2$, $\mathcal{W}^\gamma = \text{span}(d_{\mathcal{I}})_{i=1}^{m}$ and $\mathcal{V} = \Phi_{\mathcal{N}, \nu, s}$ in Lemma 6. For every $w^* \in \mathcal{W}^\gamma$, we have $w^* = \sum_{i=1}^{m} c_i d_{\mathcal{I}}$, for some $c = (c_{1}, \ldots, c_{m})^{T} \in \mathbb{R}^{m}$ without loss of generality, we assume $\|w^*\| = \sum_{i=1}^{m} |c_{i}| = 1$. Let $g_{w}$ be defined by (46). Then, it follows from Lemma 5 and Lemma 7 that there is some $h_{\otimes} \in \mathcal{V}$ such that
\[
\|g_{w} - h_{\otimes}\|_{L^{2}(\mathcal{I})} \leq C_{\gamma}^{-1} N^{-1/d}.
\]

Let $G = \left[\left(\gamma + 1\right)^{C_{\gamma}} \left(\gamma^{-1} N_{A}\right)^{T}\right]^{-1}$ for some $\gamma > 1$. We have
\[
\|g_{w} - h_{\otimes}\|_{L^{2}(\mathcal{I})} \leq \gamma^{-1} \gamma + 1
\]
This together with (i) in Lemma 7 yields
\[
\|h_{\otimes}\|_{L^{2}(\mathcal{I})} \leq \gamma^{-1} \gamma + 1 = \frac{2\gamma}{\gamma + 1}.
\]
Since $w^*$ is a linear operator and (ii) in Lemma 7 holds, there holds
\[
1 = \|w^*\| = w^*(g_{w} - h_{\otimes}) + w^*(h_{\otimes}).
\]
Hence, from
\[
\|w^*(g_{w} - h_{\otimes})\|_{L^{2}(\mathcal{I})} \leq \|w^*\|\|g_{w} - h_{\otimes}\|_{L^{2}(\mathcal{I})} \leq \frac{2\gamma}{\gamma + 1},
\]
we have
\[
w^*(h_{\otimes}) \geq 1 - \|w^*(g_{w} - h_{\otimes})\| \geq 1 - \frac{\gamma - 1}{\gamma + 1} = \frac{2}{\gamma + 1}.
\]
Consequently,
\[
\|w^*\| = 1 \leq \frac{\gamma + 1}{2 \gamma} w^*(h_{\otimes}) \leq \frac{\gamma + 1}{2 \gamma} \|w^*\|_{\Phi_{\mathcal{N}, \nu, s}} \|h_{\otimes}\|_{L^{2}(\mathcal{I})} \leq \frac{\gamma + 1}{2 \gamma} \frac{2\gamma}{\gamma + 1} \|w^*\|_{\Phi_{\mathcal{N}, \nu, s}}.
\]

Setting $\gamma = 2$, for any $f^* \in Lip^{(q_{A})}$, it follows from Lemma 6 and Lemma 5 that there exists some $h^* \in \mathcal{V} = \Phi_{\mathcal{N}, \nu, s}$ such that $h^*(x_i) = f^*(x_i)$ and
\[
\|h^* - f^*\|_{L^{2}(\mathcal{I})} \leq 5 \min_{\gamma \in \gamma_{\mathcal{Y}}} \|h - f^*\|_{L^{2}(\mathcal{I})} \leq 5C_{\gamma}^{-1} N^{-1/d}.
\]

Setting $\gamma = N^{-1/d}$ and recalling (44) and $t = \sigma(t) - \sigma(-t)$, $\Phi_{\mathcal{N}, \nu, s}$ defined in (44) can be regarded as the set of DFCNs with depth $O(\log N)$, $d_{l} = O(N^{d^2})$ and $d_{L} = O(\log N)$. Noting that there are infinitely many $\gamma \approx N^{-1/d}$, there are infinitely many DFCNs satisfying the above assertions. This completes the proof of Theorem 6.

The proofs of the other main results are simple by combining Theorem 8 with existing results.

**Proof of Theorem 7:** Theorem 7 is a direct consequence of Theorem 8.

**Proof of Corollary 2:** Setting the teacher network $g = f^\star_{\text{under global}}$ in (36), we obtain a student net $h$ based on Theorem 8 with $e = t^{-2/(2+d)}$. This derives (21). We can set the teacher network to be the UPDNN $f^\star_{\text{global}}$ in (37). Then, (22) follows from Theorem 8 and (37) directly. According to Theorem 8, it is easy to obtain a student network $h$ based on the teacher network $g = f^\star_{\text{global}}$ in (38). Then, (23) follows directly from Theorem 8 and (38). This proves Corollary 2.

**ACKNOWLEDGMENT**

The work of Ding-Xuan Zhou was completed when he was with the City University of Hong Kong.

**REFERENCES**

[1] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” Neural Comput., vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 25, 2012, pp. 1097–1105.

[3] H. Lee, P. Pham, Y. Largman, and A. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 22, 2009, pp. 1096–1104.

[4] D. Silver et al., “Mastering the game of go with deep neural networks and tree search,” Nature, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[5] R. Sun, D. Li, S. Liang, T. Ding, and R. Srikant, “The global landscape of neural networks: An overview,” IEEE Signal Process. Mag., vol. 37, no. 5, pp. 95–108, Sep. 2020.

[6] Z. Allen-Zhu, Y. Li, and Z. Song, “A convergence theory for deep learning via over-parameterization,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 242–252.

[7] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding deep learning (still) requires rethinking generalization,” Commun. ACM, vol. 64, no. 3, pp. 107–115, Mar. 2021.

[8] T. Poggio, H. Mhasakar, L. Rosasco, B. Miranda, and Q. Liao, “Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review,” Int. J. Autom. Comput., vol. 14, no. 5, pp. 503–519, Oct. 2017.
[49] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *Ann. Statist.*, vol. 48, no. 4, pp. 1875–1897, Aug. 2020.

[50] Z. Han, S. Yu, S.-B. Lin, and D.-X. Zhou, "Depth selection for deep ReLU nets in feature extraction and generalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1853–1868, Apr. 2022.

[51] C. K. Chui, S.-B. Lin, B. Zhang, and D.-X. Zhou, "Realization of spatial sparseness by deep ReLU nets with massive data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 229–243, Jan. 2022.

[52] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parametrization," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, Jun. 2019, pp. 242–252.

[53] D. Zou, Y. Cao, D. Zhou, and Q. Gu, "Gradient descent optimizes overparameterized deep ReLU networks," *Mach. Learn.*, vol. 109, no. 3, pp. 467–492, Mar. 2020.

[54] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-free Theory of Nonparametric Regression*, vol. 1, Berlin, Germany: Springer, 2002.

[55] D. Li, T. Ding, and R. Sun, "On the benefit of width for neural networks: disappearance of basins," *SIAM J. Optim.*, vol. 32, no. 3, pp. 1728–1758, Sep. 2022.

[56] Y. Cao and Q. Gu, "Generalization error bounds of gradient descent for learning over-parameterized deep ReLU networks," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 4, pp. 3349–3356.

[57] D. Zou and Q. Gu, "An improved analysis of training over-parameterized deep neural networks," in *Proc. Adv. Neural Inf. Proc. Syst.*, vol. 32, Jan. 2019, pp. 2053–2062.

[58] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," in *Proc. Int. Conf. Learn. Represent.* (ICLR), Banff, AB, Canada, 2014.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Las Condes, Chile, Dec. 2015, pp. 1026–1034.

[60] A. Daniely, "SGD learns the conjugate kernel class of the network," in *Proc. Adv. Neural Inf. Proc. Syst.*, vol. 30, Feb. 2017, pp. 2422–2430.

[61] G. Zhang, J. Martens, and R. B. Gross, "Fast convergence of natural gradient descent for over-parameterized neural networks," in *Proc. Adv. Neural Inf. Proc. Syst.*, vol. 32, May 2019, pp. 8080–8091.

[62] J. Zeng, Y. T. K. Lau, S. B. Lin, and Y. Yao, "Global convergence of block coordinate descent in deep learning," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, May 2019, pp. 7313–7323.

[63] S. Bock and M. G. Weiß, "Local convergence of adaptive gradient descent optimizers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 13, pp. 1–15, Dec. 2020.

[64] J. An and J. Lu, "Convergence of stochastic gradient descent under a local Lojasiewicz condition for deep neural networks," 2023, arXiv:2304.09221.

[65] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 32, pp. 15849–15854, Aug. 2019.

[66] M. A. Belkin, "Approximation beats concentration? An approximation view on inference with smooth radial kernels," in *Proc. Conf. Learn. Theory*, Jan. 2018, pp. 1348–1361.

[67] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 48, pp. 30063–30070, Dec. 2020.

[68] T. Liang and A. Rahklin, "Just interpolate: Kernel ‘ridgeless’ regression can generalize," *Ann. Statist.*, vol. 48, no. 3, pp. 1329–1347, 2020.

[69] M. Belkin, D. Hsu, and P. P. Mitra, "Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate," in *Proc. Adv. Neural Inf. Proc. Syst.*, vol. 31, 2018, pp. 2300–2311.

[70] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *Ann. Statist.*, vol. 50, no. 2, pp. 949–986, Apr. 2022.

[71] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai, "Harmless interpolation of noisy data in regression," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 67–83, May 2020.

[72] S.-B. Lin, X. Chang, and X. Sun, "Kernel interpolation of high dimensional scattered data," 2020, arXiv:2009.01514.

[73] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *Ann. Statist.*, vol. 48, no. 4, pp. 1875–1897, Aug. 2020.