Check for updates

J. Appl. Probab. 1–22 (2026)
*doi:*10.1017/jpr.2026.10121

STATISTICAL GUARANTEES FOR *K*-DEGREE NEURAL NETWORKS VIA SCALE REGULARIZATION

PENG CHEN,* *and*
CHENGJUN KUANG,* *Nanjing University of Aeronautics and Astronautics*
TING ZHANG ,** *Nanjing University of Information Science and Technology*

Abstract

Neural networks have been widely used in data analysis, but there is relatively little research on the statistical theory of neural networks. In this paper we generalize the scale regularization method to the k-degree neural networks, and obtain a statistical guarantee for estimators with a least-squares term and a regularizer. On this basis, we exemplify this guarantee with locally Lipschitz activation functions and l_β-regularization for $\beta \in [1, 2]$ in the case $k > 1$, showing that the corresponding prediction error decrease is of the order of $n^{-1/2}$ in the sample size n.

Keywords: k-degree neural networks; scale regularization; locally Lipschitz activation functions; prediction guarantee

2020 Mathematics Subject Classification: Primary 65C20
Secondary 62M45

1. Introduction

Neural networks have become an interesting subject in a variety of areas, including pattern recognition [1, 19], autonomous vehicles [33, 40], image segmentation [3, 24], and other fields. In recent years, extensive research on optimization and statistical guarantees has been conducted in [7], [36], [38], and the references therein. Substantial effort and contributions have been devoted to the algorithmic and computational aspects, while theoretical problems still remain to be solved urgently. One can refer to [7] and references therein for reviews. The challenges exist mainly due to the activation function and the complex structure of neural networks [44].

We begin by introducing the basic regression framework adopted throughout this paper. Consider independent observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ generated from the following non-parametric regression model:

$$y_i = g_*(\mathbf{x}_i) + \epsilon_i, \quad i \in \{1, \ldots, n\}, \tag{1.1}$$

where $g_* \colon \mathbb{R}^d \to \mathbb{R}$ denotes the unknown underlying regression function, and $\epsilon_1, \ldots, \epsilon_n \in \mathbb{R}$ represent independent random noise variables. Our objective is to approximate the target function g_* via feedforward neural networks. For a given input $\mathbf{x} \in \mathbb{R}^d$, the network output is

Received 10 April 2024 UTC; accepted 21 June 2026 UTC.
* Postal address: School of Mathematics, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China.
** Postal address: School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China. Email address: zhangtstat@nuist.edu.cn

© The Author(s), 2026. Published by Cambridge University Press on behalf of Applied Probability Trust.

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

formulated as

$$g_\Theta(\mathbf{x}) := W^L \mathbf{f}^L(\cdots W^1 \mathbf{f}^1(W^0 \mathbf{x})\cdots), \tag{1.2}$$

where the network parameter set is defined by

$$\mathcal{A} := \{\Theta = (W^L, \ldots, W^0) \colon W^l \in \mathbb{R}^{p_{l+1}\times p_l},\ l = 0, 1, \ldots, L\}.$$

Here $\mathbf{x}_i$ and y_i correspond to network inputs and outputs, respectively; the W^l stand for layer-wise weight matrices; L denotes the number of hidden layers; and $p_0, \ldots, p_{L+1}$ denote the neuron number of each layer. The mapping $\mathbf{f}^l \colon \mathbb{R}^{p_l} \to \mathbb{R}^{p_l}$ refers to the activation function of the lth layer.

As the depth of the neural network increases, the complexity of the model and the number of parameters in the untrained network continue to expand; thus the model becomes prone to overfitting (see [35] and [46]). To avoid overfitting and reduce generalization errors, a common approach is to apply regularization techniques. Regularization theory has developed accordingly, and much work has been done to improve the robustness and generalization ability of neural networks. Works such as that of Wu *et al.* [43] introduce a simple consistency training strategy (R-Drop) to regularize dropout, compensating for the non-negligible inconsistency between training and inference. Yildirim and Özkale [45] conclude that biased estimators may be powerful tools in extreme learning machine studies and that ridge-based regression estimators may outperform extreme machine learning. In [38], Taheri *et al.* investigate the scale property of 1-degree neural networks by applying the idea of scale regularization. In [39], they consider statistical guarantees for approximating stationary points using the strategy of scale regularization. Subsequently, the idea of scale regularization has been widely applied in many ways, such as connection- and node-sparse deep learning [22], and reducing computational and statistical complexity in machine learning [27].

Among existing regularization strategies, scale regularization has attracted increasing attention for homogeneous network analysis. In [38], Taheri *et al.* explore scale regularization for classical 1-degree neural networks, while in [39] they further establish statistical guarantees for approximate stationary points under the same framework. Subsequently, the scale regularization idea has been extended to sparse deep learning [22] and complexity reduction in machine learning [27].

Notably, scale regularization inherently requires activation functions to satisfy non-negative homogeneity. For theoretical analysis and statistical guarantee derivation, existing studies commonly impose the standard global Lipschitz condition on activations, with typical examples including ReLU [21] and ELU [10]. Nevertheless, most existing results are confined to 1-degree networks with globally Lipschitz activations. For general k-degree neural networks and activation functions that only satisfy local Lipschitz continuity, polynomial growth considerably elevates analytical complexity, leading to a scarcity of rigorous statistical theories in this setting.

Motivation and definition of k-degree neural networks. To provide a self-contained presentation, we first clarify the conceptual connotation and research motivation of k-degree neural networks. Conceptually, a k-degree neural network adopts activation functions with k-degree non-negative homogeneity: scaling the input vector by a non-negative constant c yields an output scaled by c^k. The case $k = 1$ recovers conventional homogeneous networks such as ReLU-based structures, whereas $k > 1$ enables the network to naturally characterize high-order

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

polynomial interactions among input features. Such modeling flexibility is particularly useful in physical system simulation, high-dimensional non-parametric regression, and geometric deep learning, where homogeneous structures and polynomial feature interactions widely exist; see more details in [12], [14], [18], [19], [20], and [29]. From a theoretical perspective, the k-degree homogeneous structure facilitates the application of scale regularization and enables us to derive sharper statistical guarantees that are unattainable in standard network frameworks.

We now formalize the relevant mathematical definition.

Definition 1.1. (*Non-negative homogeneity and positive definiteness.*) For any $d_1, d_2 \in \mathbb{N}$, we say that a function $\mathbf{f}\colon \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is non-negative homogeneous of degree $k \in (0, \infty)$ if, for any $c \in [0, \infty)$ and $\mathbf{z} \in \mathbb{R}^{d_1}$, we have

$$\mathbf{f}(c\mathbf{z}) = c^k \mathbf{f}(\mathbf{z}),$$

and we say that a function $\mathbf{f}\colon \mathbb{R}^{d_1} \to [0, \infty)$ is positive definite if

$$\mathbf{f}(\mathbf{z}) = 0 \quad \Leftrightarrow \quad \mathbf{z} = \mathbf{0}_{d_1}.$$

Novelties and contributions. Compared with [38], our contributions are non-trivial and threefold.

(i) Our framework covers general k-degree neural networks, which substantially generalize the standard networks studied in [38]. Their layered polynomial structure requires entirely new analytical tools and theoretical arguments.

(ii) The extension to general ℓ_β-regularization entails non-trivial refinements to concentration inequalities and to the decomposition of approximation and estimation errors under general convex penalties.

(iii) We establish new statistical guarantees tailored to k-degree architectures that cannot be directly inferred from ℓ_1-based theory, providing fresh theoretical insights into regularized deep networks.

Our analysis employs novel techniques to bound approximation errors and derive oracle inequalities under polynomial homogeneity. We further develop iterative constructions to handle the additional complexity caused by polynomial growth and to establish Lipschitz continuity of network mappings under locally Lipschitz activation functions.

The remainder of this paper is organized as follows. In Section 2 we establish the framework of scale regularization for k-degree neural networks and give a general prediction error for regularized least-squares estimators. In Section 3 we consider an example with l_β-regularization for $\beta \in [1, 2]$ and a sequence of sub-Gaussian noise vectors, showing that the corresponding prediction error decreases by the second root of the number of samples. In Section 4 we conclude our paper and give some discussions. In Appendix A we provide the proof of proposition and theorems in Sections 2 and 3.

2. Framework and statistical guarantee

2.1. Scale regularization for k-degree neural networks

In this subsection we conduct the theoretical analysis of the proposed network architecture. Specifically, the 0th layer corresponds to the input layer with dimension $p_0 = d$, and the $(L+1)$th layer serves as the output layer with $p_{L+1} = 1$. The total number of network

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

parameters is defined as

$$P := \sum_{l=0}^{L} p_{l+1} p_l.$$

Based on the properties of k-degree non-negative homogeneous functions and positive definite functions introduced earlier, we present the following proposition that characterizes an alternative parametrization. The proof is provided in Appendix A.1.

Proposition 2.1. *Suppose that the activation functions* $\mathbf{f}^1, \ldots, \mathbf{f}^L$ *are non-negative homogeneous of degree* k. *Let the function* $h\colon \mathcal{A} \to [0, \infty)$ *be non-negative homogeneous of degree* $\alpha \in (0, \infty)$ *and positive definite. Denote the unit ball by*

$$\mathcal{A}_h := \{\Theta \in \mathcal{A}\colon h(\Theta) \le 1\}.$$

Then, for any $\Theta \in \mathcal{A}$ *and* $\mathbf{x} \in \mathbb{R}^d$, *there exists a pair of* $\kappa \in (0, \infty)$ *and* $\Omega \in \mathcal{A}_h$ *such that*

$$g_\Theta(\mathbf{x}) = \kappa g_\Omega(\mathbf{x}). \tag{2.1}$$

Inversely, for every pair of $\kappa \in (0, \infty)$ *and* $\Omega \in \mathcal{A}_h$, *there exists a* $\Theta \in \mathcal{A}$ *such that the equality* (2.1) *holds.*

Motivated by the idea of scale regularization developed in [38] and Proposition 2.1 above, we estimate the true data generating function g_* to $[0, \infty) \times \mathcal{A}_h$ with the help of a parameter space of networks corresponding to $\{\kappa g_\Omega\colon \kappa \in [0, \infty), \Omega \in \mathcal{A}_h\}$. Hence we study the following neural networks: for any $\mathbf{x} \in \mathbb{R}^d$,

$$\kappa g_\Omega(\mathbf{x}) := \kappa U^L \mathbf{f}^L(\cdots U^1 \mathbf{f}^1(U^0 \mathbf{x})),$$

which is indexed by the scale parameters $\kappa \in [0, \infty)$ and $\Omega = (U^L, \ldots, U^0) \in \mathcal{A}_h$.

2.2. Statistical prediction guarantee

In deep learning, taking the high-dimensionality of the parameter space $\mathcal{A}$ into consideration, the usual approach to estimate the network parameters in model (1.1) is the least-squares estimator complemented by some type of regularization, and the concrete form can be written as follows:

$$\hat{\Theta}_{\mathrm{reg},h} \in \arg\min_{\Theta \in \mathcal{A}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - g_\Theta(\mathbf{x}_i))^2 + \lambda h(\Theta) \right\},$$

where $\lambda \in [0, \infty)$ is a tuning parameter and $h\colon \mathcal{A} \to [0, \infty)$ is a regularizer.

However, it seems to be difficult to analyze such estimators in network frameworks; for more detail on analyzing the estimators, we refer to [42] and [43] and the references therein. In this section we employ the alternative parametrization to incorporate regularizers. Based on the equality $g_{\hat{\Theta}_{LS}} = \hat{\kappa}_{LS} g_{\hat{\Omega}_{LS}}$ obtained by Proposition 2.1, we adopt the approach of scale regularization to transform the above estimation problem into the following framework:

$$(\hat{\kappa}_h, \hat{\Omega}_h) \in \arg\min_{\substack{\kappa \in [0,\infty) \\ \Omega \in \mathcal{A}_h}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \kappa g_\Omega(\mathbf{x}_i))^2 + \lambda \kappa \right\}. \tag{2.2}$$

Thus we turn to solving the regularization on the scale parameter κ.

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

Now we focus on the statistical guarantees for the parameter estimator described above. On the one hand, we consider the prediction error

$$\mathrm{err}(\kappa g_\Omega) := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\kappa g_\Omega(\mathbf{x}_i) - g_*(\mathbf{x}_i))^2} \quad \text{for } \kappa \in [0,\infty),\ \Omega \in \mathcal{A}_h,$$

which measures the accuracy of the network κg_Ω fitting the unknown function g_* on the data. On the other hand, for a new sample $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$, we measure the generalization error or prediction risk by

$$\mathrm{risk}(\kappa g_\Omega) := \mathbb{E}_{(\mathbf{x},y)}[(\kappa g_\Omega(\mathbf{x}) - y)^2] \quad \text{for } \kappa \in [0,\infty),\ \Omega \in \mathcal{A}_h.$$

Similar to high-dimensional linear regression, the core quantity of prediction guarantees is the effective noise [23], and thus we define the quantile $\lambda_{h,t}$ of the effective noise for given level $t \in [0, 1]$ through

$$\lambda_{h,t} = \min\{\delta \in [0,\infty) \colon \mathbb{P}(z_h \le \delta) \ge 1 - t\},$$

where

$$z_h := \sup_{\Omega\in\mathcal{A}_h}\left\{\frac{2}{n}\sum_{i=1}^{n} g_\Omega(\mathbf{x}_i)\epsilon_i\right\},$$

which plays an important role in exploring the properties of the function class $\{g_\Omega \colon \Omega \in \mathcal{A}_h\}$, such as the Gaussian complexity and Rademacher complexity [5].

By the same argument as the proof of [38, Theorem 2], we can immediately obtain the following prediction guarantee in terms of the prediction error.

Theorem 2.1. *Assume that* $\lambda \ge \lambda_{h,t}$ *for* $t \in [0, 1]$. *Then*

$$\mathrm{err}^2(\hat{\kappa}_h g_{\hat{\Omega}_h}) \le \inf_{\substack{\kappa\in[0,\infty)\\ \Omega\in\mathcal{A}_h}} \{\mathrm{err}^2(\kappa g_\Omega) + 2\lambda\kappa\},$$

with probability at least $1 - t$.

The error bound suggests that the squared prediction error of the regularized estimator is bounded by the squared bias term $\mathrm{err}^2(\kappa g_\Omega)$ and the variance term $2\lambda\kappa$. A small prediction error can be achieved with relatively small quantile $\lambda_{h,t}$ and scale κ.

3. Example: l_β-regularization for $\beta \in [1, 2]$

In this section we consider the case $k > 1$ and illustrate a concrete example of l_β-regularization to present an explicit bound. Specifically, we consider a weak regularizer proposed in [42], which has proved that the sample efficiency can indeed depend on the presence of the regularizer, and the weak regularizer often makes the optimization problem achieve the best results (the maximum normalized margin solution; see more details in [42, Section 4]). Recently, the weak regularizer has been widely used in [6], [26], and the references therein. Hence, similar to [42, (4.1)], for any $\beta \in [1, 2]$, we define the regularizer h as

$$h(\Omega) := \|\Omega\|_F^\beta := \left(\sum_{l=0}^{L}\sum_{k=1}^{p_{l+1}}\sum_{j=1}^{p_l}|U_{kj}^l|^2\right)^{\beta/2}.$$

Then h is non-negative homogeneous of degree β.

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

Before stating our main conclusions, we first give the definition of the Euclidean norm: for any $d \geq 1$ and $v \in \mathbb{R}^d$,

$$|\mathbf{v}| := \sqrt{\sum_{i=1}^{d} |v_i|^2}.$$

Then we suppose that the activation functions $\mathbf{f}^1, \ldots, \mathbf{f}^L$ satisfy the following assumption.

Assumption 3.1. For all $l \in \{1, \ldots, L\}$, let the activation functions $\mathbf{f}^l$ be non-negative homogeneous of degree $k \in (1, \infty)$ and positive definite. We further suppose that the activation functions satisfy $\mathbf{f}^l(\mathbf{0}_{p_l}) = \mathbf{0}_{p_l}$ and are locally Lipschitz-continuous for a constant $M_k \in (1, \infty)$ and with respect to the Euclidean norms on their input and output spaces:

$$|\mathbf{f}^l(\mathbf{z}) - \mathbf{f}^l(\mathbf{z}')| \leq M_k(|\mathbf{z}|^{k-1} + |\mathbf{z}'|^{k-1} + 1)|\mathbf{z} - \mathbf{z}'| \quad \text{for all } \mathbf{z}, \mathbf{z}' \in \mathbb{R}^{p_l}. \tag{3.1}$$

Assumption 3.1 is satisfied by many nonlinear activation functions. For example, the square ReLU function $z \mapsto \max\{z, 0\}^2$ [18] satisfies Assumption 3.1 with $M_2 = 2$, and when $k \geq 2$, the ReP_k function $z \mapsto \max\{z, 0\}^k$ [19] satisfies Assumption 3.1 with $M_k = k \times 2^{k-2}$. For more details on the activation functions, we refer the readers to [2], [9], [13], [17], [21], [31], [34], [37], and the references therein.

Furthermore, we assume the noise variables ϵ_i in model (1.1) satisfy the uniformly sub-Gaussian assumption [15, Chapter 8].

Assumption 3.2. For $i = 1, \ldots, n$, let the noise variables ϵ_i be independent, centered, and uniformly sub-Gaussian for constants $K_1, \gamma_1 \in (0, \infty)$:

$$\max_{i \in \{1,\ldots,n\}} K_1^2\big(\mathbb{E}e^{|\epsilon_i|^2/K_1^2} - 1\big) \leq \gamma_1^2.$$

For any $\mathbf{x} \in \mathbb{R}^d$, we define

$$B(\mathbf{x}) = 2^{L-1+\frac{(L+6)k^{L+3}}{(k-1)^2}} M_k^{\frac{(L+7)k^{L+3}}{k-1}} \left(\frac{2}{L}\right)^{L/2} \big(1 + \big(|\mathbf{x}|^{k-1} + |\mathbf{x}|^{(k-1)k^{L+1}}\big)^2\big)$$
$$\times \Big(1 + 2^{\frac{L^2(L+10)k^{L+2}}{(k-1)^2}} M_k^{\frac{(3L^2+1)Lk^{L+2}}{k-1}} \big(|\mathbf{x}| + |\mathbf{x}|^{Lk^{L+1}}\big)^L\Big)\big(|\mathbf{x}| + |\mathbf{x}|^{k^L}\big).$$

With the shorthand

$$\mathcal{A}_\beta := \big\{\Theta \in \mathcal{A} \colon \|\Omega\|_F^\beta \leq 1\big\} \quad \text{and} \quad \|\mathbf{x}\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\mathbf{x}_i|^2},$$

we extend our analysis to general ℓ_β-regularization. This generalization necessitates non-trivial refinements of concentration inequalities as well as the decomposition of approximation and estimation errors under general convex penalties. The following prediction guarantee for the estimator in (2.2) is then established, and its proof is provided in Appendix A.2.

Theorem 3.1. *Assume Assumptions 3.1 and 3.2 hold. Suppose that*

$$\lambda \geq a M_k^L \sqrt{\frac{1}{n}\sum_{i=1}^{n} (B(\mathbf{x}_i))^2} \sqrt{L \log(2P)} \frac{\log(2n)}{\sqrt{n}},$$

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

where $a \in (0, \infty)$ is a constant that depends only on the noise parameters K_1 and γ_1. Then, for n large enough,

$$\mathrm{err}^2(\hat{\kappa}_h g_{\hat{\Omega}_h}) \leq \inf_{\substack{\kappa \in [0,\infty) \\ \Omega \in \mathcal{A}_1}} \{\mathrm{err}^2(\kappa g_\Omega) + 2\lambda\kappa\},$$

with probability at least $1 - \frac{1}{n}$.

Remark 3.1. From the locally Lipschitz-continuous condition, we know the activation function has polynomial growth. As a result, the complexity of neural networks is greatly increased, which makes it very difficult to derive the continuity and boundary properties of the neural networks in $\mathcal{A}_\beta$. To address this issue, we construct some iterative sequences during the proof of Theorem 3.1, then utilize some common techniques for handling locally Lipschitz-continuous functions. Finally, we obtain the Lipschitz-continuous property of the neural networks, and thus Theorem 3.1 is derived. However, when the input $|\mathbf{x}| > 1$, the polynomial growth of the activation function leads to the bound obtained in Theorem 3.1 being exponentially increasing with respect to the number of layers L, and it seems difficult to improve this result by using the iterative method in this paper.

As an illustration, we can simplify Theorem 3.1 in a parametric setting.

Corollary 3.1. *Assume that Assumptions 3.1 and 3.2 hold. Suppose that*

$$\lambda = aM_k^L \sqrt{\frac{1}{n}\sum_{i=1}^{n} (B(\mathbf{x}_i))^2} \sqrt{L \log(2P)} \frac{\log(2n)}{\sqrt{n}},$$

and there exist parameters $(\kappa_*, \Omega_*) \in [0, \infty) \times \mathcal{A}_1$ such that $\kappa_* g_{\Omega_*}(\mathbf{x}_i) = g_*(\mathbf{x}_i)$ for all $i \in \{1, \ldots, n\}$. Then, for n large enough,

$$\mathrm{err}^2(\hat{\kappa}_h g_{\hat{\Omega}_h}) \leq 2a\kappa_* M_k^L \sqrt{\frac{1}{n}\sum_{i=1}^{n} (B(\mathbf{x}_i))^2} \sqrt{L \log(2P)} \frac{\log(2n)}{\sqrt{n}}$$

with probability at least $1 - 1/(2n)$.

Remark 3.2. For the k-degree neural networks with l_β-regularization, Theorem 3.1 and Corollary 3.1 establish essentially an $n^{-1/2}$-decrease of the error in the sample size n; this conclusion validates the conjecture in [38, Corollary 4], and it is further verified that the method of scale regularization is still applicable to the case of k-degree neural networks.

Moreover, based on Theorem 3.1 and Corollary 3.1 above, and similar to the proof of [38, Theorem 5], we can immediately obtain the following bound for the prediction risk.

Theorem 3.2. *Keep the same assumptions and notation as in Corollary 3.1. Suppose that the inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. random vectors. Then, for n large enough,*

$$\mathrm{risk}(\hat{\kappa}_h g_{\hat{\Omega}_h}) \leq 1.01\mathrm{risk}(\kappa_* g_{\Omega_*}) + a\kappa_* M_k^L \sqrt{\frac{1}{n}\sum_{i=1}^{n} (B(\mathbf{x}_i))^2} \sqrt{L \log(2P)} \frac{\log(2P)}{\sqrt{n}}$$

$$+ a(\kappa_*)^2 M_k^{2L} \sqrt{L^2 \log(2p) \sum_{i=1}^{n} (B(\mathbf{x}_i))^4} \frac{\log(2n)}{n},$$

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

with probability at least $1 - 1/n$*, where* $a \in (0, \infty)$ *is a constant that depends only on the sub-Gaussian parameters* K_1 *and* γ_1 *of the noise.*

Theorem 3.2 proves the fact that for properly chosen tuning parameter λ and large enough sample size n, any reasonable optimal point of (2.2) generalizes essentially as well as (κ_*, Ω_*). Our result generalizes [38, Theorem 5] in the context of k-degree neural networks, and is applicable to the global optimum of the objective function.

4. Discussion

This paper establishes statistical guarantees for k-degree neural networks via scale regularization. Our theoretical results show that the prediction error decrease is of the order of $n^{-1/2}$ in the sample size n. We also illustrate an explicit bound in the case of l_β-regularization, locally Lipschitz activation functions, and sub-Gaussian noise. Indeed, our new statistical guarantees might be useful in deriving bounds for other regularizers [11, 26, 28, 42], activation functions [8, 21, 25, 30], and noise distributions [4, 16, 32, 41] with different rates. Theoretical properties still merit further research within the framework of neural networks.

Appendix A. Proofs in Sections 2 and 3

A.1. Proof of Proposition 2.1

We first prove the sufficiency of the proposition. For any fixed $\Theta = (W^L, \dots, W^0) \in \mathcal{A}$, we assume that $W^l = \mathbf{0}_{p_{l+1} \times p_l}$ for an $l \in \{0, \dots, L\}$. By (1.2) and the fact that the activation functions are non-negative homogeneous, we have

$$\begin{aligned} g_\Theta(\mathbf{x}) &= W^L \mathbf{f}^L(\cdots \mathbf{0}_{p_{l+1} \times p_l} \mathbf{f}^l(\cdots W^1 \mathbf{f}^1(W^0 \mathbf{x}))) \\ &= W^L \mathbf{f}^L(\cdots 0 \cdot \mathbf{0}_{p_{l+1} \times p_l} \mathbf{f}^l(\cdots W^1 \mathbf{f}^1(W^0 \mathbf{x}))) \\ &= 0 \quad \text{for all } \mathbf{x} \in \mathbb{R}^d. \end{aligned}$$

Therefore, when $W^l = \mathbf{0}_{p_{l+1} \times p_l}$ for $l \in \{0, \dots, L\}$, let $\kappa := 0$, and the equality $\kappa g_\Omega = g_\Theta$ holds for all $\Omega \in \mathcal{A}_h$.

When $W^l \neq \mathbf{0}_{p_{l+1} \times p_l}$ for all $l \in \{0, \dots, L\}$, denote

$$\kappa := (h(\Theta))^{(k^L + k^{L-1} \cdots + 1)/\alpha}.$$

Notice that h is positive definite; we have $h(\Theta) \in (0, \infty)$, which implies $\kappa \in (0, \infty)$. Let

$$\Omega := \Theta / \kappa^{1/(k^L + k^{L-1} \cdots + 1)} = \left(W^L / \kappa^{1/(k^L + k^{L-1} \cdots + 1)}, \dots, W^0 / \kappa^{1/(k^L + k^{L-1} \cdots + 1)}\right).$$

Since h is non-negative homogeneous of degree α, it is straightforward to calculate that

$$h(\Omega) = h\left(\Theta / \kappa^{1/(k^L + k^{L-1} \cdots + 1)}\right) = \kappa^{-\alpha/(k^L + k^{L-1} \cdots + 1)} h(\Theta) = (h(\Theta))^{-1} h(\Theta) = 1,$$

that is, $\Omega \in \mathcal{A}_h$. Now we only need to prove the equality (2.1). Since $\mathbf{f}^1, \dots, \mathbf{f}^L$ are non-negative homogeneous of degree k, for any $\mathbf{x} \in \mathbb{R}^d$, one can derive that

$$\begin{aligned} \kappa g_\Omega(\mathbf{x}) &= \kappa \frac{W^L}{\kappa^{1/(k^L + k^{L-1} \cdots + 1)}} \mathbf{f}^L\left(\cdots \frac{W^1}{\kappa^{1/(k^L + k^{L-1} \cdots + 1)}} \mathbf{f}^1\left(\frac{W^0}{\kappa^{1/(k^L + k^{L-1} \cdots + 1)}} \mathbf{x}\right)\right) \\ &= \kappa \frac{W^L}{\kappa^{1/(k^L + k^{L-1} \cdots + 1)}} \mathbf{f}^L\left(\cdots \frac{W^1}{\kappa^{k+1/(k^L + k^{L-1} \cdots + 1)}} \mathbf{f}^1\left(W^0 \mathbf{x}\right)\right) \end{aligned}$$

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

$$
\begin{aligned}
&= \cdots \\
&= \frac{\kappa}{\kappa^{(k^L+k^{L-1}\cdots+1)/(k^L+k^{L-1}\cdots+1)}} W^L \mathbf{f}^L(\cdots W^1 \mathbf{f}^1(W^0\mathbf{x})) \\
&= W^L \mathbf{f}^L(\cdots W^1 \mathbf{f}^1(W^0\mathbf{x})) \\
&= g_\Theta(\mathbf{x}),
\end{aligned}
$$

and thus (2.1) is proved.

Inversely, for a fixed $\kappa \in [0, \infty)$ and a $\Omega = (U^L, \ldots, U^0) \in \mathcal{A}_h$, let

$$
\Theta := \kappa^{1/(k^L+k^{L-1}\cdots+1)}\Omega = \big(\kappa^{1/(k^L+k^{L-1}\cdots+1)}U^L, \ldots, \kappa^{1/(k^L+k^{L-1}\cdots+1)}U^0\big).
$$

Then, for any $\mathbf{x} \in \mathbb{R}^d$, one can derive from the non-negative homogeneity of degree k of the activation functions that

$$
\begin{aligned}
g_\Theta(\mathbf{x}) &= \kappa^{1/(k^L+k^{L-1}\cdots+1)} U^L \mathbf{f}^L\big(\cdots \kappa^{1/(k^L+k^{L-1}\cdots+1)} U^1 \mathbf{f}^1\big(\kappa^{1/(k^L+k^{L-1}\cdots+1)} U^0\mathbf{x}\big)\big) \\
&= \kappa^{1/(k^L+k^{L-1}\cdots+1)} U^L \mathbf{f}^L\big(\cdots \kappa^{(k+1)/(k^L+k^{L-1}\cdots+1)} U^1 \mathbf{f}^1(U^0\mathbf{x})\big) \\
&= \cdots \\
&= \kappa^{(k^L+k^{L-1}\cdots+1)/(k^L+k^{L-1}\cdots+1)} U^L \mathbf{f}^L(\cdots U^1 \mathbf{f}^1(U^0\mathbf{x})) \\
&= \kappa\, g_\Omega(\mathbf{x}),
\end{aligned}
$$

and the desired result follows.

A.2. Proof of Theorem 3.1

We first introduce some notations. Define the Frobenius norms of the parameter and weight matrices by

$$
\|\Theta\|_F := \sqrt{\sum_{l=0}^{L} \|W^l\|_F^2} \quad \text{and} \quad \|W^l\|_F := \sqrt{\sum_{k=1}^{p_{l+1}} \sum_{j=1}^{p_l} \big(W^l_{kj}\big)^2}.
$$

For any networks g_Θ and g_Γ with $\Theta, \Gamma \in \mathcal{A}$, the prediction distance is defined as

$$
\|g_\Theta - g_\Gamma\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n} (g_\Theta(\mathbf{x}_i) - g_\Gamma(\mathbf{x}_i))^2},
$$

and similarly

$$
\|g_\Theta\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^{n} (g_\Theta(\mathbf{x}_i))^2}.
$$

We divide the networks into inner and outer subnetworks and then obtain the locally Lipschitz property of neural networks on $\mathcal{A}$.

The inner subnetworks of a network $g_\Theta \in \mathcal{G} := \{g_\Theta \colon \Theta \in \mathcal{A}\}$ are vector-valued functions defined by

$$
S_0 g_\Theta(\mathbf{x}) := \mathbf{x} \quad \text{for all } \mathbf{x} \in \mathbb{R}^d,
$$

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

and

$$S_l g_\Theta(\mathbf{x}) := \mathbf{f}^l(W^{l-1}\mathbf{f}^{l-1}(\cdots W^1\mathbf{f}^1(W^0\mathbf{x}))) \in \mathbb{R}^{p_l} \quad \text{for all } \mathbf{x} \in \mathbb{R}^d,$$

for $l \in \{1, \ldots, L\}$. Similarly, the outer subnetworks of g_Θ are real-valued functions defined by

$$S^l g_\Theta(\mathbf{z}) := W^L\mathbf{f}^L(\cdots W^l\mathbf{f}^l(W^{l-1}\mathbf{z})) \in \mathbb{R} \quad \text{for all } \mathbf{z} \in \mathbb{R}^{p_{l-1}},$$

for $l \in \{1, \ldots, L\}$, and

$$S^{L+1} g_\Theta(\mathbf{z}) := W^L\mathbf{z} \in \mathbb{R} \quad \text{for all } \mathbf{z} \in \mathbb{R}^{p_L}.$$

The initial network can be split into an inner and outer network along every layer $l \in \{1, \ldots, L+1\}$:

$$g_\Theta(\mathbf{x}) = S^l g_\Theta(S_{l-1} g_\Theta(\mathbf{x})).$$

Since the locally Lipschitz property of neural networks is much more complicated than the global Lipschitz ($k = 1$) case, we need to define some iterative sequences.

For any $r \geq 0$ and $\Theta = (W^L, \ldots, W^0) \in \mathcal{A}$, let $c_0(r, \Theta) = r + \|W^0\|_F^{k-1} r^k$ and

$$c_l(r, \Theta) = c_{l-1}(r, \Theta) + M_k^{l(k-1)} \prod_{j=0}^{l} \|W^j\|_F^{k-1} c_{l-1}(r, \Theta)^k \quad \text{for } 1 \leq l \leq L.$$

Based on the above sequence, for any fixed $l \in \{1, \ldots, L\}$, we define

$$\tilde{c}_j(r, \Theta) = 0 \quad \text{for } 0 \leq j < l,$$

$$\tilde{c}_l(r, \Theta) = r + \|W^l\|_F^{k-1} r^k,$$

$$\tilde{c}_j(r, \Theta) = \tilde{c}_{j-1}(r, \Theta) + M_k^{(j-l)(k-1)} \prod_{q=l}^{j} \|W^q\|_F^{k-1} \tilde{c}_{j-1}^k(r, \Theta) \quad \text{for } l+1 \leq j \leq L.$$

In addition, for any $\Gamma = (V^L, \ldots, V^0) \in \mathcal{A}$, we define

$$d_l(r, \Theta, \Gamma) := M_k^l \prod_{j=0}^{l-2} \|V^j\|_F \|W^{l-1}\|_F c_{l-2}(r, \Gamma) + M_k^{1+(l-1)k} \prod_{j=0}^{l-2} \|V^j\|_F^k \|W^{l-1}\|_F^k c_{l-2}^k(r, \Gamma),$$

where we set (note the unusual indexes of the product) $\prod_{j=0}^{-1} b_j := 1$ for any sequence b_j.

According to the above definition, we first show an upper bound for the inner subnetworks and a Lipschitz property for the outer subnetworks.

Lemma A.1. *Suppose that the activation functions $\mathbf{f}^l : \mathbb{R}^{p_l} \to \mathbb{R}^{p_l}$ satisfy Assumption 3.1. Then it holds for every $\mathbf{x} \in \mathbb{R}^d$ and $\Theta = (W^L, \ldots, W^0)$ that*

$$|S_{l-1} g_\Theta(\mathbf{x})| \leq \left(M_k^{l-1} \prod_{j=0}^{l-2} \|W^j\|_F c_{l-2}(|\mathbf{x}|, \Theta) \right) \vee |\mathbf{x}|,$$

where $a \vee b = \max\{a, b\}$ for any $a, b \in \mathbb{R}$.

Proof. Using the assumption that $\mathbf{z} \mapsto \mathbf{f}^{l-1}(\mathbf{z})$ is locally Lipschitz, we get for every $\Theta = (W^L, \ldots, W^0)$ and $\mathbf{x} \in \mathbb{R}^d$ that

$$\begin{aligned} |S_{l-1} g_\Theta(\mathbf{x})| &= |\mathbf{f}^{l-1}\big(W^{l-2} S_{l-2} g_\Theta(\mathbf{x})\big)| \\ &\leq M_k\big(1 + |W^{l-2} S_{l-2} g_\Theta(\mathbf{x})|^{k-1}\big) |W^{l-2} S_{l-2} g_\Theta(\mathbf{x})| \end{aligned}$$

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

$$= M_k |W^{l-2} S_{l-2} g_\Theta(\mathbf{x})| + M_k |W^{l-2} S_{l-2} g_\Theta(\mathbf{x})|^k$$
$$\le M_k \|W^{l-2}\|_F |S_{l-2} g_\Theta(\mathbf{x})| + M_k \|W^{l-2}\|_F^k |S_{l-2} g_\Theta(\mathbf{x})|^k.$$

Then, when $l = 2$,

$$\begin{aligned} |S_1 g_\Theta(\mathbf{x})| &\le M_k \|W^0\|_F |S_0 g_\Theta(\mathbf{x})| + M_k \|W^0\|_F^k |S_0 g_\Theta(\mathbf{x})|^k \\ &\le M_k \|W^0\|_F |\mathbf{x}| + M_k \|W^0\|_F^k |\mathbf{x}|^k \\ &= M_k \|W^0\|_F \big(|\mathbf{x}| + \|W^0\|_F^{k-1} |\mathbf{x}|^k\big) \\ &= M_k \|W^0\|_F c_0(|\mathbf{x}|, \Theta), \end{aligned}$$

when $l = 3$,

$$\begin{aligned} |S_2 g_\Theta(\mathbf{x})| &\le M_k \|W^1\|_F |S_1 g_\Theta(\mathbf{x})| + M_k \|W^1\|_F^k |S_1 g_\Theta(\mathbf{x})|^k \\ &\le M_k \|W^1\|_F M_k \|W^0\|_F c_0(|\mathbf{x}|, \Theta) + M_k \|W^1\|_F^k M_k^k c_0(|\mathbf{x}|, \Theta)^k \|W^0\|_F^k \\ &= M_k^2 \|W^1\|_F \|W^0\|_F \big(c_0(|\mathbf{x}|, \Theta) + M_k^{k-1} \|W^1\|_F^{k-1} \|W^0\|_F^{k-1} c_0(|\mathbf{x}|, \Theta)^k\big) \\ &= M_k^2 \|W^1\|_F \|W^0\|_F c_1(|\mathbf{x}|, \Theta). \end{aligned}$$

Inductively, for any $l \ge 2$,

$$|S_{l-1} g_\Theta(\mathbf{x})| \le M_k^{l-1} \prod_{j=0}^{l-2} \|W^j\|_F c_{l-2}(|\mathbf{x}|, \Theta).$$

In addition, notice that $|S_0 g_\Theta(\mathbf{x})| = |\mathbf{x}|$, so one can verify readily that $\mathbf{x} \mapsto S_{l-1} g_\Theta(\mathbf{x})$ and $\mathbf{x} \mapsto S_0 g_\Theta(\mathbf{x})$ are bounded by

$$M_k^{l-1} \prod_{j=0}^{l-2} \|W^j\|_F c_{l-2}(|\mathbf{x}|, \Theta) \quad \text{and} \quad |\mathbf{x}|,$$

respectively. □

Lemma A.2. *Assume that the activation functions* $\mathbf{f}^l \colon \mathbb{R}^{p_l} \to \mathbb{R}^{p_l}$ *satisfy Assumption 3.1. Then it holds for every* $\mathbf{x} \in \mathbb{R}^d$ *and* $\Theta = (W^L, \dots, W^0)$ *that*

$$\begin{aligned} &|S^{l+1} g_\Theta(\mathbf{z}_1) - S^{l+1} g_\Theta(\mathbf{z}_2)| \\ &\le M_k^{L-l} |\mathbf{z}_1 - \mathbf{z}_2| \big(1 + \|W^l\|_F^{k-1} |\mathbf{z}_1|^{k-1} + \|W^l\|_F^{k-1} |\mathbf{z}_2|^{k-1}\big) \|W^{l+1}\|_F \|W^l\|_F \\ &\quad \times \prod_{j=l+1}^{L-1} \left(1 + M_k^{j-l} \prod_{q=l}^{j} \|W^q\|_F (\tilde{c}_{j-1}(|\mathbf{z}_1|, \Theta) + \tilde{c}_{j-1}(|\mathbf{z}_2|, \Theta))\right) \|W^{j+1}\|_F. \end{aligned} \tag{A.1}$$

Proof. For every $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{p_l}$ and $l \in \{0, \dots, L\}$, we have

$$\begin{aligned} &|S^{l+1} g_\Theta(\mathbf{z}_1) - S^{l+1} g_\Theta(\mathbf{z}_2)| \\ &= \big|W^L \mathbf{f}^L\big(\cdots W^{l+1} \mathbf{f}^{l+1}\big(W^l \mathbf{z}_1\big)\big) - W^L \mathbf{f}^L\big(\cdots W^{l+1} \mathbf{f}^{l+1}\big(W^l \mathbf{z}_2\big)\big)\big| \\ &\le \|W^L\|_F \big|\mathbf{f}^L\big(\cdots W^{l+1} \mathbf{f}^{l+1}\big(W^l \mathbf{z}_1\big)\big) - \mathbf{f}^L\big(\cdots W^{l+1} \mathbf{f}^{l+1}\big(W^l \mathbf{z}_2\big)\big)\big|. \end{aligned}$$

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

Then Assumption 3.1 implies that

$$
\begin{aligned}
&|S^{l+1}g_\Theta(\mathbf{z}_1) - S^{l+1}g_\Theta(\mathbf{z}_2)| \\
&\quad \le \|W^L\|_F M_k \big|W^{L-1}\mathbf{f}^{L-1}\big(\cdots W^{l+1}\mathbf{f}^{l+1}\big(W^l\mathbf{z}_1\big)\big) - W^{L-1}\mathbf{f}^{L-1}\big(\cdots W^{l+1}\mathbf{f}^{l+1}\big(W^l\mathbf{z}_2\big)\big)\big| \\
&\qquad \times \big(1 + \big|W^{L-1}\mathbf{f}^{L-1}\big(\cdots W^{l+1}\mathbf{f}^{l+1}\big(W^l\mathbf{z}_1\big)\big)\big|^{k-1} + \big|W^{L-1}\mathbf{f}^{L-1}\big(\cdots W^{l+1}\mathbf{f}^{l+1}\big(W^l\mathbf{z}_2\big)\big)\big|^{k-1}\big).
\end{aligned}
$$

Hence it can be inferred by induction that

$$
\begin{aligned}
&|S^{l+1}g_\Theta(\mathbf{z}_1) - S^{l+1}g_\Theta(\mathbf{z}_2)| \\
&\quad \le M_k^{L-l}|\mathbf{z}_1 - \mathbf{z}_2|\big(1 + |W^l\mathbf{z}_1|^{k-1} + |W^l\mathbf{z}_2|^{k-1}\big)\|W^{l+1}\|_F\|W^l\|_F \\
&\qquad \times \prod_{j=l+1}^{L-1}\big(1 + \big|W^j\mathbf{f}^j\big(\cdots W^{l+1}\mathbf{f}^{l+1}\big(W^l\mathbf{z}_1\big)\big)\big|^{k-1} + \big|W^j\mathbf{f}^j\big(\cdots W^{l+1}\mathbf{f}^{l+1}\big(W^l\mathbf{z}_2\big)\big)\big|^{k-1}\big)\|W^{j+1}\|_F.
\end{aligned}
$$

Then we need to bound $|\mathbf{f}^j(\cdots W^{l+1}\mathbf{f}^{l+1}(W^l\mathbf{z}))|$ for any $l+1 \le j \le L-1$. Similar to the proof of Lemma A.1, we have

$$
|\mathbf{f}^j(\cdots W^{l+1}\mathbf{f}^{l+1}(W^l\mathbf{z}))| \le M_k^{j-l}\prod_{q=l}^{j-1}\|W^q\|_F\tilde{c}_{j-1}(|\mathbf{z}|, \Theta).
$$

Hence we obtain

$$
\big|W^j\mathbf{f}^j(\cdots W^{l+1}\mathbf{f}^{l+1}(W^l\mathbf{z}))\big| \le M_k^{j-l}\prod_{q=l}^{j}\|W^q\|_F\tilde{c}_{j-1}(|\mathbf{z}|, \Theta) \quad \text{for any } l+1 \le j \le L-1,
$$

and the desired result follows. □

Now we define

$$
\begin{aligned}
&B_1(r, \Theta, \Gamma) \\
&= \max_{l\in\{1,\dots,L+1\}}\Bigg\{\Bigg(1 + \|W^l\|_F^{k-1}d_l(r, \Theta, \Gamma)^{k-1} + \|W^l\|_F^{k-1}\prod_{j=0}^{l-1}\|V^j\|_F^{k-1}c_{l-1}^{k-1}(r, \Gamma)\Bigg)c_{l-2}(r, \Gamma) \\
&\times \Bigg(1 + \big(\|W^{l-1}\|_F^{k-1} + \|V^{l-1}\|_F^{k-1}\big)M_k^{(l-1)(k-1)}\prod_{q=0}^{l-2}\|V^q\|_F^{k-1}c_{l-2}^{k-1}(r, \Gamma)\Bigg) \\
&\times \prod_{\substack{j=0\\ j\ne l-1}}^{L}\Bigg(1 + M_k^{j-l}\prod_{q=l}^{j}\|W^q\|_F\Bigg(\tilde{c}_{j-1}(d_l(r, \Theta, \Gamma), \Theta) + \tilde{c}_{j-1}\Bigg(M_k^l\prod_{j=0}^{l-1}\|V^j\|_F c_{l-1}(r, \Gamma), \Theta\Bigg)\Bigg)\Bigg) \\
&\times (\|W^j\|_F \vee \|V^j\|_F)\Bigg\},
\end{aligned}
\tag{A.2}
$$

and give the following locally Lipschitz property of neural networks.

Lemma A.3. *Assume that the activation functions* $\mathbf{f}^l\colon \mathbb{R}^{p_l} \to \mathbb{R}^{p_l}$ *satisfy Assumption* 3.1. *Then it holds for every* $\mathbf{x} \in \mathbb{R}^d$ *and* $\Theta = (W^L, \dots, W^0)$, $\Gamma = (V^L, \dots, V^0) \in \mathcal{A}$ *that*

$$
|g_\Theta(\mathbf{x}) - g_\Gamma(\mathbf{x})| \le 2M_k^L\sqrt{L}B_1(|\mathbf{x}|, \Theta, \Gamma)\|\Theta - \Gamma\|_F.
$$

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

Similarly, it holds that

$$\|g_\Theta - g_\Gamma\|_n \le 2M_k^L\sqrt{L}\sqrt{\frac{1}{n}\sum_{i=1}^n (B_1(|\mathbf{x}_i|, \Theta, \Gamma))^2}\|\Theta - \Gamma\|_F.$$

Proof. We consider two networks g_Θ and g_Γ with parameters $\Theta = (W^L, \dots, W^0) \in \mathcal{A}$ and $\Gamma = (V^L, \dots, V^0) \in \mathcal{A}$, respectively. Then, similar to the proof of Proposition 6 in [38], we have

$$\begin{aligned}|g_\Theta(\mathbf{x}) - g_\Gamma(\mathbf{x})| \le \sum_{l=1}^L &\left|\left(S^{l+1}g_\Theta\left(\mathbf{f}^l\left(W^{l-1}S_{l-1}g_\Gamma(\mathbf{x})\right) - S^{l+1}g_\Theta\left(\mathbf{f}^l\left(V^{l-1}S_{l-1}g_\Gamma(\mathbf{x})\right)\right)\right)\right| \\ &+ |(W^L - V^L)S_L g_\Gamma(\mathbf{x})|.\end{aligned}$$

Then (3.1) and (A.1) imply

$$\begin{aligned}&|g_\Theta(\mathbf{x}) - g_\Gamma(\mathbf{x})| \\ &\le \sum_{l=1}^L M_k^{L-l}\left(1 + \|W^l\|_F^{k-1}\underbrace{|\mathbf{f}^l(W^{l-1}S_{l-1}g_\Gamma(\mathbf{x}))|^{k-1}}_{①} + \|W^l\|_F^{k-1}\underbrace{|\mathbf{f}^l(V^{l-1}S_{l-1}g_\Gamma(\mathbf{x}))|^{k-1}}_{②}\right) \\ &\times \prod_{j=l+1}^{L-1}\left(1 + M_k^{j-l}\prod_{q=l}^j \|W^q\|_F\left(\underbrace{\tilde{c}_{j-1}(|\mathbf{f}^l(W^{l-1}S_{l-1}g_\Gamma(\mathbf{x}))|, \Theta)}_{③} + \underbrace{\tilde{c}_{j-1}(|\mathbf{f}^l(V^{l-1}S_{l-1}g_\Gamma(\mathbf{x}))|, \Gamma)}_{④}\right)\right) \\ &\times \|W^{j+1}\|_F\|W^{l+1}\|_F\|W^l\|_F \\ &\times M_k\left(1 + \underbrace{|W^{l-1}S_{l-1}g_\Gamma(\mathbf{x})|^{k-1}}_{⑤} + \underbrace{|V^{l-1}S_{l-1}g_\Gamma(\mathbf{x})|^{k-1}}_{⑥}\right)\|W^{l-1} - V^{l-1}\|_F\underbrace{|S_{l-1}g_\Gamma(\mathbf{x})|}_{⑦} \\ &+ \|(W^L - V^L)\|\underbrace{|S_L g_\Gamma(\mathbf{x})|}_{⑧}.\end{aligned}$$

Then we need to bound the above eight terms. For terms ① and ②, similar to the proof of Lemma A.1, one can write, by (3.1),

$$\begin{aligned}&|\mathbf{f}^l(W^{l-1}S_{l-1}g_\Gamma(\mathbf{x}))| \\ &\quad= |\mathbf{f}^l(W^{l-1}\mathbf{f}^{l-1}(V^{l-2}\mathbf{f}^{l-2}(\cdots V^1\mathbf{f}^1(V^0\mathbf{x}))))| \\ &\quad\le M_k^l\prod_{j=0}^{l-2}\|V^j\|_F\|W^{l-1}\|_F c_{l-2}(|\mathbf{x}|, \Gamma) + M_k^{1+(l-1)k}\prod_{j=0}^{l-2}\|V^j\|_F^k\|W^{l-1}\|_F^k c_{l-2}^k(|\mathbf{x}|, \Gamma) \\ &\quad= d_l(|\mathbf{x}|, \Theta, \Gamma)\end{aligned}$$

and

$$|\mathbf{f}^l(V^{l-1}S_{l-1}g_\Gamma(\mathbf{x}))| = |S_l g_\Gamma(\mathbf{x})| \le M_k^l\prod_{j=0}^{l-1}\|V^j\|_F c_{l-1}(|\mathbf{x}|, \Gamma).$$

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

For terms ③ and ④, according to the definition of $\tilde{c}_j(|\mathbf{z}|, \Theta)$, we know the function is increasing in $|\mathbf{z}|$, so we only need to bound $|\mathbf{f}^l(W^{l-1}S_{l-1}g_\Gamma(\mathbf{x}))|$. One can derive from terms ① and ② that

$$\tilde{c}_{j-1}(|\mathbf{f}^l(W^{l-1}S_{l-1}g_\Gamma(\mathbf{x}))|, \Theta) \leq \tilde{c}_{j-1}(d_l(|\mathbf{x}|, \Theta, \Gamma), \Theta)$$

and

$$\tilde{c}_{j-1}(|\mathbf{f}^l(V^{l-1}S_{l-1}g_\Gamma(\mathbf{x}))|, \Theta) \leq \tilde{c}_{j-1}\left(M_k^l \prod_{j=0}^{l-1} \|V^j\|_F c_{l-1}(|\mathbf{x}|, \Gamma), \Theta\right).$$

For the last four terms, it is easy to imply from Lemmas A.1 and A.2 that

$$|W^{l-1}S_{l-1}g_\Gamma(\mathbf{x})| \leq \|W^{l-1}\|_F |S_{l-1}g_\Gamma(\mathbf{x})| \leq \|W^{l-1}\|_F M_k^{l-1} \prod_{j=0}^{l-2} \|V^j\|_F c_{l-2}(|\mathbf{x}|, \Gamma),$$

$$|V^{l-1}S_{l-1}g_\Gamma(\mathbf{x})| \leq M_k^{l-1} \prod_{j=0}^{l-1} \|V^j\|_F c_{l-2}(|\mathbf{x}|, \Gamma),$$

$$|S_{l-1}g_\Gamma(\mathbf{x})| \leq M_k^{l-1} \prod_{j=0}^{l-2} \|V^j\|_F c_{l-2}(|\mathbf{x}|, \Gamma),$$

$$|S_L g_\Gamma(\mathbf{x})| \leq M_k^L \prod_{j=0}^{L-1} \|V^j\|_F c_{L-1}(|\mathbf{x}|, \Gamma).$$

Therefore we have

$$\begin{aligned}
&|g_\Theta(\mathbf{x}) - g_\Gamma(\mathbf{x})| \\
&\leq M_k^L \sum_{l=1}^{L} \left(1 + \|W^l\|_F^{k-1} d_l(|\mathbf{x}|, \Theta, \Gamma)^{k-1} + \|W^l\|_F^{k-1} M_k^{l(k-1)} \prod_{j=0}^{l-1} \|V^j\|_F^{k-1} c_{l-1}^{k-1}(|\mathbf{x}|, \Gamma)\right) \|W^{l+1}\|_F \\
&\times \prod_{j=l+1}^{L-1} \left(1 + M_k^{j-l} \prod_{q=l}^{j} \|W^q\|_F \left(\tilde{c}_{j-1}(d_l(|\mathbf{x}|, \Theta, \Gamma), \Theta) + \tilde{c}_{j-1}\left(M_k^l \prod_{j=0}^{l-1} \|V^j\| c_{l-1}(|\mathbf{x}|, \Gamma), \Theta\right)\right)\right) \\
&\times \|W^{j+1}\|_F \|W^l\|_F \left(1 + (\|W^{l-1}\|_F^{k-1} + \|V^{l-1}\|_F^{k-1}) M_k^{(l-1)(k-1)} \prod_{q=0}^{l-2} \|V^q\|_F^{k-1} c_{l-2}^{k-1}(|\mathbf{x}|, \Gamma)\right) \\
&\times \|W^{l-1} - V^{l-1}\|_F \prod_{j=0}^{l-2} \|V^j\|_F c_{l-2}(|\mathbf{x}|, \Gamma) + \|(W^L - V^L)\|_F M_k^L \prod_{j=0}^{L-1} \|V^j\|_F c_{L-1}(|\mathbf{x}|, \Gamma).
\end{aligned}$$

Then we obtain

$$\begin{aligned}
&|g_\Theta(\mathbf{x}) - g_\Gamma(\mathbf{x})| \\
&\leq M_k^L \sum_{l=1}^{L} \left(1 + \|W^l\|_F^{k-1} d_l(|\mathbf{x}|, \Theta, \Gamma)^{k-1} + \|W^l\|_F^{k-1} \prod_{j=0}^{l-1} \|V^j\|_F^{k-1} c_{l-1}^{k-1}(|\mathbf{x}|, \Gamma)\right) c_{l-2}(|\mathbf{x}|, \Gamma) \\
&\quad \times \left(1 + (\|W^{l-1}\|_F^{k-1} + \|V^{l-1}\|_F^{k-1}) M_k^{(l-1)(k-1)} \prod_{q=0}^{l-2} \|V^q\|_F^{k-1} c_{l-2}^{k-1}(|\mathbf{x}|, \Gamma)\right)
\end{aligned}$$

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

$$\times \prod_{\substack{j=0 \\ j\neq l-1}}^{L}\left(1+M_k^{j-l}\prod_{q=l}^{j}\|W^q\|_F\left(\tilde{c}_{j-1}(d_l(|\mathbf{x}|,\Theta,\Gamma),\Theta)+\tilde{c}_{j-1}\left(M_k^l\prod_{j=0}^{l-1}\|V^j\|_F c_{l-1}(|\mathbf{x}|,\Gamma),\Theta\right)\right)\right)$$

$$\times(\|W^j\|_F\vee\|V^j\|_F)\times\|W^{l-1}-V^{l-1}\|_F+M_k^L c_{L-1}(|\mathbf{x}|,\Gamma)\prod_{j=0}^{L-1}(\|W^j\|_F\vee\|V^j\|_F)\|W^L-V^L\|_F.$$

Furthermore, recall (A.2).Then, consolidating and rearranging yields

$$\begin{aligned}
&|g_\Theta(\mathbf{x})-g_\Gamma(\mathbf{x})|\\
&\leq M_k^L\sum_{l=1}^{L+1}\left(1+\|W^l\|_F^{k-1}d_l(|\mathbf{x}|,\Theta,\Gamma)^{k-1}+\|W^l\|_F^{k-1}\prod_{j=0}^{l-1}\|V^j\|_F^{k-1}c_{l-1}^{k-1}(|\mathbf{x}|,\Gamma)\right)c_{l-2}(|\mathbf{x}|,\Gamma)\\
&\times\left(1+\left(\|W^{l-1}\|_F^{k-1}+\|V^{l-1}\|_F^{k-1}\right)M_k^{(l-1)(k-1)}\prod_{q=0}^{l-2}\|V^q\|_F^{k-1}c_{l-2}^{k-1}(|\mathbf{x}|,\Gamma)\right)\\
&\times\prod_{\substack{j=0 \\ j\neq l-1}}^{L}\left(1+M_k^{j-l}\prod_{q=l}^{j}\|W^q\|_F\left(\tilde{c}_{j-1}(d_l(|\mathbf{x}|,\Theta,\Gamma),\Theta)+\tilde{c}_{j-1}\left(M_k^l\prod_{j=0}^{l-1}\|V^j\|_F c_{l-1}(|\mathbf{x}|,\Gamma),\Theta\right)\right)\right)\\
&\times(\|W^j\|_F\vee\|V^j\|_F)\times\|W^{l-1}-V^{l-1}\|_F\\
&\leq M_k^L B_1(|\mathbf{x}|,\Theta,\Gamma)\sum_{m=0}^{L}\|W^m-V^m\|_F.
\end{aligned}$$

For the last sum in the above bound, similar to the proof of Proposition 6 in [38], we obtain

$$\sum_{m=0}^{L}\|W^m-V^m\|_F\leq 2\sqrt{L}\|\Theta-\Gamma\|_2\leq 2\sqrt{L}\|\Theta-\Gamma\|_F.$$

Combining all of the above, we have

$$|g_\Theta(\mathbf{x})-g_\Gamma(\mathbf{x})|\leq 2\sqrt{L}M_k^L B_1(|\mathbf{x}|,\Theta,\Gamma)\|\Theta-\Gamma\|_F,$$

and the first claim is proved. The second claim then follows readily:

$$\begin{aligned}
\|g_\Theta-g_\Gamma\|_n&=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(g_\Theta(\mathbf{x}_i)-g_\Gamma(\mathbf{x}_i))^2}\\
&\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}(2\sqrt{L}M_k^L B_1(|\mathbf{x}_i|,\Theta,\Gamma))^2}\|\Theta-\Gamma\|_F\\
&=2\sqrt{L}M_k^L\sqrt{\frac{1}{n}\sum_{i=1}^{n}(B_1(|\mathbf{x}_i|,\Theta,\Gamma))^2}\|\Theta-\Gamma\|_F.
\end{aligned}$$

□

Now we give the following Lipschitz property and boundary on $\mathcal{A}_\beta$.

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

Lemma A.4. *Assume that the activation functions* $\mathbf{f}^l\colon \mathbb{R}^{p_l} \to \mathbb{R}^{p_l}$ *satisfy Assumption* 3.1. *Then it holds for every* $\mathbf{x} \in \mathbb{R}^d$ *and* $\Theta = (W^L, \ldots, W^0)$, $\Gamma = (V^L, \ldots, V^0) \in \mathcal{A}_\beta$ *that*

$$\|g_\Theta - g_\Gamma\|_n \le 2M_k^L \sqrt{L} \sqrt{\frac{1}{n}\sum_{i=1}^n (B(|\mathbf{x}_i|))^2}\|\Theta - \Gamma\|_F$$

and

$$\|g_\Theta\|_n \le 2M_k^L \sqrt{L} \sqrt{\frac{1}{n}\sum_{i=1}^n (B(\mathbf{x}_i))^2}.$$

Proof. The proof follows from Lemma A.3 and we need to bound $B_1(|\mathbf{x}|, \Theta, \Gamma)$ in Lemma A.3 in the case $h(\Theta) \le 1$ and $h(\Gamma) \le 1$. Recalling (A.2), we have

$$\begin{aligned}
&B_1(|\mathbf{x}|, \Theta, \Gamma) \qquad\qquad (A.3)\\
&\le M_k^2 \max_{l\in\{1,\ldots,L+1\}} \Big\{\big(1 + d_l(|\mathbf{x}|, \Theta, \Gamma)^{k-1} + c_{l-1}(|\mathbf{x}|, \Gamma)^{k-1}\big)c_{l-2}(|\mathbf{x}|, \Gamma)\\
&\quad \times \big(1 + 2M_k^{(l-1)(k-1)} c_{l-2}(|\mathbf{x}|, \Gamma)^{k-1}\big)\Big\} \max_{l\in\{1,\ldots,L+1\}} \left\{\prod_{\substack{j=0\\ j\ne l-1}}^{L} \|W^j\|_F \vee \|V^j\|_F\right\}\\
&\quad \times \max_{l\in\{1,\ldots,L+1\}} \left\{\prod_{\substack{j=0\\ j\ne l-1}}^{L} \big(1 + M_k^{j-l}\big(\tilde{c}_{j-1}(d_l(|\mathbf{x}|, \Theta, \Gamma), \Theta) + \tilde{c}_{j-1}\big(M_k^l c_{l-1}(|\mathbf{x}|, \Gamma), \Theta\big)\big)\big)\right\}\\
&:= \mathcal{K}_1 + \mathcal{K}_2,
\end{aligned}$$

so we need to bound the three sequences $c_l(|\mathbf{x}|, \Gamma)$, $d_l(|\mathbf{x}|, \Theta, \Gamma)$ and $\tilde{c}_{j-1}(u, \Theta)$ appearing in the above inequality.

We first derive the upper bound of the sequence $c_l(|\mathbf{x}|, \Gamma)$. For any $l \ge 1$, by the inequality of the arithmetic and geometric means, we have

$$\begin{aligned}
c_l(|\mathbf{x}|, \Gamma) &= c_{l-1}(|\mathbf{x}|, \Gamma) + M_k^{l(k-1)} \prod_{j=0}^{l} \|V^j\|_F^{k-1} c_{l-1}(|\mathbf{x}|, \Gamma)^k\\
&\le c_{l-1}(|\mathbf{x}|, \Gamma) + M_k^{l(k-1)} \left(\frac{1}{l+1}\sum_{j=0}^{l} \|V^j\|_F^2\right)^{(l+1)(k-1)/2} c_{l-1}(|\mathbf{x}|, \Gamma)^k\\
&\le c_{l-1}(|\mathbf{x}|, \Gamma) + M_k^{l(k-1)} (l+1)^{-(l+1)(k-1)/2} c_{l-1}(|\mathbf{x}|, \Gamma)^k\\
&\le c_{l-1}(|\mathbf{x}|, \Gamma) + \left(\frac{M_k}{2}\right)^{l(k-1)} c_{l-1}(|\mathbf{x}|, \Gamma)^k.
\end{aligned}$$

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

Then

$$\begin{aligned}
c_0(|\mathbf{x}|, \Gamma) &= |\mathbf{x}| + \|V^0\|_F^{k-1}|\mathbf{x}|^k \le |\mathbf{x}| + |\mathbf{x}|^k,\\
c_1(|\mathbf{x}|, \Gamma) &= c_0(|\mathbf{x}|, \Gamma) + \left(\frac{M_k}{2}\right)^{k-1} c_0(|\mathbf{x}|, \Gamma)^k\\
&\le |\mathbf{x}| + |\mathbf{x}|^k + \left(\frac{M_k}{2}\right)^{k-1} (|\mathbf{x}| + |\mathbf{x}|^k)^k\\
&\le |\mathbf{x}| + |\mathbf{x}|^k + \left(\frac{M_k}{2}\right)^{k-1} 2^{k-1}\big(|\mathbf{x}| + |\mathbf{x}|^{k^2}\big)\\
&= |\mathbf{x}| + |\mathbf{x}|^k + {M_k}^{k-1}\big(|\mathbf{x}| + |\mathbf{x}|^{k^2}\big)\\
&\le 2\big(1 + {M_k}^{k-1}\big)\big(|\mathbf{x}| + |\mathbf{x}|^{k^2}\big).
\end{aligned}$$

By the method of induction, we have

$$c_l(|\mathbf{x}|, \Gamma) \le 2\big(1 + M_k^{k-1}\big)^{k^l + 2k^{l-1} + \cdots + lk + l + 1}\big(|\mathbf{x}| + |\mathbf{x}|^{k^{l+1}}\big).$$

Hence, for any $0 \le m \le L+1$,

$$\begin{aligned}
\max_{l\in\{1,\ldots,L+1\}} c_{l-m}(|\mathbf{x}|, \Gamma) &\le 2\big(1 + M_k^{k-1}\big)^{k^{L+1-m} + 2k^{L-m} + \cdots + (L+1-m)k + L + 2 - m}\big(|\mathbf{x}| + |\mathbf{x}|^{k^{L+2-m}}\big)\\
&:= I_1(m)\big(|\mathbf{x}| + |\mathbf{x}|^{k^{L+2-m}}\big). \qquad \text{(A.4)}
\end{aligned}$$

Next we bound $d_l(|\mathbf{x}|, \Theta, \Gamma)$. By (A.4), we have

$$\begin{aligned}
&d_l(|\mathbf{x}|, \Theta, \Gamma)\\
&= M_k^l \prod_{j=0}^{l-2} \|V^j\|_F \|W^{l-1}\|_F c_{l-2}(|\mathbf{x}|, \Gamma) + M_k^{1+(l-1)k} \prod_{j=0}^{l-2} \|V^j\|_F^k \|W^{l-1}\|_F^k c_{l-2}^k(|\mathbf{x}|, \Gamma)\\
&\le M_k^l \prod_{j=0}^{l-2} \|V^j\|_F \|W^{l-1}\|_F I_1(2)\big(|\mathbf{x}| + |\mathbf{x}|^{k^L}\big)\\
&\quad + M_k^{1+(l-1)k} \prod_{j=0}^{l-2} \|V^j\|_F^k \|W^{l-1}\|_F^k I_1(2)^k \big(|\mathbf{x}| + |\mathbf{x}|^{k^L}\big)^k.
\end{aligned}$$

Since $\Theta, \Gamma \in \mathcal{A}_\beta$, we have $\|V^j\|_F \le 1$ and $\|W^{l-1}\|_F \le 1$, so that

$$\begin{aligned}
d_l(|\mathbf{x}|, \Theta, \Gamma) &\le M_k^l\big(I_1(2)\big(|\mathbf{x}| + |\mathbf{x}|^{k^L}\big) + M_k^{(l-1)(k-1)} I_1(2)^k 2^{k-1}\big(|\mathbf{x}| + |\mathbf{x}|^{k^{L+1}}\big)\big)\\
&\le M_k^l I_1(2)\big(2 + M_k^{(l-1)(k-1)} I_1(2)^{k-1} 2^{k-1}\big)\big(|\mathbf{x}| + |\mathbf{x}|^{k^{L+1}}\big)\\
&\le M_k^{L+1} I_1(2)\big(2 + M_k^{L(k-1)} I_1(2)^{k-1} 2^{k-1}\big)\big(|\mathbf{x}| + |\mathbf{x}|^{k^{L+1}}\big)\\
&:= I_{k,L}\big(|\mathbf{x}| + |\mathbf{x}|^{k^{L+1}}\big). \qquad \text{(A.5)}
\end{aligned}$$

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

Denoting $d_l(|\mathbf{x}|, \Theta, \Gamma) := u$, for $\tilde{c}_{j-1}(u, \Theta)$, we have

$$\tilde{c}_l(u, \Theta) = u + \|W^l\|_F^{k-1} u^k \le u + u^k,$$

and, for any $j \ge l+2$, the inequality of the arithmetic and geometric means implies that

$$\begin{aligned}\tilde{c}_{j-1}(u, \Theta) &= \tilde{c}_{j-2}(u, \Theta) + M_k^{(j-1-l)(k-1)} \prod_{q=l}^{j-1} \|W^q\|^{k-1} \tilde{c}_{j-2}^k(u, \Theta) \\ &\le \tilde{c}_{j-2}(u, \Theta) + \left(\frac{M_k}{2}\right)^{(j-1-l)(k-1)} \tilde{c}_{j-2}(u, \Theta)^k,\end{aligned}$$

which implies from (A.4) that

$$\tilde{c}_{j-1}(u, \Theta) \le I_1(l+1)\big(u + u^{k^{L+1-l}}\big). \tag{A.6}$$

Hence one can derive from (A.4) and (A.5) that

$$\begin{aligned}\mathcal{K}_1 \le{}& M_k^2\big(1 + I_{k,L}^{k-1}\big(|\mathbf{x}| + |\mathbf{x}|^{k^{L+1}}\big)^{k-1} + I_1(1)^{k-1}\big(|\mathbf{x}| + |\mathbf{x}|^{k^{L+1}}\big)^{k-1}\big) I_1(2)\big(|\mathbf{x}| + |\mathbf{x}|^{k^L}\big) \\ &\times \big(1 + 2M_k^{L(k-1)} I_1(2)^{k-1}\big(|\mathbf{x}| + |\mathbf{x}|^{k^L}\big)^{k-1}\big) \max_{l\in\{1,\dots,L+1\}} \left\{ \prod_{\substack{j=0 \\ j\ne l-1}}^{L} \|W^j\|_F \vee \|V^j\|_F \right\},\end{aligned}$$

whereas by (A.3), (A.4), (A.5), and (A.6),

$$\begin{aligned}\mathcal{K}_2 \le{}& \max_{l\in\{1,\dots,L\}} \big\{1 + 2M_k^{L-l} I_1(l+1)\big(I_{k,L} + M_k^l I_1(1) + 2^{L-l}\big(I_{k,L}^{L+1-l} + M_k^{l(L+1-l)} I_1(1)^{L+1-l}\big)\big) \\ &\times \big(|\mathbf{x}| + |\mathbf{x}|^{(L+1-l)k^{L+1}}\big)\big\}^{L-l+1} \\ \le{}& \max_{l\in\{1,\dots,L\}} \big\{1 + 4M_k^{L-1} I_1(l+1)\big(I_{k,L} + M_k^L I_1(1) + 2^{L-1}\big(I_{k,L}^L + M_k^{L^2} I_1(1)^L\big)\big) \\ &\times \big(|\mathbf{x}| + |\mathbf{x}|^{Lk^{L+1}}\big)\big\}^L.\end{aligned}$$

Notice that for any positive integer r, we have

$$k^r + 2k^{r-1} + \cdots + rk + r + 1 \le \frac{k^{r+2}}{(k-1)^2}.$$

We therefore have

$$I_1(m) \le 2\big(1 + M_k^{k-1}\big)^{\frac{k^{L+3-m}}{(k-1)^2}}, \tag{A.7}$$

and by the inequality of the arithmetic and geometric means,

$$\begin{aligned}\max_{l\in\{1,\dots,L+1\}} \left\{ \prod_{\substack{j=0 \\ j\ne l-1}}^{L} \|W^j\|_F \vee \|V^j\|_F \right\} &\le \max_{l\in\{1,\dots,L+1\}} \left\{ \frac{1}{L} \sum_{j=0, j\ne l}^{L} \|W^j\|_F^2 \vee \|V^j\|_F^2 \right\}^{L/2} \\ &\le \left(\frac{1}{L} \sum_{j=0}^{L} \|W^j\|_F^2 \vee \|V^j\|_F^2 \right)^{L/2}\end{aligned}$$

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

$$\leq \left(\frac{1}{L}\sum_{j=0}^{L}\left(\|W^j\|_F^2 + \|V^j\|_F^2\right)\right)^{L/2}$$
$$\leq \left(\frac{2}{L}\right)^{L/2}.$$

These imply

$$\begin{aligned}\mathcal{K}_1 &\leq 2M_k^2\left(\frac{2}{L}\right)^{L/2}\left(1 + I_{k,L}^{k-1}\left(|\mathbf{x}| + |\mathbf{x}|^{k^{L+1}}\right)^{k-1} + 2^{k-1}\left(1 + M_k^{k-1}\right)^{\frac{k^{L+2}}{k-1}}\left(|\mathbf{x}| + |\mathbf{x}|^{k^{L+1}}\right)^{k-1}\right)\\ &\quad\times\left(1 + M_k^{k-1}\right)^{\frac{k^{L+1}}{(k-1)^2}}\left(|\mathbf{x}| + |\mathbf{x}|^{k^L}\right)\left(1 + 2^k M_k^{L(k-1)}\left(1 + M_k^{k-1}\right)^{\frac{k^{L+1}}{k-1}}\left(|\mathbf{x}| + |\mathbf{x}|^{k^L}\right)^{k-1}\right)\\ &\leq 2M_k^2\left(\frac{2}{L}\right)^{L/2}\left(1 + \left(2^{k-1}I_{k,L}^{k-1} + 4^{k-1}\left(1 + M_k^{k-1}\right)^{\frac{k^{L+2}}{k-1}}\right)\left(|\mathbf{x}|^{k-1} + |\mathbf{x}|^{(k-1)k^{L+1}}\right)\right)\\ &\quad\times\left(1 + M_k^{k-1}\right)^{\frac{k^{L+1}}{(k-1)^2}}\left(|\mathbf{x}| + |\mathbf{x}|^{k^L}\right)\left(1 + 2^{2k-1}M_k^{L(k-1)}\left(1 + M_k^{k-1}\right)^{\frac{k^{L+1}}{k-1}}\left(|\mathbf{x}|^{k-1} + |\mathbf{x}|^{(k-1)k^L}\right)\right)\end{aligned}$$

and

$$\begin{aligned}\mathcal{K}_2 &\leq \Big(1 + 8M_k^{L-1}\left(1 + M_k^{k-1}\right)^{\frac{k^{L+1}}{(k-1)^2}}\left(|\mathbf{x}| + |\mathbf{x}|^{Lk^{L+1}}\right)\\ &\quad\times\left(I_{k,L} + M_k^L 2\left(1 + M_k^{k-1}\right)^{\frac{k^{L+2}}{(k-1)^2}} + 2^{L-1}\left(I_{k,L}^L + M_k^{L^2}2^L\left(1 + M_k^{k-1}\right)^{\frac{Lk^{L+2}}{(k-1)^2}}\right)\right)\Big)^L.\end{aligned}$$

In addition, (A.7) and $M_k \geq 1$ imply that

$$\begin{aligned}I_{k,L} &\leq 2M_k^{L+1}\left(1 + M_k^{k-1}\right)^{\frac{L+1}{(k-1)^2}}\left(2 + 4^{k-1}M_k^{L(k-1)}\left(1 + M_k^{k-1}\right)^{\frac{L+1}{k-1}}\right)\\ &\leq 2^{1+\frac{L+1}{(k-1)^2}}M_k^{\frac{k(L+1)}{k-1}}2^{1+\frac{2(k-1)^2+L+1}{k-1}}M_k^{Lk+1}\\ &\leq 2^{2k+\frac{k(L+1)}{(k-1)^2}}M_k^{\frac{Lk^2+2k-1}{k-1}}.\end{aligned} \tag{A.8}$$

Therefore, by the same argument as the proof of (A.8), estimating, consolidating, and rearranging yields

$$\begin{aligned}B_1(|\mathbf{x}|, \Theta, \Gamma) &\leq 2^{\frac{(L+6)k^{L+3}}{(k-1)^2}}M_k^{\frac{(L+7)k^{L+3}}{k-1}}\left(\frac{2}{L}\right)^{L/2}\left(1 + \left(|\mathbf{x}|^{k-1} + |\mathbf{x}|^{(k-1)k^{L+1}}\right)^2\right)\left(|\mathbf{x}| + |\mathbf{x}|^{k^L}\right)\\ &\quad\times\left(1 + 2^{\frac{L(L+10)k^{L+2}}{(k-1)^2}}M_k^{\frac{(3L^2+1)k^{L+2}}{k-1}}\left(|\mathbf{x}| + |\mathbf{x}|^{Lk^{L+1}}\right)\right)^L\\ &\leq B(\mathbf{x}).\end{aligned}$$

Then the first inequality follows from Lemma A.3 and the second inequality can be derived from setting Γ equal to the all-zeros parameter in the first inequality. The proof is complete. □

Finally, we move towards the proof of Theorem 3.1.

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

Proof of Theorem 3.1. Combining with Lemmas A.3 and A.4, and applying the same arguments discussed in the proof of Theorem 3 of [38], yields the desired result. □

Acknowledgements

We wish to thank the editor and two referees for their insightful comments and suggestions on this article, which have led to significant improvements.

Funding information

Peng Chen's research was supported by the National Natural Science Foundation of China (12301176) and the International (Regional) Cooperation and Exchange Program of NSFC under NSFC-RS Joint Program (W2521122). Ting Zhang's research was supported by the National Natural Science Foundation of China (12301342).

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] Albanese, L., Alemanno, F., Alessandrelli, A. and Barra, A. (2022). Replica symmetry breaking in dense Hebbian neural networks. *J. Statist. Phys.* **189**, 24.

[2] Apicella, A., Donnarumma, F., Isgrò, F. and Prevete, R. (2021). A survey on modern trainable activation functions. *Neural Netw.* **138**, 14–32.

[3] Badrinarayanan, V., Kendall, A. and Cipolla, R. (2017). SegNet: A deep convolutional encoder–decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intellig.* **39**, 2481–2495.

[4] Bakhshizadeh, M., Maleki, A. and de la Pena, V. H. (2023). Sharp concentration results for heavy-tailed distributions. *Inf. Inference* **12**, 1655–1685.

[5] Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3**, 463–482.

[6] Bartlett, P. L, Foster, D. J. and Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems* **31**, pp. 6241–6250. Curran Associates.

[7] Bartlett, P. L., Montanari, A. and Rakhlin, A. (2021). Deep learning: A statistical viewpoint. *Acta Numer.* **30**, 87–201.

[8] Castaneda, G., Morris, P. and Khoshgoftaar, T. M. (2019). Evaluation of maxout activations in deep learning across several big data domains. *J. Big Data* **6**, 1–35.

[9] Chung, H., Lee, S. J. and Park, J. G. (2016). Deep neural network using trainable activation functions. In *2016 International Joint Conference on Neural Networks (IJCNN),* pp. 348–352. IEEE.

[10] Clevert, D. A., Unterthiner, T. and Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (ELUs). Available at arXiv:1511.07289.

[11] Dai, Z., Karzand, M. and Srebro, N. (2021). Representation costs of linear neural networks: Analysis and design. In *Advances in Neural Information Processing Systems* **35**, pp. 26884–26896. Curran Associates.

[12] Diaa, A., Fenaux, L., Humphries, T., Dietz, M., Ebrahimianghazani, F., Kacsmar, B. and Kerschbaum, F. (2024). Fast and private inference of deep neural networks by co-designing activation functions. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 2191–2208.

[13] Dubey, S. R., Singh, S. K. and Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomput.* **503**, 92–108.

[14] Feng, H. S. and Yang, C. H. (2023). PolyLU: A simple and robust polynomial-based linear unit activation function for deep learning. *IEEE Access.* **11**, 101347–101358.

[15] Geer, S. A. (2000). *Empirical Processes in M-Estimation* (Cambridge Series in Statistical and Probabilistic Mathematics **6**). Cambridge University Press.

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

[16] Gorbunov, E., Danilova, M., Dobre, D., Dvurechenskii, P., Gasnikov, A. and Gidel, G. (2022). Clipped stochastic methods for variational inequalities with heavy-tailed noise. In *Advances in Neural Information Processing Systems* **36**, pp. 31319–31332. Curran Associates.
[17] Janjua, J. I., Zulfiqar, S., Khan, T. A. and Ramay, S. A. (2023). Activation function conundrums in the modern machine learning paradigm. In *2023 International Conference on Computer and Applications (ICCA)*, pp. 1–8. IEEE.
[18] Klusowski, J. M. and Barron, A. R. (2018). Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Trans. Inform. Theory* **64**, 7649–7656.
[19] Krotov, D. and Hopfield, J. J. (2016). Dense associative memory for pattern recognition. In *Advances in Neural Information Processing Systems* **30**, pp. 1172–1180. Curran Associates.
[20] Kubjas, K., Li, J. and Wiesmann, M. (2024). Geometry of polynomial neural networks. *Algebraic Statist.* **15**, 295–328.
[21] Lederer, J. (2021). Activation functions in artificial neural networks: A systematic overview. Available at arXiv:2101.09957.
[22] Lederer, J. (2024). Statistical guarantees for sparse deep learning. *AStA Adv. Statist. Anal.* **108**, 231–258.
[23] Lederer, J. and Vogt, M. (2021). Estimating the Lasso's effective noise. *J. Mach. Learn. Res.* **22**, 12658–12689.
[24] Li, X., Ma, S., Xu, J., Tang, J., He, S. and Guo, F. (2024). TranSiam: Aggregating multi-modal visual features with locality for medical image segmentation. *Expert Syst. Appl.* **237**, 121574.
[25] Lyu, K. and Li, J. (2019). Gradient descent maximizes the margin of homogeneous neural networks. Available at arXiv:1906.05890.
[26] McCulloch, J. A., Pierre, S. R. S., Linka, K. and Kuhl, E. (2024). On sparse regression, L_p-regularization, and automated model discovery. *Internat. J. Numer. Methods Engrg* **125**, e7481.
[27] Mohades, A. and Lederer, J. (2023). Reducing computational and statistical complexity in machine learning through cardinality sparsity. Available at arXiv:2302.08235.
[28] Neyshabur, B., Tomioka, R. and Srebro, N. (2015). Norm-based capacity control in neural networks. In *Proceedings of the 28th Conference on Learning Theory*, pp. 1376–1401. PMLR.
[29] Pusztaházi, L. S., Eigner, G. and Csiszár, O. (2024). Parametric activation functions for neural networks: A tutorial survey. *IEEE Access* **12**, 168626–168644.
[30] Ramachandran, P., Zoph, B. and Le, Q. V. (2017). Searching for activation functions. Available at arXiv:1710.05941.
[31] Rasamoelina, A. D., Adjailia, F. and Sinčák, P. (2020). A review of activation function for artificial neural network. In *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pp. 281–286. IEEE.
[32] Rigollet, P. and Hütter, J. C. (2023). High-dimensional statistics. Available at arXiv:2310.19244.
[33] Serban, A., Poll, E. and Visser, J. (2020). Adversarial examples on object recognition: A comprehensive survey. *ACM Comput. Surv.* **53**, 1–38.
[34] So, D. R., Mańke, W., Liu, H., Dai, Z., Shazeer, N. and Le, Q. V. (2021). Primer: Searching for efficient transformers for language modeling. Available at arXiv:2109.08668.
[35] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958.
[36] Sun, R. (2019). Optimization for deep learning: Theory and algorithms. Available at arXiv:1912.08957.
[37] Szandala, T. (2021). Review and comparison of commonly used activation functions for deep neural networks. *Bio-Inspired Neurocomput.* **903**, 203–224.
[38] Taheri, M., Xie, F. and Lederer, J. (2021). Statistical guarantees for regularized neural networks. *Neural Netw.* **142**, 148–161.
[39] Taheri, M., Xie, F. and Lederer, J. (2022). Statistical guarantees for approximate stationary points of simple neural networks. Available at arXiv:2205.04491.
[40] Tian, Y., Pei, K., Jana, S. and Ray, B. (2018). DeepTest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th International Conference on Software Engineering*, pp. 303–314. ACM.
[41] Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science* (Cambridge Series in Statistical and Probabilistic Mathematics **47**). Cambridge University Press.
[42] Wei, C., Lee, J. D., Liu, Q. and Ma, T. (2019). Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems* **33**. Curran Associates.
[43] Wu, L., Li, J., Wang, Y., Meng, Q., Qin, T., Chen, W. and Liu, T. Y. (2021). R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems* **35**, pp. 10890–10905. Curran Associates.

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press

[44] XIANG, W., TRAN, H. D. AND JOHNSON, T. T. (2018). Output reachable set estimation and verification for multilayer neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **29**, 5777–5783.
[45] YILDIRIM, H. AND ÖZKALE, M. R. (2019). The performance of ELM based ridge regression via the regularization parameters. *Expert Syst. Appl.* **134**, 225–233.
[46] ZHANG, L., SONG, J., GAO, A., CHEN, J., BAO, C. AND MA, K. (2019). Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3713–3722.

https://doi.org/10.1017/jpr.2026.10121 Published online by Cambridge University Press