Nonasymptotic Theory for Two-Layer Neural Networks: Beyond the Bias–Variance Trade-Off

Huiyuan Wang HUIYUAN.WANG@PENNMEDICINE.UPENN.EDU
Department of Biostatistics, Epidemiology and Informatics
Perelman School of Medicine
University of Pennsylvania
Philadelphia, PA 19104, USA

Wei Lin WEILIN@MATH.PKU.EDU.CN
School of Mathematical Sciences and Center for Statistical Science
Peking University
Beijing 100871, China

Editor: My editor

Abstract

Large neural networks have proved remarkably effective in modern deep learning practice, even in the overparametrized regime where the number of active parameters is large relative to the sample size. This contradicts the classical perspective that a machine learning model must trade off bias and variance for optimal generalization. To resolve this conflict, we present a nonasymptotic generalization theory for two-layer neural networks with ReLU activation function by incorporating scaled variation regularization. Interestingly, the regularizer is equivalent to ridge regression from the angle of gradient-based optimization, but plays a similar role to the group lasso in controlling the model complexity. By exploiting this "ridge–lasso duality," we obtain new prediction bounds for any finite network width, which reproduce the double descent phenomenon. Moreover, the overparametrized minimum risk is lower than its underparametrized counterpart when the signal is strong, and is minimax optimal over a suitable class of functions. By contrast, we show that overparametrized random feature models suffer from the curse of dimensionality and thus are suboptimal.

Keywords: Double descent, generalization, neural network, overparametrization, regularization

1 Introduction

During the past decade, deep learning has demonstrated superiority over traditional machine learning techniques for representation learning and prediction in a wide variety of tasks, including object recognition in computer vision (He et al., 2016), machine translation and text generation in natural language processing (Sutskever et al., 2014), general game playing (Schrittwieser et al., 2020), and disease diagnosis in clinical research (Esteva et al., 2017). Many such successful applications build on large neural networks that operate in the overparametrized regime, where the number of parameters is relatively large compared to the number of training samples. For instance, the convolutional neural network AlexNet (Krizhevsky et al., 2012) had 60 million parameters trained on 1.2 million images; the more recent large language model GPT-3 was trained with 175 billion parameters and 300 billion training tokens (Brown et al., 2020).

Theoretical insights into overparametrized neural networks have been obtained from the optimization viewpoint (Arora et al., 2018; Soltanolkotabi et al., 2019), suggesting that overparametrization can speed up convergence or improve the optimization landscape. The benefits of overparametrization to generalization in deep learning, however, remain mysterious. Numerical evidence indicates that deep neural networks easily fit random labels but still generalize well even without explicit regularization (Zhang et al., 2021). These empirical findings deeply challenge the conventional wisdom that optimal generalization should be achieved by trading off bias and variance. The so-called "double descent" curve (Belkin et al., 2019; Nakkiran et al., 2021a) was proposed and conjectured as a ubiquitous phenomenon for unifying the generalization behaviors of machine learning models across the underparametrized and overparametrized regimes, but so far has not been theoretically justified for realistic neural networks.

While the notion of overparametrization is not new and has long been studied in high-dimensional statistics (Wainwright, 2019), there are some fundamental differences between the usual high-dimensional models and overparametrized deep learning models. In high-dimensional problems, although the number of parameters can be large or even exponentially growing, it is almost always assumed that certain parsimonious structures (e.g., sparsity and low-rankness) exist and can be exploited. For example, recent work has shown that minimum norm interpolators have near-optimal prediction risk and hence overfitting is not detrimental in linear regression when the parameters are sparse or the design matrix is low-rank (Bartlett et al., 2020; Muthukumar et al., 2020; Hastie et al., 2022; Chinot et al., 2022). Such parsimony and the regularization for achieving it play two roles: (i) to control the model complexity for balancing bias and variance, and (ii) to ensure model identifiability so that prediction and estimation are essentially equivalent. These ideas, however, do not readily extend to overparametrized neural networks, because: (i) sparsity-inducing regularization is often not required in deep learning or not strong enough (e.g., in dropout) to bring the dimensionality down to a level below the sample size (Srivastava et al., 2014); and (ii) neural networks are intrinsically unidentifiable owing to weight space symmetry and many other equivalent parametrizations (Goodfellow et al., 2016, p. 277).

Neural networks are pure prediction algorithms in the sense of Efron (2020), which operate in a nonparametric and nonparsimonious way. Nonparametric theory for neural networks was pioneered by Barron (1994), who derived risk bounds in terms of the network width for complexity-regularized two-layer sigmoidal networks. For different function classes and the now popular ReLU activation function, recent developments have shown that deep neural networks can deliver minimax optimal rates of convergence and in certain cases circumvent the curse of dimensionality (Schmidt-Hieber, 2020; Hayakawa and Suzuki, 2020; Farrell et al., 2021; Kohler and Langer, 2021). The architectural constraints imposed by this line of work, however, require the networks to be sparse or of small size, restricting the number of nonzero or active parameters to a smaller order than the sample size. Therefore, although these results demonstrate the efficiency of deep architectures, they are still confined to the underparametrized regime and do not go beyond the bias–variance trade-off.

Another line of work controls the model complexity of neural networks via norm-based regularization and obtains complexity and risk bounds in terms of various norms of the estimated network parameters. Neyshabur et al. (2015a) and Golowich et al. (2020), among others, considered group norm and matrix norm regularization and derived size-independent bounds on the Rademacher complexity. However, as observed empirically by Neyshabur et al. (2019), these complexity measures increase with the network size and do not correlate with the test error. As a result, they may lead to vacuous bounds for large networks and are not sufficient to explain the role of overparametrization. Recognizing these gaps, Neyshabur et al. (2019) presented complexity bounds that empirically decrease with the network size and could potentially explain the benefits of large networks. Nevertheless, norm-based complexity measures implicitly depend on the network size and the training process, which are difficult to analyze precisely and control tightly.

This paper contributes to the ongoing debate about the role of overparametrization in deep learning by developing a nonasymptotic theory for two-layer neural networks across the underparametrized and overparametrized regimes. Our theory is intended to be as transparent as possible, relying on no sparsity assumptions and giving rise to sharp risk bounds in terms of the sample size, dimensionality, and network width. Building on this theory, we aim to gain insight into the following questions:

- How does the network perform in the overparametrized regime differently from in the underparametrized regime?
- How does the overparametrized minimum risk compare with its underparametrized counterpart and how far is it from optimal?

Specifically, suppose that we observe predictors $\mathbf{x}_i \in \mathbb{R}^d$ and responses $y_i \in \mathbb{R}$ generated from the nonparametric regression model

$$y_i = f^*(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where f^* is an unknown function to be estimated and ε_i are random errors. Let $\sigma(z) = \max(z, 0)$ be the rectified linear unit (ReLU) activation function (Jarrett et al., 2009). We consider a two-layer neural network with m hidden units, $g(\cdot;\boldsymbol{\theta})\colon \mathbb{R}^d \to \mathbb{R}$, of the form

$$g(\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{m} a_k \sigma(\mathbf{v}_k^T \mathbf{x} + b_k) \tag{2}$$

with parameters $\boldsymbol{\theta} = (a_1, \ldots, a_m, \mathbf{v}_1^T, \ldots, \mathbf{v}_m^T, b_1, \ldots, b_m)^T$. Without loss of generality, we do not include an intercept term. In addition to the usual assumptions on $\mathbf{x}_i$ and ε_i, our key assumption is that f^* belongs to a suitably defined class of functions with certain integral representations, which allows us to bypass the curse of dimensionality. Detailed assumptions will be given in Section 2.3.

By incorporating a *scaled variation* regularizer to be defined in Section 2.2, our main result (Theorem 10) shows that the prediction (or generalization) error of the regularized network estimator $g(\cdot;\widehat{\boldsymbol{\theta}})$ is of order

$$\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2) \min\left(\frac{md\log n}{n}, \sqrt{\frac{d\log n}{n}}\right), \tag{3}$$

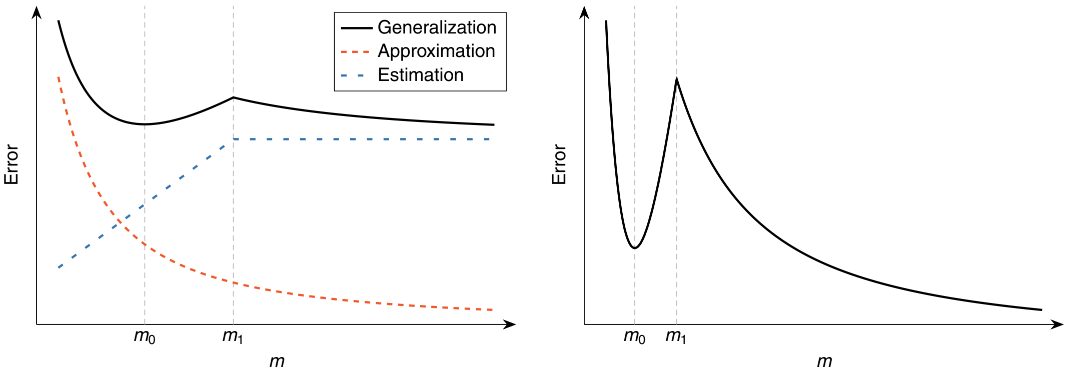


Figure 1: Risk curves for varying network width m from the prediction bound (3) with $\|f^*\|_{\mathcal{S}}^2/\sigma_\varepsilon^2 = 1$, $d = 6$, and $n = 1000$. The left panel shows the decomposition of prediction error into approximation and estimation errors. The right panel shows the same plot but with larger m, from which it is apparent that the second valley is lower than the first.

where $\|f^*\|_{\mathcal{S}}$ is the $\mathcal{S}$-norm of f^* (Definition 1) and σ_ε^2 is the variance of ε_i. We emphasize that this result holds for all $m \geq 1$ and any global minimizer of the regularized empirical risk. The prediction bound (3) consists of two terms: the first term represents the approximation error, which decreases with the network width m, while the second term represents the estimation error, which increases with m up to some critical point $m_1 \asymp \sqrt{n/(d \log n)}$ and thereafter stays constant. An intriguing consequence of this unusual trade-off is a double descent risk curve, as shown in Figure 1. To answer our question regarding optimality, we find the first valley or underparametrized minimum risk to be $O((d \log n/n)^{(d+3)/(2d+3)})$, which occurs at $m_0 \asymp (n/(d \log n))^{d/(2d+3)}$, by matching the approximation and estimation errors in (3). While this rate is slightly better than that of the second valley or overparametrized minimum risk, $O(\sqrt{d \log n/n})$, the asymptotic comparison can be reversed in finite samples, as shown in the right panel of Figure 1. When the signal-to-noise ratio $\|f^*\|_{\mathcal{S}}^2/\sigma_\varepsilon^2$ is large, the second valley tends to be lower than the first; a precise condition is given in (18). We further prove that the overparametrized minimum risk is minimax rate-optimal over a suitable class of functions (Theorem 11). By contrast, overparametrized random feature models suffer from the curse of dimensionality and thus are suboptimal (Proposition 12). Overall, our results lend theoretical support to the benefits of overparametrization in deep learning and shed light on the currently debated double descent phenomenon.

Intuitively, the number of parameters or the network width m is not an appropriate measure of model complexity for the network (2) in the overparametrized regime, and one must seek alternatives. The idea of our approach to achieving model complexity control while allowing m to grow unbounded is to exploit the *ridge–lasso duality* of the scaled variation regularizer. On the one hand, by the positive homogeneity of the ReLU function, a reparametrization yields the equivalence of scaled variation regularization to ridge regression, or (standard) *weight decay* in deep learning (Krogh and Hertz, 1991), which in general does not induce sparsity. This characterization plays a conceptual role in ensuring that our improved complexity control is not due to sparsity, so that an intrinsic overparametrization is possible. On the other hand, a linearization of the ReLU function by parameter space partitioning transforms the regularized problem into a group lasso form. This gives a key insight into the geometry of the global minima: the estimated network weights residing in the same region must be parallel to each other. Such collinearity greatly reduces the effective number of parameters and enables us to measure the model complexity in terms of the number of nonparallel directions. This implicit (within-group) formation and (between-group) breaking of symmetry lies at the heart of our theoretical analysis.

Our proofs leading to the prediction bound (3) consist of several ingredients. We first investigate the approximation properties of two-layer ReLU networks and obtain sharp approximation bounds in Theorem 2. The group lasso formulation mentioned above and detailed in Section 3.3 allows us to borrow techniques from high-dimensional statistics for deriving our nonasymptotic theory. Deviating from the standard theory for high-dimensional linear regression, we make no sparsity or eigenvalue assumptions. Instead, we leverage a tight control of the scaled variation norm of the best-approximating finite-width network to achieve a convergence rate slow than the usual $O(n^{-1})$ rate but nevertheless minimax optimal (Theorem 11). To do so, we first bound the empirical error for fixed designs (Theorem 7) and then prove similar prediction bounds for random designs (Theorem 8) via a maximal inequality. Noting that sparsity-based complexity control is more effective for narrower networks, we establish prediction bounds in the underparametrized regime via a metric entropy argument (Theorem 9). Combining Theorems 8 and 9 yields the ultimate result (Theorem 10) across the underparametrized and overparametrized regimes.

1.1 Related Work

Although not the focus of this paper, approximation theory is often an integral part and first step of establishing statistical guarantees for neural networks. Sharp approximation bounds can be obtained for target functions that are well represented by two-layer neural networks, for which purpose various function spaces have been proposed. For sigmoidal networks, the seminal work of Barron (1993) considered a class of functions that have an integral representation involving the Fourier transform. The idea was further developed by, for example, Bach (2017) and Siegel and Xu (2024) to define variation spaces and norms for positively homogeneous activation functions, including ReLU. Other recent work (Ongie et al., 2020; Parhi and Nowak, 2021) has introduced an equivalent characterization of the variation space for two-layer ReLU networks via the Radon transform and has related it to more classical function spaces (Parhi and Nowak, 2023). Our choice of the target function space and its associated norm is similar to but slightly extends those of Ongie et al. (2020) and Parhi and Nowak (2023) to allow the identification of affine functions.

There is an enormous literature on inference in nonparametric or infinite-dimensional models, providing sharp risk bounds for a variety of nonparametric estimation problems; see, for example, Tsybakov (2009), Giné and Nickl (2016), and references therein. However, this well-developed theory of nonparametric statistics primarily concerns the classical function spaces such as Hölder and Sobolev spaces, with minimax rates that are too slow to distinguish between the performance of neural networks and other nonparametric methods. Moreover, most results for classical nonparametric problems involve a bias–variance tradeoff that seems inevitable (Derumigny and Schmidt-Hieber, 2023). Our work departs from

Source	Rate of convergence	Regularization
Barron (1994)	$m^{-1} + md \log n/n$	Complexity
E et al. (2019)	$m^{-1} + \log n \sqrt{\log d/n}$	ℓ_1 path norm
Parhi and Nowak (2023)	$n^{-(d+3)/(2d+3)}$ (up to logarithmic factors)	Total variation
This work	$m^{-(d+3)/d} + \min(md \log n/n, \sqrt{d \log n/n})$	Scaled variation

Table 1: Comparisons of generalization bounds for two-layer neural networks.

this research by exploiting a more relevant function space for two-layer neural networks and demonstrating unconventional behavior of the risk curve.

Generalization bounds have been derived for two-layer neural networks in certain variation spaces. Most of the existing work focuses on variational formulations of the constrained empirical risk minimization problem. For example, Bach (2017) and Parhi and Nowak (2023) considered a variational problem by constraining the network estimator to within a ball in the function space. Although Parhi and Nowak (2023) showed that such network estimators are nearly minimax optimal, little information is provided on the network width for their estimator to attain the optimality. A representer theorem of Parhi and Nowak (2021) implies that the network width of the solution is smaller than the sample size; in this sense the minimax optimality of the constrained network estimator provides no clue about the effect of overparametrization. One exception is the work of E et al. (2019), which obtained generalization bounds for regularized two-layer networks that allow the network width to grow unbounded. The ℓ_1 path norm regularization that they adopted, however, induces sparsity in the network parameters, casting doubt on the implication of their results for intrinsically overparametrized networks. Comparisons between our results and some representative ones from the literature are summarized in Table 1.

Mean-field and neural tangent kernel theories are two popular frameworks for analyzing the training dynamics of two-layer neural networks in the infinite-width limit. The mean-field theory shows that the stochastic gradient descent dynamics of two-layer networks is asymptotically described by a nonlinear partial differential equation (PDE), and approximation results such as laws of large numbers and central limit theorems can be derived (Mei et al., 2018; Sirignano and Spiliopoulos, 2020; Rotskoff and Vanden-Eijnden, 2022). The generalization behavior of the PDE model, however, is difficult to study except in some specific examples. Under a different scaling, overparametrized two-layer networks are shown to behave as their linearizations at random initialization, and optimization and generalization properties can be investigated by exploiting the neural tangent kernel (Jacot et al., 2018) and the associated kernel methods. This "lazy training" regime (Chizat et al., 2019) entails a large performance gap between realistic and linearized networks and hence does not explain the power of fully trained neural networks (E et al., 2020; Ghorbani et al., 2021). Dou and Liang (2021) went a step further and developed an adaptive theory for neural network training with data-adaptive kernels. Nevertheless, the impact of adaptivity on generalization remains unclear.

Since the conceptualization of the double descent curve by Belkin et al. (2019), several theoretical models and explanations have been developed for the phenomenon. A majority of these efforts have focused on minimum norm least squares and ridge regression in linear and random feature models (Belkin et al., 2020b; Hastie et al., 2022; Muthukumar et al., 2020; Mei and Montanari, 2022). Random matrix theory is the backbone of most of these results, which concerns the high-dimensional asymptotic regime where $n, d \to \infty$ with $n \asymp d$. Similar asymptotics have also been derived for classification problems (Deng et al., 2022; Liang and Sur, 2022). Li and Meng (2021) and Liang et al. (2020) demonstrated a "multiple descent" phenomenon in infinite-dimensional linear regression and kernel ridgeless regression. Despite these important developments, it still seems difficult to isolate a general mechanism for double descent and determine whether the phenomenon results from specific model assumptions or asymptotic regimes.

Geman et al. (1992) pioneered research on the bias–variance trade-off in neural networks by measuring bias and variance individually. Recent work along this direction exploits the bias–variance decomposition and suggests unconventional behavior of the bias and variance components. For example, Yang et al. (2020) and Liu et al. (2021) developed theory allowing the risk curve to have qualitatively different shapes, with the double descent curve arising from a unimodal variance. Adlam and Pennington (2020) and Lin and Dobriban (2021) performed a fine-grained ANOVA decomposition of the variance into components associated with randomness from sampling, initialization, and label noise. Chen et al. (2024) presented a phenomenon called the bias–variance alignment and connected it with calibration and neural collapse. Most of these mechanisms, however, were theoretically verified only under simplified settings or restrictive assumptions. It remains elusive whether and how they extend to realistic neural networks and fit in with a modern understanding of the bias–variance trade-off (Derumigny and Schmidt-Hieber, 2023).

Our analysis is carried out through the approximation–estimation decomposition, which is closely related to but subtly different from the bias–variance decomposition (Brown and Ali, 2024). More precisely, the approximation error reflects the *model bias* that arises from approximating the true data-generating process by a restricted family of prediction rules (e.g., finite-width neural networks). The estimation error measures the difference of errors between a sample-based prediction rule and the best achievable within the restricted family. It contributes both an *estimation bias* and an *estimation variance* to the prediction error (Hastie et al., 2009, Section 7.3). Thus, the *total bias*, which consists of the model bias and estimation bias, may behave differently from the approximation error, which is the model bias alone. While the bias–variance decomposition is more ready for numerical evaluation, the approximation–estimation decomposition is more amenable to theoretical analysis, especially for complex nonparametric models and from the nonasymptotic viewpoint.

1.2 Organization of the Paper

Section 2 introduces the definitions of two-layer ReLU networks and the target function class. Theoretical assumptions and approximation properties are also described. Section 3 presents the regularized estimation framework and formalizes the ridge–lasso duality. Our main results, including nonasymptotic generalization guarantees and minimax optimality, are developed in Section 4. Section 5 discusses the random feature model and points out its suboptimality. Section 6 provides some further discussion. Proofs are deferred to the Appendices.

2 Preliminaries

2.1 Notation

For $1 \leq q < \infty$, let $\| \cdot \|_q$ denote the ℓ_q-norm of a vector. Let $\mathbb{B}^d$ and $\mathbb{S}^{d-1}$ denote the unit ℓ_2-ball and ℓ_2-sphere, respectively, in $\mathbb{R}^d$. For a matrix $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$, define the $\ell_{2,1}$-norm $\|\mathbf{B}\|_{2,1} = \sum_{j=1}^J \|\boldsymbol{\beta}_j\|_2$. Denote by $\mathcal{M}(D)$ the set of signed measures α on D with finite total variation $|\alpha|(D)$. In particular, the Dirac measure $\delta_{\mathbf{x}} \in \mathcal{M}(D)$ if $\mathbf{x} \in D$. For a function f, let $\|f\|_\infty$ denote the L_∞-norm on $\mathbb{B}^d$, and $\|f\|_2$ and $\|f\|_n$ the $L_2(\mu)$-norm and its empirical counterpart, respectively.

2.2 Neural Networks and the Target Function Class

We consider the two-layer neural network $g(\cdot; \boldsymbol{\theta})$ with ReLU activation function and width m given by (2). Let $\boldsymbol{\Theta}_m$ denote the parameter space. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}_m$, define the *scaled variation norm* of $g(\cdot; \boldsymbol{\theta})$ by

$$\nu(\boldsymbol{\theta}) = \sum_{k=1}^m |a_k| \|\mathbf{w}_k\|_2, \tag{4}$$

where $\mathbf{w}_k = (\mathbf{v}_k^T, b_k)^T$. Using (4) as a regularizer for two-layer ReLU networks has been considered by Parhi and Nowak (2021) and Parhi and Nowak (2023). The scaled variation norm is similar to but different from the ℓ_p path norm introduced by Neyshabur et al. (2015a). In fact, it coincides with the ℓ_1 path norm when the $d+1$ input nodes collapse into a single node; for $d \geq 1$, however, it is not separable in the first-layer weights and hence not a path norm. We have coined the name to emphasize its connection with the total variation norm and distinguish it from the usual ℓ_1 path norm considered in previous work. Moreover, the scaled variation norm is a finite-dimensional version of the $\mathcal{S}$-norm to be introduced in Definition 1 below, except for the additional inclusion of the bias term. We will show in Section 3 that this regularizer has some desirable properties that are key to our theoretical analysis.

To develop intuition for our target function space, note that the finite-width network $g(\cdot; \boldsymbol{\theta})$ has an integral representation with respect to a discrete signed measure. Specifically, if we define $\alpha_m = \sum_{k=1}^m a_k \delta_{\mathbf{w}_k}$, then

$$g(\mathbf{x}; \boldsymbol{\theta}) = \int_{\mathbb{R}^{d+1}} \big(\sigma(\mathbf{v}^T \mathbf{x} + b) - \sigma(b)\big)\, d\alpha_m(\mathbf{w}) + g(\mathbf{0}; \boldsymbol{\theta}),$$

where $\mathbf{w} = (\mathbf{v}^T, b)^T$. Motivated by this representation, we are interested in general functions associated with a signed measure $\alpha \in \mathcal{M}(\mathbb{R}^{d+1})$, which are of the form

$$g_\alpha(\mathbf{x}) = \int_{\mathbb{R}^{d+1}} \big(\sigma(\mathbf{v}^T \mathbf{x} + b) - \sigma(b)\big)\, d\alpha(\mathbf{w}) + g_\alpha(\mathbf{0}).$$

This can be viewed as an *infinite-width* two-layer ReLU network and naturally represents those functions that can be approximated by *finite-width* ReLU networks. For $g_\alpha(\cdot)$ to be well defined, a sufficient condition is

$$\int_{\mathbb{R}^{d+1}} \|\mathbf{v}\|_2 \, d|\alpha|(\mathbf{w}) < \infty,$$

since by the Lipschitz continuity of the ReLU function, $|\sigma(\mathbf{v}^T\mathbf{x}+b)-\sigma(b)| \leq |\mathbf{v}^T\mathbf{x}| \leq \|\mathbf{v}\|_2\|\mathbf{x}\|_2$. Treating functions that differ by a constant as identical, we consider the space of functions modulo constants

$$\mathcal{G} = \left\{\mathbf{x} \mapsto \int_{\mathbb{R}^{d+1}} \big(\sigma(\mathbf{v}^T\mathbf{x}+b)-\sigma(b)\big)\, d\alpha(\mathbf{w}) : \int_{\mathbb{R}^{d+1}} \|\mathbf{v}\|_2\, d|\alpha|(\mathbf{w}) < \infty\right\}. \tag{5}$$

Interestingly, there is a one-to-one correspondence between $\mathcal{G}$ and $\mathcal{M}_2(\mathbb{R}^{d+1}) \equiv \big\{\alpha \in \mathcal{M}(\mathbb{R}^{d+1}) : \int_{\mathbb{R}^{d+1}} \|\mathbf{v}\|_2\, d|\alpha|(\mathbf{w}) < \infty\big\}$. A formal statement is given by Proposition 19 in Appendix D.2. At first sight, the definition of $\mathcal{G}$ depends on the ReLU activation function $\sigma(\cdot)$. However, this dependence can be eliminated by an equivalent characterization via the Radon transform (Ongie et al., 2020; Parhi and Nowak, 2021).

To equip the function space $\mathcal{G}$ with a norm, we introduce the following definition.

Definition 1 *The $\mathcal{S}$-norm of $f \in \mathcal{G}$ is defined as $\|f\|_{\mathcal{S}} = \int_{\mathbb{R}^{d+1}} \|\mathbf{v}\|_2\, d|\alpha_f|(\mathbf{w})$, where the signed measure $\alpha_f \in \mathcal{M}_2(\mathbb{R}^{d+1})$ is uniquely determined by*

$$f(\mathbf{x}) = \int_{\mathbb{R}^{d+1}} \big(\sigma(\mathbf{v}^T\mathbf{x}+b)-\sigma(b)\big)\, d\alpha_f(\mathbf{w}) + f(\mathbf{0}).$$

We give some examples of functions in $\mathcal{G}$ to illustrate the $\mathcal{S}$-norm:

- For any linear function $f(\mathbf{x}) = \boldsymbol{\beta}^T\mathbf{x}$, note that $f(\mathbf{x}) = \sigma(\boldsymbol{\beta}^T\mathbf{x}) - \sigma(-\boldsymbol{\beta}^T\mathbf{x})$ and hence $\|f\|_{\mathcal{S}} = 2\|\boldsymbol{\beta}\|_2$.
- For the two-layer ReLU network $g(\cdot;\boldsymbol{\theta})$ of width m in (2), we have $\|g(\cdot;\boldsymbol{\theta})\|_{\mathcal{S}} = \sum_{k=1}^m |a_k|\|\mathbf{v}_k\|_2$.
- Let ρ be a probability measure on $\mathbb{S}^{d-1}\times[-1,1]$. Consider the reproducing kernel Hilbert space (RKHS)

$$\mathcal{H}_\rho = \left\{\mathbf{x} \mapsto \int_{\mathbb{S}^{d-1}\times[-1,1]} a(\mathbf{w})\sigma(\mathbf{v}^T\mathbf{x}+b)\, d\rho(\mathbf{w}) : \int_{\mathbb{S}^{d-1}\times[-1,1]} |a(\mathbf{w})|^2\, d\rho(\mathbf{w}) < \infty\right\} \tag{6}$$

  associated with the kernel

$$H_\rho(\mathbf{x},\mathbf{z}) = \mathbb{E}_{\mathbf{w}\sim\rho}\big(\sigma(\mathbf{v}^T\mathbf{x}+b)\sigma(\mathbf{v}^T\mathbf{z}+b)\big). \tag{7}$$

  If $f \in \mathcal{H}_\rho$, then $\|f\|_{\mathcal{S}} = \mathbb{E}_{\mathbf{w}\sim\rho}(|a(\mathbf{w})|)$; see Lemma 21.
- Let $\phi\colon \mathbb{R}\to\mathbb{R}$ be twice differentiable with $\phi'' \in L_1(\mathbb{R})$, and fix some $\mathbf{v}_0 \in \mathbb{S}^{d-1}$. Then the ridge function $f(\mathbf{x}) = \phi(\mathbf{v}_0^T\mathbf{x})$ satisfies

$$\|f\|_{\mathcal{S}} = 2|\phi'(0)| + \int_{-\infty}^{\infty} |\phi''(t)|\, dt < \infty;$$

  see Lemma 22.

Our definition of the target function space is inspired by and related to several previously studied spaces for two-layer neural networks. In particular, $\mathcal{G}$ is equivalent to the variation spaces in Bach (2017) and Siegel and Xu (2024) and the Radon bounded variation spaces in Ongie et al. (2020) and Parhi and Nowak (2021). In contrast to classical smoothness spaces such as Hölder and Besov spaces (Farrell et al., 2021; Suzuki, 2019), this function space is among the few that allow us to avoid the curse of dimensionality and show a performance gap between neural networks and other nonparametric methods such as random feature models. Meanwhile, it is flexible enough to include (i) sufficiently smooth functions in the spectral Barron space and Sobolev spaces, and (ii) more structured functions such as ridge functions and those in certain reproducing kernel Hilbert spaces.

Our definition of the $\mathcal{S}$-norm is more transparent than those in previous work since it is defined explicitly as a functional of α_f, a uniquely determined signed measure. Moreover, it slightly improves on previous proposals in several respects. Notably, for an affine function $f(\mathbf{x}) = \boldsymbol{\beta}^T\mathbf{x}+c$, $\|f\|_{\mathcal{S}} = 2\|\boldsymbol{\beta}\|_2$ instead of being zero. This has two important consequences: (i) the $\mathcal{S}$-norm is a norm rather than a seminorm; and (ii) there is no need to introduce a skip connection in a representer theorem (cf. Ongie et al., 2020; Parhi and Nowak, 2021). The latter is compatible with deep learning practice since skip connections are only necessary in deep neural networks such as residual networks (He et al., 2016). More mathematical details can be found in Appendix D.

2.3 Assumptions

We consider the nonparametric regression model (1) and impose the following conditions:

(C1) $f^* \in \mathcal{G}_M \equiv \{f \in \mathcal{G} : \|f\|_{\mathcal{S}} \le M\}$ for some constant $M > 0$;

(C2) $\mathbf{x}_i \sim \mu$ independently, where μ is supported in $\mathbb{B}^d$;

(C3) $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ independently and are independent of $\mathbf{x}_i$.

Condition (C2) is mild and standard in the machine learning literature since the predictors are usually bounded and can be normalized. Under Condition (C2), it suffices to consider the restrictions of functions in $\mathcal{G}$ (or $\mathcal{G}_M$) to $\mathbb{B}^d$; denote the space of such restrictions by $\mathcal{G}(\mathbb{B}^d)$ (or $\mathcal{G}_M(\mathbb{B}^d)$). An important consequence is that, for any $f \in \mathcal{G}(\mathbb{B}^d)$, there exists a signed measure $\widetilde{\alpha}_f \in \mathcal{M}(\mathbb{S}^{d-1} \times [-1,1])$ and $c \in \mathbb{R}$ such that

$$f(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} \sigma(\mathbf{v}^T\mathbf{x} + b)\, d\widetilde{\alpha}_f(\mathbf{w}) + c, \quad \mathbf{x} \in \mathbb{B}^d; \tag{8}$$

see Proposition 20 in Appendix D.2. Compared with a similar integral representation in Parhi and Nowak (2023, Remark 3), note that no skip connection appears in (8). Thus, functions in $\mathcal{G}(\mathbb{B}^d)$ have a simpler integral representation

$$\mathcal{G}(\mathbb{B}^d) = \left\{\mathbf{x} \mapsto \int_{\mathbb{S}^{d-1}\times[-1,1]} \sigma(\mathbf{v}^T\mathbf{x} + b)\, d\alpha(\mathbf{w}) : |\alpha|(\mathbb{S}^{d-1} \times [-1,1]) < \infty\right\},$$

which will allow us to obtain a sharp approximation bound.

2.4 Approximation Properties

Approximation rates for two-layer neural networks of width m have been derived in various function spaces. A classical probabilistic argument, first applied to neural networks by Barron (1993), yields an approximation rate of $O(1/\sqrt{m})$ in the L_2-norm; see also Jones (1992) and Siegel and Xu (2020). The approximation rate has been improved by Makovoz (1996), Bach (2017), and Klusowski and Barron (2018), among others. In particular, Bach (2017) obtained an $O(m^{-(d+3)/(2d)})$ rate in the L_∞-norm by using a result from geometric discrepancy theory (Matoušek, 1996); see also Siegel (2025). Moreover, this rate is not improvable (Bourgain et al., 1989). As for our target function space $\mathcal{G}_M$, we have the following approximation result, which is a direct consequence of Bach (2017, Proposition 1) and the integral representation (8).

Theorem 2 *For any $f \in \mathcal{G}_M(\mathbb{B}^d)$, there exists a network $g(\cdot;\boldsymbol{\theta})$ of width m in the form of (2) such that $\nu(\boldsymbol{\theta}) \le 6\|f\|_{\mathcal{S}}$ and*

$$\|f - g(\cdot;\boldsymbol{\theta})\|_\infty \le C\|f\|_{\mathcal{S}} m^{-(d+3)/(2d)}$$

for some constant $C > 0$ depending only on d.

The construction in Theorem 2 has a tight control on the scaled variation norm of the network parameter. This suggests using the scaled variation norm as a regularizer for the network estimation problem, as we will discuss in the next section.

3 Methodology and the Ridge–Lasso Duality

In this section we introduce our regularized estimation problem and formalize the notion of the ridge–lasso duality through two different reparametrizations. The reparametrization into ridge regression argues from the optimization angle that the regularized problem can enter the overparametrized regime without enforcing sparsity. The reparametrization into the group lasso provides an effective tool for parameter reduction and complexity control, which will be needed in our proofs for Section 4.

3.1 Regularized Estimation

In order to learn f^* from the training sample, we adopt the penalized empirical risk minimization (ERM) framework and seek to minimize

$$J_n(\boldsymbol{\theta};\lambda) = \frac{1}{2n}\sum_{i=1}^{n}\big(y_i - g(\mathrm{x}_i;\boldsymbol{\theta})\big)^2 + \lambda\nu(\boldsymbol{\theta}),$$

where $g(\cdot;\boldsymbol{\theta})$ is the two-layer ReLU network of width m in (2), $\nu(\boldsymbol{\theta})$ is the scaled variation norm in (4), and $\lambda > 0$ is a regularization parameter. The regularized network estimator is given by $g(\cdot;\widehat{\boldsymbol{\theta}})$, where

$$\widehat{\boldsymbol{\theta}} = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_m} J_n(\boldsymbol{\theta};\lambda). \tag{9}$$

In a related work, Parhi and Nowak (2023) studied a variational problem in the variation space associated with two-layer ReLU networks, where constraint-based regularization is imposed on the variation norm of the network function. A representer theorem guarantees the existence of a finitely supported solution of width $m \leq n-(d+1)$ to the variational problem. However, the finite-dimensional network learning problem is equivalent to the variational problem only when $m \geq n-(d+1)$. See their Theorem 5 and Section III.B. Therefore, their results still fall within the underparametrized regime and do not fully characterize the influence of the network width. By contrast, we provide a direct attack on the finite-dimensional network learning problem (9) and allow the network width m to vary freely.

3.2 Equivalence to Ridge Regression

In this and the next subsections, we explore some useful reformulations of the optimization problem (9), which allow the scaled variation regularizer, when coupled with the ReLU function, to inherit some crucial properties from ridge regression (Hoerl, 2020) and the group lasso (Yuan and Lin, 2006), two familiar forms of regularization in statistics. We start by recasting (9) as the ℓ_2-regularized ERM problem

$$\widehat{\boldsymbol{\theta}}_{\ell_2} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_m}{\arg\min}\left\{\frac{1}{2n}\sum_{i=1}^{n}\big(y_i - g(\mathbf{x}_i;\boldsymbol{\theta})\big)^2 + \frac{\lambda}{2}\sum_{k=1}^{m}(a_k^2 + \|\mathbf{w}_k\|_2^2)\right\}. \tag{10}$$

To see this, consider the reparametrization $\widetilde{\boldsymbol{\theta}} = \mathcal{T}_1(\boldsymbol{\theta})$ defined by

$$\widetilde{a}_k = a_k\sqrt{\frac{\|\mathbf{w}_k\|_2}{|a_k|}}, \qquad \widetilde{\mathbf{w}}_k = \mathbf{w}_k\sqrt{\frac{|a_k|}{\|\mathbf{w}_k\|_2}}$$

if $|a_k|\|\mathbf{w}_k\|_2 \neq 0$, and $(\widetilde{a}_k, \widetilde{\mathbf{w}}_k^T) = \mathbf{0}$ otherwise. After the reparametrization, we have $|\widetilde{a}_k| = \|\widetilde{\mathbf{w}}_k\|_2$ and the regularizer becomes

$$\sum_{k=1}^{m}|\widetilde{a}_k|\|\widetilde{\mathbf{w}}_k\|_2 = \frac{1}{2}\sum_{k=1}^{m}(\widetilde{a}_k^2 + \|\widetilde{\mathbf{w}}_k\|_2^2).$$

Meanwhile, the positive homogeneity of the ReLU function implies that $a_k\sigma((\mathbf{x}_k^T,1)\mathbf{w}_k) = \widetilde{a}_k\sigma((\mathbf{x}_k^T,1)\widetilde{\mathbf{w}}_k)$, so that the network function is invariant under the reparametrization. Note further that any solution $\widehat{\boldsymbol{\theta}}_{\ell_2}$ to the problem (10) must satisfy $\widehat{\boldsymbol{\theta}}_{\ell_2} = \mathcal{T}_1(\widehat{\boldsymbol{\theta}}_{\ell_2})$, because otherwise it could be improved by a rescaling. Using these facts, we obtain the following equivalence result.

Proposition 3 *Any solution $\widehat{\boldsymbol{\theta}}_{\ell_2}$ to the optimization problem* (10) *is a solution to the problem* (9)*. Conversely, if $\widehat{\boldsymbol{\theta}}$ is a solution to the optimization problem* (9)*, then $\mathcal{T}_1(\widehat{\boldsymbol{\theta}})$ is a solution to the problem* (10)*.*

Proposition 3 says that the solutions to the ℓ_2-regularized problem lie on a submanifold of the solution manifold of the original problem that is invariant under the reparametrization $\mathcal{T}_1$. What is the implication of this equivalence for neural network training dynamics with, for example, gradient descent? The following result assures us that the gradient flow trajectories for the two problems are indeed identical when initialized with a reparametrization $\mathcal{T}_1$.

Proposition 4 *Consider the gradient flow for the optimization problem* (9) *defined by*

$$\frac{d}{dt}\boldsymbol{\theta}(t) = -\nabla_{\boldsymbol{\theta}} J_n(\boldsymbol{\theta}(t);\lambda)$$

and for the problem (10) *defined similarly, both initialized at* $\boldsymbol{\theta}(0) = \mathcal{T}_1(\boldsymbol{\theta}_0)$ *for an arbitrary* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_m$*. Then the trajectories of the two gradient flows coincide.*

Although the gradient flow characterization in Proposition 4 is not explicitly used in the proofs of our main results in Section 4, it entails that the optimization problem (9), as in ridge regression, does not produce a sparse solution. This helps to ensure that our improved complexity control for the overparametrized regime is not, even implicitly, due to sparsity, and in this sense that the overparametrization is intrinsic.

Observations similar to Proposition 3 have been noted in slightly different forms by, for example, Neyshabur et al. (2015b, Theorem 1) and Parhi and Nowak (2021, Theorem 8). Initialization with the reparametrization $\mathcal{T}_1$ and its stationarity along the gradient flow have been exploited by Dou and Liang (2021) for studying the ℓ_2-regularized ERM problem, where it is referred to as a "balanced condition." In general, ℓ_2 regularization does not induce entrywise sparsity in the parameters, but see Srebro et al. (2004) for an unusual example where it does induce sparsity in spectral structures. Moreover, several implicit regularization strategies for deep learning, such as noise injection and early stopping, have been shown to be equivalent to ℓ_2 regularization (Bishop, 1995; Sjöberg and Ljung, 1995), which may help bridge the gap between our method and implicit regularization.

3.3 Connection to the Group Lasso

One major obstacle in analyzing the generalization performance of neural networks is the excessive redundancy and nonidentifiability of the network parameters under the usual nonconvex formulation. The ReLU activation function, on the other hand, is simple enough in that it reduces to a linear function once the sign of $\mathbf{v}_k^T\mathbf{x} + b_k$ is fixed. This naturally suggests a partitioning of the parameter space $\mathbb{R}^{d+1}$ for $\mathbf{w}$ by certain hyperplanes into regions within which the signs of $\mathbf{x}_i^T\mathbf{v} + b$ are all determined. The partition will then allow us to reveal a strong symmetry in the estimated network parameters and recast the optimization problem (9) in a group lasso form, which will be convenient for the derivation of generalization properties in the next section.

Specifically, denote by $\mathbf{X} = ((\mathbf{x}_1^T, 1)^T, \ldots, (\mathbf{x}_n^T, 1)^T)^T$ the $n \times (d+1)$ design matrix, and $\mathbf{D} = \operatorname{diag}(I(\mathbf{X}\mathbf{w} \geq 0))$ the diagonal indicator matrix for the positivity of $\mathbf{X}\mathbf{w}$. Consider the hyperplanes in $\mathbb{R}^{d+1}$ passing through the origin and orthogonal to $\mathbf{x}_i$, defined by $\mathbf{x}_i^T\mathbf{v}+b = 0$. These n hyperplanes divide the parameter space $\mathbb{R}^{d+1}$ into finitely many regions, denoted by $R_1, \ldots, R_p$, such that $\mathbf{D}$ stays constant over (the interior of) each R_j. It is well known (Cover, 1965, Theorem 2) that the number of these regions

$$p \leq 2\sum_{j=0}^{d} \binom{n-1}{j} \leq 2n^d, \tag{11}$$

where the first upper bound is sharp when $\mathbf{X}$ has full rank. Taking into account the sign of a, we thus partition the parameter space $\mathbb{R}^{d+2}$ for $(a, \mathbf{w}^T)^T$ into $2p$ regions

$$Q_j = [0,\infty) \times R_j, \quad Q_{p+j} = (-\infty, 0) \times R_j, \quad j = 1, \ldots, p,$$

and define $\mathbf{D}_{p+j} = -\mathbf{D}_j$. Clearly, R_j and Q_j are convex cones. The linearity of the ReLU function over each Q_j and the optimality of $\widehat{\boldsymbol{\theta}}$ entail the following collinearity property.

Proposition 5 *For any solution $\widehat{\boldsymbol{\theta}} = (\widehat{a}_1, \ldots, \widehat{a}_m, \widehat{\mathbf{w}}_1^T, \ldots, \widehat{\mathbf{w}}_m^T)^T$ to the optimization problem* (9)*, if $(\widehat{a}_k, \widehat{\mathbf{w}}_k^T)^T$ and $(\widehat{a}_\ell, \widehat{\mathbf{w}}_\ell^T)^T$ lie in the interior of the same cone Q_j, then $\widehat{\mathbf{w}}_k$ and $\widehat{\mathbf{w}}_\ell$ must be collinear, that is, $\widehat{\mathbf{w}}_k = c_0 \widehat{\mathbf{w}}_\ell$ for some constant $c_0 > 0$.*

Define $S_j = \{1 \le k \le m : (a_k, \mathbf{w}_k^T)^T \in Q_j\}$, $s_j = 1$, and $s_{p+j} = -1$ for $j = 1, \ldots, p$. To understand why the collinearity must hold, note that the "conewise collinearization" $\widetilde{\boldsymbol{\theta}} = \mathcal{T}_2(\boldsymbol{\theta})$ defined by

$$\widetilde{a}_k = s_j, \quad \widetilde{\mathbf{w}}_k = \frac{1}{|S_j|} \sum_{\ell \in S_j} |a_\ell| \mathbf{w}_\ell, \quad k \in S_j \tag{12}$$

does not change the value of the network function on the training sample, but would yield a smaller scaled variation norm by the triangle inequality if the network weights in Q_j were not all collinear. A simple example with $d = 1$ and $n = 2$ illustrating the reparametrization $\mathcal{T}_2$ is given in Figure 2. Proposition 5 provides a useful geometric insight into the regularization effect of scaled variation norm: it favors the most symmetric (yet not parsimonious) representation among many equivalent parametrizations within the same cone.

The parameter redundancy suggested by Proposition 5 motivates us to collect the network weights falling within the same cone and define the aggregated parameters $\mathbf{B}(\boldsymbol{\theta}) = (\boldsymbol{\beta}_1(\boldsymbol{\theta}), \ldots, \boldsymbol{\beta}_{2p}(\boldsymbol{\theta}))$ with

$$\boldsymbol{\beta}_j(\boldsymbol{\theta}) = \sum_{k \in S_j} |a_k| \mathbf{w}_k.$$

With this new set of parameters, the network function on the training sample can be written in the linear form

$$\sum_{k=1}^m a_k \sigma(\mathbf{X}\mathbf{w}_k) = \sum_{j=1}^{2p} \mathbf{D}_j \mathbf{X} \boldsymbol{\beta}_j(\boldsymbol{\theta}), \tag{13}$$

where $\sigma(\cdot)$ applies componentwise. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}_m$, the triangle inequality implies that

$$\|\mathbf{B}(\boldsymbol{\theta})\|_{2,1} = \sum_{j=1}^{2p} \|\boldsymbol{\beta}_j(\boldsymbol{\theta})\|_2 \le \sum_{j=1}^{2p} \sum_{k \in S_j} |a_k| \|\mathbf{w}_k\|_2 = \nu(\boldsymbol{\theta}),$$

where the equality holds under the reparametrization $\mathcal{T}_2$. In particular, since the estimator $\widehat{\boldsymbol{\theta}}$ satisfies the collinearity property, we can replace $\nu(\widehat{\boldsymbol{\theta}})$ by $\|\mathbf{B}(\widehat{\boldsymbol{\theta}})\|_{2,1}$ and reformulate (9) as a group lasso problem. Denote by $\mathbf{y} = (y_1, \ldots, y_n)^T$ the response vector. We summarize the above discussion in the following proposition.

Proposition 6 *For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}_m$, the reparametrization $\widetilde{\boldsymbol{\theta}} = \mathcal{T}_2(\boldsymbol{\theta})$ defined in* (12) *satisfies $g(\mathbf{x}_i; \widetilde{\boldsymbol{\theta}}) = g(\mathbf{x}_i; \boldsymbol{\theta})$ for $i = 1, \ldots, n$ and $\|\mathbf{B}(\widetilde{\boldsymbol{\theta}})\|_{2,1} = \nu(\widetilde{\boldsymbol{\theta}}) \le \nu(\boldsymbol{\theta})$. Furthermore, the solution $\widehat{\boldsymbol{\theta}}$ to the optimization problem* (9) *satisfies*

$$J_n(\widehat{\boldsymbol{\theta}}; \lambda) = \frac{1}{2n} \left\| \mathbf{y} - \sum_{j=1}^{2p} \mathbf{D}_j \mathbf{X} \boldsymbol{\beta}_j(\widehat{\boldsymbol{\theta}}) \right\|_2^2 + \lambda \|\mathbf{B}(\widehat{\boldsymbol{\theta}})\|_{2,1}.$$

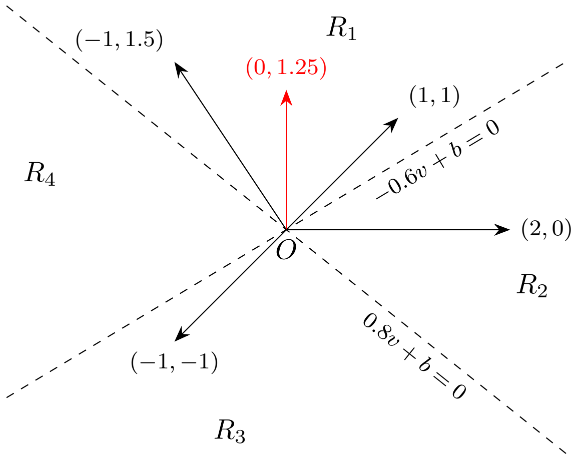


Figure 2: A simple example to illustrate the "conewise collinearization" reparametrization $\mathcal{T}_2$. Consider $n = 2$ data points with $x_1 = 0.8$ and $x_2 = -0.6$, and a two-layer ReLU network $g(x; \boldsymbol{\theta}) = \sum_{k=1}^{4} \sigma(v_k x + b_k)$ with $(v_1, b_1) = (-1, 1.5)$, $(v_2, b_2) = (1, 1)$, $(v_3, b_3) = (2, 0)$, and $(v_4, b_4) = (-1, -1)$. The parameter space $\mathbb{R}^2$ for (v, b) is partitioned by two straight lines $vx_1 + b = 0$ and $vx_2 + b = 0$ into four cones (R_1, R_2, R_3, R_4). Parameters falling within each cone share the same pattern of activations on the data points, and become collinear after reparametrization. For example, the two hidden units with parameters $(v_1, b_1) = (-1, 1.5)$ and $(v_2, b_2) = (1, 1)$ in R_1 are activated for both x_1 and x_2, which are collinearized to $(\widetilde{v}_1, \widetilde{b}_1) = (\widetilde{v}_2, \widetilde{b}_2) = (0, 1.25)$ after reparametrization. Moreover, $\|(\widetilde{v}_1, \widetilde{b}_1)\|_2 + \|(\widetilde{v}_2, \widetilde{b}_2)\|_2 \leq \|(v_1, b_1)\|_2 + \|(v_2, b_2)\|_2$ by the triangle inequality.

The geometric insight and group lasso formulation in Propositions 5 and 6 allow us to borrow techniques from high-dimensional statistics for deriving sharp generalization bounds. In particular, it removes parameter redundancy in the original neural network formulation, leading to a better complexity control and improved generalization bounds in the over-parametrized regime, as will be shown in Theorems 7 and 8. We emphasize, however, that the parameter space partition and the resulting group structure are data-adaptive and not known a priori. Hence, despite the connection to the group lasso, two-layer ReLU networks are radically different from linear models and hold the potential for better generalization.

Similar connections between ℓ_2-regularized two-layer ReLU networks and the group lasso have been explored by Pilanci and Ergen (2020) and Wang et al. (2022) from a purely optimization standpoint. A complete equivalence result, however, requires m to be sufficiently large; see Theorem 1 of Pilanci and Ergen (2020). Our key observation is that for our purposes it suffices to have the weaker result of Proposition 6, which places no restriction on the minimum network width.

3.4 Numerical Experiments

We have shown in theory that the global solution to the regularized problem (9) has the nice collinearity property: the first-layer parameters tend to cluster into groups with the same directions. We now present some numerical experiments to verify that this clustering effect does occur in practice. To this end, we generated a training sample of size $n = 1000$ from the model $y_i = \cos(2\pi x_i) + \varepsilon_i$, with x_i sampled uniformly from $[-1, 1]$ and ε_i drawn from $N(0, 0.01)$. With different regularization strengths specified by $\lambda = 0.01$, 0.001, and 0 (no regularization), we trained a two-layer ReLU network of width $m = 100$ using full-batch Adam with a cosine annealing learning rate scheduler in PyTorch. The network was randomly initialized using PyTorch's default method.

The scaled first-layer parameters $(|a_k|v_k, |a_k|b_k)$, recorded at 50, 500, and 10,000 epochs, are represented as points in a plane and plotted in polar coordinates in Figure 3. We see that, under relatively strong regularization ($\lambda = 0.01$), the neurons tend to concentrate around a few directions after 500 epochs and become perfectly aligned within each group after 10,000 epochs. The clustering effect diminishes as λ decreases. When no regularization is imposed ($\lambda = 0$), the final parameters are roughly dispersed over two fairly wide sectors and show no clear clustering patterns. Interestingly, however, the early stopped parameters (e.g., at 500 epochs) exhibit a clustering effect similar to that observed with proper regularization. This is not surprising in view of the equivalence of both scaled variation regularization and early stopping to ℓ_2 regularization (Sjöberg and Ljung, 1995).

A similar phenomenon for neural network training, even with plain gradient descent, has been described in the literature under various names such as quantization (Maennel et al., 2018), alignment (Ji and Telgarsky, 2019), and condensation (Luo et al., 2021). Our theory and experiments show that this phenomenon is well captured by our regularized problem, which allows us to exploit this important property in generalization analysis.

4 Main Results

In this section we establish statistical guarantees for two-layer ReLU networks. In Section 4.1 we present nonasymptotic bounds on the prediction error of the regularized network estimator, and in Section 4.2 show that they are minimax optimal.

4.1 Nonasymptotic Generalization Bounds

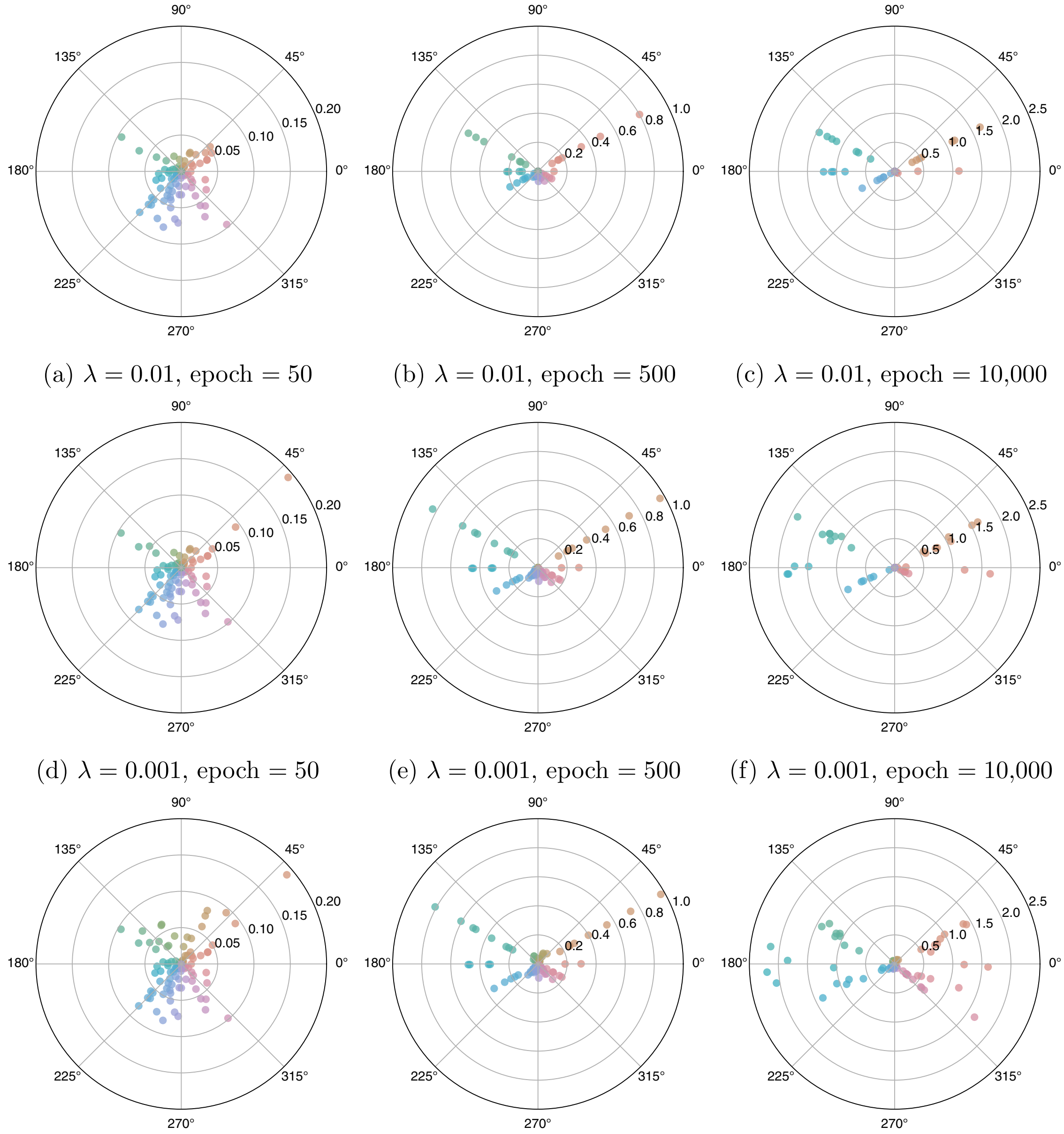
For the nonparametric regression model (1) and the regularized network estimator $g(\cdot; \widehat{\boldsymbol{\theta}})$ defined by (9), we are interested in bounding the empirical error

$$\|g(\cdot; \widehat{\boldsymbol{\theta}}) - f^*\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}\big(g(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}) - f^*(\mathbf{x}_i)\big)^2$$

in the fixed design case, and the prediction (or generalization) error

$$\|g(\cdot; \widehat{\boldsymbol{\theta}}) - f^*\|_2^2 = \mathbb{E}_{\mathbf{x}\sim\mu}\big(g(\mathbf{x}; \widehat{\boldsymbol{\theta}}) - f^*(\mathbf{x})\big)^2$$

in the random design case. Our main techniques for proving the nonasymptotic bounds are inspired by and synthesize those in previous work on high-dimensional linear models and two-layer neural networks. We first note that the technical arguments best suited to the underparametrized and overparametrized regimes may be rather different. For underparametrized networks, complexity control via metric entropy (e.g., Barron, 1994; Parhi and Nowak, 2023) can be effective and give sharp bounds. Moving into the overparametrized



(a) $\lambda = 0.01$, epoch = 50

(b) $\lambda = 0.01$, epoch = 500

(c) $\lambda = 0.01$, epoch = 10,000

(d) $\lambda = 0.001$, epoch = 50

(e) $\lambda = 0.001$, epoch = 500

(f) $\lambda = 0.001$, epoch = 10,000

(g) $\lambda = 0$, epoch = 50

(h) $\lambda = 0$, epoch = 500

(i) $\lambda = 0$, epoch = 10,000

Figure 3: Numerical experiments on simulated data to validate the regularization effect of scaled variation norm. Randomly initialized two-layer ReLU networks of width $m = 100$ were trained on a sample with $d = 1$ and $n = 1000$ using full-batch Adam with different regularization strengths ($\lambda = 0.1$, 0.01, and 0). The scaled first-layer parameters $(|a_k|v_k, |a_k|b_k)$ at 50, 500, and 10,000 epochs are plotted in polar coordinates.

regime, however, entropy-based bounds tend to be too loose and pessimistic since they do not take into account the parameter redundancy growing with the network width. We therefore turn to the group lasso formulation outlined in Section 3.3 and borrow ideas from (group) ℓ_1-regularized linear regression and norm-based complexity control. Our first result concerns the empirical error of the regularized network estimator.

Theorem 7 *Under Conditions (C1) and (C3) and the assumption that* $\max_i \|\mathbf{x}_i\|_2 \leq 1$, *the regularized network estimator* $g(\cdot; \widehat{\boldsymbol{\theta}})$ *with* $\lambda = C_1 \sigma_\varepsilon \sqrt{d \log n / n}$ *satisfies*

$$\|g(\cdot; \widehat{\boldsymbol{\theta}}) - f^*\|_n^2 \leq C \left\{ \|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2) \sqrt{\frac{d \log n}{n}} \right\} \tag{14}$$

with probability at least $1 - O(n^{-C_2})$, *and*

$$\mathbb{E}\|g(\cdot; \widehat{\boldsymbol{\theta}}) - f^*\|_n^2 \leq C \left\{ \|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2) \sqrt{\frac{d \log n}{n}} \right\}, \tag{15}$$

for some constants $C_1, C_2, C > 0$.

Throughout this section, the constants are independent of m and n, but may depend on d, M, and σ_ε; we have suppressed the dependence for simplicity, which can be made explicit by inspecting our proofs. Our technique for proving Theorem 7 differs from the standard group lasso theory for sparse linear models in several aspects. First, the linear model here is formed by a reparametrization of the two-layer ReLU network, which is only an approximation to the target function. Second, there is no assumption that this linear model is sparse or its design matrix satisfies a compatibility or restricted eigenvalue condition. Finally, the reparametrization $\mathcal{T}_2$ in (12) is data-dependent, and hence the group lasso formulation does not uniformly hold for all samples and parameters.

In the proof of Theorem 7, we address these challenges as follows: (i) To deal with the approximate nature of the linear model, we focus on the $L_\infty(\mathbb{B}^d)$-approximation of f^* from Theorem 2 and incorporate an extra term representing the approximation error throughout our analysis. (ii) Without sparsity or eigenvalue assumptions, the parameters of the linear model are unidentifiable; however, we can still proceed with the analysis of the prediction error at the price of a slower convergence rate than the usual $O(n^{-1})$ rate. This extends the classical persistence results for the lasso without assuming sparsity; see Greenshtein (2006, Corollary 2) and Bühlmann and van de Geer (2011, Corollary 6.1). (iii) Note that it suffices for the group lasso formulation to hold on the training sample for both the optimal solution $\widehat{\boldsymbol{\theta}}$ and (a reparametrization of) the best-approximating parameters $\boldsymbol{\theta}^*$. While the former is an exact equality, the latter comes from properties of the reparametrization $\mathcal{T}_2$ for all $\boldsymbol{\theta}$, as readily shown by Proposition 6.

The error bounds in Theorem 7 decompose into the approximation error that arises from using a finite-width neural network to approximate the nonparametric model (1), and the estimation error that accounts for the variability in estimating the finite-width network. The most surprising fact about this decomposition is that there is *no* trade-off between the two terms: as the network width m increases, the approximation error always decreases, while the estimation error remains constant. To appreciate why this is possible, note first that the estimation error scales as $O(\sqrt{\log p/n})$ as a consequence of the lack of parameter identifiability. Moreover, the *effective* dimension p is bounded by $O(n^d)$ from (11), which does not depend on the network width m. In fact, when the design matrix $\mathbf{X}$ is of rank $r < n$, one can further replace d by r (Cover, 1965). In other words, no matter how large m grows, the number of distinct (nonparallel) features extracted by the first layer of the network is finite, leading to an upper bound for the estimation error.

More insights can be gained by decomposing the error bounds in Theorem 7 in another way. Note that both $\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d}$ and $\|f^*\|_{\mathcal{S}}^2\sqrt{d\log n/n}$ contribute to the bias, while $\sigma_\varepsilon^2\sqrt{d\log n/n}$ constitutes the variance. Thus, the bias and variance exhibit a similar trend to that of the approximation–estimation decomposition: the bias decreases and the variance stays constant as m increases. Moreover, the estimation bias and estimation variance are balanced by the optimal choice of λ, thereby achieving an optimal estimation error. Overall, our result extends beyond the classical bias–variance trade-off and demonstrates the virtues of overparametrization in two-layer neural networks.

Combining Theorem 7 with a maximal inequality, we obtain similar bounds on the prediction error of the regularized network estimator.

Theorem 8 *Under Conditions (C1)–(C3), the regularized network estimator* $g(\cdot;\widehat{\boldsymbol{\theta}})$ *with* $\lambda = \lambda_1 \equiv C_1 \max(\|f^*\|_{\mathcal{S}} m^{-(d+3)/d}, \sigma_\varepsilon\sqrt{d\log n/n})$ *satisfies*

$$\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_2^2 \le C\left\{\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2)\sqrt{\frac{d\log n}{n}}\right\} \tag{16}$$

with probability at least $1 - O(n^{-C_2})$*, and*

$$\mathbb{E}\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_2^2 \le C\left\{\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2)\sqrt{\frac{d\log n}{n}}\right\}, \tag{17}$$

for some constants $C_1, C_2, C > 0$ *and sufficiently large* n*.*

It is worthwhile to compare our results with those in the literature on overparametrized two-layer ReLU networks. Recent research has focused on the neural tangent kernel regime and showed that sufficiently wide two-layer ReLU networks trained by gradient descent with random initialization achieve a generalization error of $O(n^{-1/2})$ up to logarithmic factors; see, for example, Arora et al. (2019), E et al. (2020), and Ji and Telgarsky (2020). While these results deliver roughly the same rates as ours, the target functions they considered fall in a certain reproducing kernel Hilbert space, which constitutes only a small subset of our target function space. In addition, our analysis concerns any global optimum rather than the solution obtained by a specific optimization algorithm.

E et al. (2019) considered explicit regularization for two-layer ReLU networks and obtained generalization bounds of $O(m^{-1} + n^{-1/2})$ up to logarithmic factors, which are of a similar nature to ours. However, several differences are notable. First, they employed the ℓ_1 path norm, which penalizes on the ℓ_1-norm of the first-layer weights and promotes sparsity. Accordingly, they considered the so-called Barron space

$$\mathcal{B}_2 = \left\{\mathbf{x} \mapsto \int_{\mathbb{S}_1^{d-1}} a(\mathbf{v})\sigma(\mathbf{v}^T\mathbf{x})\,d\rho(\mathbf{v}) : \int_{\mathbb{S}_1^{d-1}} |a(\mathbf{v})|^2\,d\rho(\mathbf{v}) < \infty\right\},$$

where $\mathbb{S}_1^{d-1}$ is the ℓ_1 unit sphere in $\mathbb{R}^d$. To compare with our definition of $\mathcal{G}$ in (5), let $d\alpha_\rho(\mathbf{w}) = a(\mathbf{v})I(\mathbf{v} \in \mathbb{S}_1^{d-1}, b = 0)\, d\rho(\mathbf{v})$, where $I(\cdot)$ is the indicator function. Then

$$\int_{\mathbb{R}^{d+1}} \|\mathbf{v}\|_2 \, d|\alpha_\rho|(\mathbf{w}) \le \sqrt{d} \int_{\mathbb{S}_1^{d-1}} |a(\mathbf{v})| \, d\rho(\mathbf{v}) \le \sqrt{d} \biggl(\int_{\mathbb{S}_1^{d-1}} |a(\mathbf{v})|^2 \, d\rho(\mathbf{v}) \biggr)^{1/2} < \infty$$

by the Cauchy–Schwarz inequality. Thus, we see that $\mathcal{B}_2$ is a subset of our target function space $\mathcal{G}$. Furthermore, they resorted to a truncated risk to deal with the noisy case, which introduces some technicalities that seem unnecessary.

The group lasso approach and the size-independent upper bound (11) for p, albeit effective in the overparametrized regime, tend to overestimate the variance for sufficiently narrow networks. In this case, a standard metric entropy argument may be more appropriate and give a sharper estimate of the variance that increases with the network width. Adapting this argument to our regularization problem yields the following result, which demonstrates a classical U-shaped risk curve.

Theorem 9 *Under Conditions (C1)–(C3), the regularized network estimator* $g(\cdot; \widehat{\boldsymbol{\theta}})$ *with* $\lambda = \lambda_2 \equiv C_1 \max(\|f^*\|_{\mathcal{S}} m^{-(d+3)/d}, \sigma_\varepsilon m d \log n / n)$ *satisfies*

$$\|g(\cdot; \widehat{\boldsymbol{\theta}}) - f^*\|_2^2 \le C \biggl\{ \|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2) \frac{md \log n}{n} \biggr\}$$

with probability at least $1 - O(n^{-C_2})$ *for some constants* $C_1, C_2, C > 0$.

The proof technique used for Theorem 9 differs substantially from those in previous work, since we are analyzing a penalized rather than constrained problem and do not impose any boundedness constraints on the network function or parameters; cf. Schmidt-Hieber (2020), Farrell et al. (2021), and Parhi and Nowak (2023).

Finally, noting that the ranges of allowable m in Theorems 8 and 9 partially overlap, we put them together to obtain a complete picture of the generalization behavior of two-layer ReLU networks, as stated in the following encompassing result.

Theorem 10 *Under Conditions (C1)–(C3), the regularized network estimator* $g(\cdot; \widehat{\boldsymbol{\theta}})$ *with* $\lambda = \min(\lambda_1, \lambda_2)$, *where* λ_1 *and* λ_2 *are defined in Theorems 8 and 9, respectively, satisfies*

$$\|g(\cdot; \widehat{\boldsymbol{\theta}}) - f^*\|_2^2 \le C \biggl\{ \|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2) \min\biggl(\sqrt{\frac{d \log n}{n}}, \frac{md \log n}{n} \biggr) \biggr\}$$

with probability at least $1 - O(n^{-C_1})$ *for some constants* $C_1, C > 0$ *and sufficiently large* n.

The implications of Theorem 10 have been discussed in the Introduction. In particular, it gives rise to the double descent risk curve illustrated in Figure 1 and provides a simple yet appealing explanation for the curious phenomenon. In the underparametrized regime, the network estimator behaves as the usual nonparametric methods, with the network width m controlling the trade-off between approximation and estimation. A too small or too large m will result in an inferior performance, and a narrow valley around $m_0 \asymp (n/(d \log n))^{d/(2d+3)}$ lies in between. As m continues to increase and exceeds some threshold

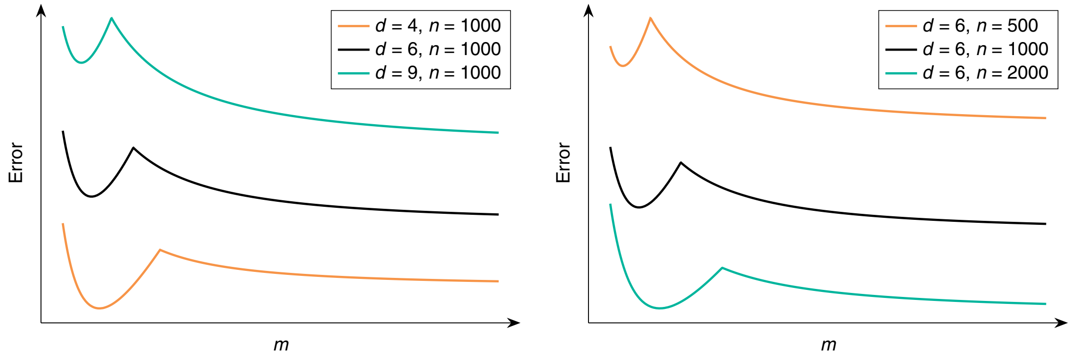


Figure 4: Risk curves from Theorem 10 with $\|f^*\|_{\mathcal{S}}^2/\sigma_\varepsilon^2 = 1$, varying d, and varying n.

$m_1 \asymp \sqrt{n/(d\log n)}$, the intrinsic model complexity and hence the estimation error of the network estimator become saturated and remain constant, while the approximation error diminishes consistently. This leads to a second, flat valley extending toward infinity.

Comparisons between the two valleys yield further insights. Asymptotically, the convergence rate of the first valley or underparametrized minimum risk, $O((d\log n/n)^{(d+3)/(2d+3)})$, is slightly smaller than that of the second valley or overparametrized minimum risk, $O(\sqrt{d\log n/n})$. In finite samples, however, this comparison can be reversed. A little algebra shows that the second valley is lower than the first whenever

$$\kappa \equiv \frac{\|f^*\|_{\mathcal{S}}^2}{\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2} > \left(\frac{1}{2}\right)^{(2d+3)/d}\left(\frac{n}{d\log n}\right)^{3/(2d)}. \tag{18}$$

When $d \gg \log n$, the above condition approximately becomes $\kappa > 1/4$, or the signal-to-noise ratio $\|f^*\|_{\mathcal{S}}^2/\sigma_\varepsilon^2 = \kappa/(1-\kappa) > 1/3$. This makes intuitive sense since the reduction in approximation error plays a more important role when the signal is stronger. An example of the risk curve with $\kappa = 1/2$, $d = 6$, and $n = 1000$ was given in Figure 1. Additional settings with varying d and n are shown in Figure 4. The advantages of overparametrization are more visible when d is relatively large or n is relatively small. From the practitioner's perspective, the overparametrized regime is also more attractive in that it provides an infinitely wide sweet spot that avoids the choice of an optimal network width.

A related phenomenon was observed by Hastie et al. (2022) for minimum ℓ_2-norm interpolators in linear regression, where the risk curve oscillates with varying signal-to-noise ratio. In their case, however, the infimum of the second descent can occur at a finite point and the risk curve can approach the null risk from below when the signal-to-noise ratio is high. This is in sharp contrast to our case, where the risk curve always decreases, and the signal-to-noise ratio only affects the comparison between the two infima.

Now we are ready to revisit the bias–variance trade-off in light of Theorem 10. As the sum of model bias and estimation bias, the total bias

$$\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + \|f^*\|_{\mathcal{S}}^2 \min\left(\sqrt{\frac{d\log n}{n}}, \frac{md\log n}{n}\right)$$

inherits the double descent shape from the risk curve, first decreasing and then increasing in the underparametrized regime and decreasing again in the overparametrized regime. On the other hand, the variance $\sigma_\varepsilon^2 \min(\sqrt{d \log n/n}, md \log n/n)$ exhibits the same trend as the estimation error, first increasing and then staying constant. This observation is radically different from those previously suggested (e.g., Hastie et al., 2022; Mei and Montanari, 2022; Yang et al., 2020) where the double descent curve is mainly due to a unimodal variance.

Finally, we note that it would be difficult to directly assess the numerical performance of the global solution to the nonconvex problem (9). Any practical method for obtaining an approximate solution would inevitably involve some forms of implicit regularization such as random initialization and early stopping. Most of the existing numerical studies on the double descent phenomenon (e.g., Belkin et al., 2019; Nakkiran et al., 2021a) employed ad hoc procedures to control for the randomness and artifacts of implicit regularization. Our limited numerical exploration suggests that these factors may interfere with the risk curve and lead to unstable or misleading results. A further study needs to incorporate the influence of nonconvex optimization, which we hope to report elsewhere.

4.2 Minimax Lower Bounds

We have revealed that the risk curve of our estimator has two valleys. The convergence rate of the first valley is known to be minimax optimal over the function class $\mathcal{G}_M$ (Parhi and Nowak, 2023). In fact, the underparametrized result (Theorem 9) relies crucially on the assumption that M is finite; otherwise, the entropy calculations may be affected. In this subsection, we investigate the optimality of the second valley. Although it cannot be optimal over $\mathcal{G}_M$, we will show that it is minimax optimal over the larger function class $\mathcal{G}$.

Recall that minimax lower bounds characterize the best worst-case performance achievable by any procedure, which is an intrinsic property of the target function class; for reviews, see Tsybakov (2009, Chapter 2) and Wainwright (2019, Chapter 15). To gain intuition for the minimax rate in our case, we consider the RKHS defined in (6). If the target function $f^* \in \mathcal{H}_{\rho^*}$ for some known ρ^*, then the problem of recovering f^* reduces to kernel ridge regression. It was shown by Caponnetto and De Vito (2007) that the minimax optimal rate for learning functions in an RKHS is $n^{-\gamma/(\gamma+1)}$ when the jth eigenvalue of the kernel decays at the rate of $j^{-\gamma}$ for $\gamma > 1$. Note that $\mathcal{H}_\rho \subset \mathcal{G}$ for all ρ, and the minimax rate approaches $n^{-1/2}$ as $\gamma \to 1$. This heuristic argument is formalized in the following result.

Theorem 11 *Under Conditions (C2) and (C3) with μ being the uniform distribution on $\mathbb{S}^{d-1}$, there exists a constant $C > 0$ such that*

$$\inf_{\widehat{f}} \sup_{f^* \in \mathcal{G}(\mathbb{S}^{d-1})} \mathbb{E}\|\widehat{f} - f^*\|_2^2 \geq C\sigma_\varepsilon \sqrt{\frac{\log n}{n}},$$

where the infimum is taken over all estimators.

This result says that the upper bounds on the overparametrized minimum risk in Theorems 8 and 10 are sharp. Without requiring the existence of a finite M, these bounds are essentially unimprovable, which corroborates the effectiveness of overparametrized two-layer ReLU networks.

5 Suboptimality of Random Feature Models

In the previous section we have seen that two-layer neural networks are minimax optimal for learning target functions in $\mathcal{G}$ or $\mathcal{G}_M$. But can the optimal rates be achieved by some other, computationally simpler, estimators? One such method worthy of investigation are random feature models (Rahimi and Recht, 2007), which provide a stochastic approximation to kernel methods by first mapping the input into a randomized feature space and then applying standard linear methods. They can be viewed as two-layer neural networks with random first-layer weights and, as such, often serve as a prototype for studying the generalization behavior of realistic neural networks. For example, Mei and Montanari (2022) computed the prediction error of random feature regression that recovers the double descent curve in the asymptotic regime where $m, n, d \to \infty$ with $m \asymp n \asymp d$. We now show, however, that random feature models are not sufficient to explain the generalization power of fully trained two-layer networks by proving that they are suboptimal over our target function space.

Specifically, we consider the random feature model

$$h_{\rho_0}(\mathbf{x}; \mathbf{a}) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} a_k \sigma(\mathbf{v}_k^T \mathbf{x} + b_k),$$

where $\mathbf{w}_k = (\mathbf{v}_k^T, b_k)^T \sim \rho_0$ independently for some fixed ρ_0 and $\mathbf{a} = (a_1, \ldots, a_m)^T$ is the vector of parameters to be estimated. Minimizing the ℓ_2-regularized empirical risk

$$\frac{1}{2n} \sum_{k=1}^{m} \big(y_i - h_{\rho_0}(\mathbf{x}_i; \mathbf{a})\big)^2 + \frac{\lambda}{2} \|\mathbf{a}\|_2^2$$

gives the solution $\widehat{\mathbf{a}}(\lambda) = (\mathbf{K} + n\lambda \mathbf{I}_n)^{-1}\mathbf{y}$ and the random feature estimator $h_{\rho_0}(\cdot; \widehat{\mathbf{a}}(\lambda))$, where $\mathbf{K} = (K_{ij}) \in \mathbb{R}^{n \times n}$ is the kernel matrix with entries

$$K_{ij} = \frac{1}{m} \sum_{k=1}^{m} \sigma(\mathbf{v}_k^T \mathbf{x}_i + b_k) \sigma(\mathbf{v}_k^T \mathbf{x}_j + b_k).$$

In fact, $K_{ij} \to H_{\rho_0}(\mathbf{x}_i, \mathbf{x}_j)$ as $m \to \infty$ for the kernel H_ρ defined in (7). To demonstrate the suboptimality of random feature models, it suffices to establish a performance lower bound that is larger than the upper bounds for two-layer neural networks. The following result gives the desired lower bound on the worst-case performance of optimally tuned random feature estimators.

Proposition 12 *Under Conditions (C2) and (C3) with μ being the uniform distribution on $[-1/\sqrt{d}, 1/\sqrt{d}]^d$, there exists a universal constant $C > 0$ such that*

$$\sup_{f^* \in \mathcal{G}_M} \inf_{\lambda > 0} \mathbb{E}\|h_{\rho_0}(\cdot; \widehat{\mathbf{a}}(\lambda)) - f^*\|_2 \geq \frac{CM}{d\{\min(m, n)\}^{2/d}}.$$

The proof of Proposition 12 builds on the argument of Barron (1993, Theorem 6) for approximation by linear supspaces with fixed basis functions. Similar lower bounds have been obtained by E et al. (2020) for random feature models trained by gradient descent with noiseless data. Compared with the underparametrized rate $O(n^{-(d+3)/(2d+3)})$ and overparametrized rate $O(n^{-1/2})$ from Theorem 10, the rate in Proposition 12 is much slower for $d \geq 3$, revealing a substantial performance gap between two-layer neural networks and random feature models. The exponential dependence on $1/d$ manifests the curse of dimensionality for nonparametric methods using fixed rather than adaptive basis functions.

6 Discussion

The ongoing debate over the double descent phenomenon and the virtues of overparametrization casts a cloud on the trustworthiness of modern deep learning methods and undermines the foundations of large machine learning models. We have developed a nonasymptotic generalization theory for finite-width two-layer neural networks without resorting to mean-field or neural tangent kernel approximations. As far as we are aware, this provides the first complete explanation for the double descent phenomenon beyond linear and kernel-type (e.g., random feature) methods. Compared with the existing literature, our theoretical framework has the following advantages: (i) we take a nonparametric viewpoint and consider target functions in a large function space, which allows us to define approximation and estimation errors in an appropriate manner and directly tackle the problem of bias–variance trade-off; (ii) unlike previous asymptotic studies, our nonasymptotic approach helps separate the effects of diverging dimensionality and overparametrization on generalization performance; (iii) the explicit regularization strategy we have adopted naturally extends classical and kernel ridge regression, making our results independent of the algorithmic specifics of nonconvex optimization.

Our theory yields insights that have not been previously obtained under simpler models or asymptotic regimes. We highlight some important ones as follows:

- *Impact of dimensionality.* In linear regression, the number of parameters coincides with the dimensionality, and hence it is impossible to decouple their effects. For kernel methods, Liang et al. (2020) and Montanari and Zhong (2022) relaxed the proportional asymptotics on n and d, but still required d to be polynomially growing with n. Our results show that for two-layer neural networks the double descent curve persists even when d is fixed and, therefore, the phenomenon is not tied to high dimensionality. Nevertheless, the dimensionality does play a role in determining the superiority of the overparametrized regime. Specifically, as seen from (18), a moderately large d suffices to ensure the global optimality of infinite overparametrization over a wide range of signal-to-noise ratio.

- *Double descent with optimal regularization.* In linear and random feature models, it has been shown that optimal ridge regularization eliminates double descent (Hastie et al., 2022; Nakkiran et al., 2021b; Mei and Montanari, 2022). This raises the concern of whether double descent should be treated as a pathological behavior due to insufficient regularization and hence should be avoided or mitigated in practice. Contrary to this view, our theory, which has been derived under optimal choices of the regularization parameter, provides a radically different framework in which double descent is an intrinsic feature rather than an artifact and cannot be eliminated by optimal regularization.

- *Complexity control.* As pointed out by Belkin et al. (2020a), the most interesting aspects of double descent is not the peaking phenomenon itself but its connection to classical ideas of the bias–variance trade-off and complexity control. Unfortunately, previous work offers little insight in this regard and does not clarify the mechanism behind the superiority of overparametrization. By contrast, our theory gives a clear explanation of what drives double descent: the partition of the parameter space into finitely many regions and the emergence of collinearity within each region reduce the effective dimensionality, thereby achieving adaptive complexity control in the overparametrized regime.

- *Bias–variance trade-off.* The literature presents a mixed picture for bias and variance in linear and random feature models (Hastie et al., 2022; Mei and Montanari, 2022; Yang et al., 2020). The somewhat peculiar behaviors demonstrated in these studies are partly due to the assumption that the data-generating model is parametric and varies with the dimensionality. Neural networks, however, are intrinsically nonparametric, and the bias–variance trade-off should be discussed within this framework (Geman et al., 1992). Embracing this viewpoint, our results suggest a double descent bias curve and an initially increasing and then constant variance. In particular, the decreasing risk in the overparametrized regime is caused by the decreasing bias and the constant variance. Although there is no apparent trade-off between bias and variance, the general principle of bias–variance trade-off (Derumigny and Schmidt-Hieber, 2023) still seems to hold: the bias and variance cannot simultaneously tend to zero. More subtly, there is in fact a bias–variance trade-off in estimating the best-approximating finite-width neural network, with λ effectively controlling the trade-off. With the optimal choice of λ, the classical U-shaped risk curve is absent because the bias and variance are not strictly monotonic functions of m.

Our framework may be extended in several directions. The most important would be to develop a general theory for deep neural networks, by finding a convex reformulation and analyzing the symmetric structures arising from overparametrization. Such a formulation does not seem to be readily available, but see Ergen and Pilanci (2021) for useful results in some special cases. To this end, an extension of the target function class $\mathcal{G}$ is required, which may be carried out in a similar spirit to the compositional function space in Schmidt-Hieber (2020) and the neural tree space in E and Wojtowytsch (2020). For simplicity, we have considered only explicit regularization and the theoretical optimal solution to the regularized problem. An interesting direction is to take into account practical algorithms and implicit regularization, possibly by exploring the connection of our problem to ℓ_2 regularization. Moreover, optimization error may be addressed by providing uniform convergence guarantees for all approximate solutions. Finally, it would be worthwhile to extend our theory to classification problems and more network architectures such as convolutional and recurrent neural networks. For example, if one considers multi-class classification with cross-entropy loss as in Bos and Schmidt-Hieber (2022), the group lasso formulation is still valid. One can then, in principle, follow the ideas of regularized M-estimation for high-dimensional regression (Negahban et al., 2012) to derive the theory. Additional efforts are required to address challenges similar to those discussed in Section 4.1.

Acknowledgments

This research was supported by National Natural Science Foundation of China grants 12171012, 12292980, and 12292981.

Appendices

These appendices contain proofs of the main results and technical details. We provide the proofs of Propositions 3–6 in Appendix A. The proofs of generalization bounds under the overparametrized regime are presented in Appendix B. Appendix C includes the proofs of results in the underparametrized regime. Appendix D contains mathematical details of the target function class and the proof of Theorem 2. Proofs of lower bounds are provided in Appendix E. Appendix F collects some technical lemmas that are needed to prove the main results.

Appendix A. Proofs for Section 3

In this section we provide the proofs of Propositions 3–6. To simplify the notation, we write $\widetilde{\mathbf{x}}_i = (\mathbf{x}_i^T, 1)^T$.

Proof of Proposition 3 Let $\widehat{\boldsymbol{\theta}}_{\ell_2}$ be an arbitrary solution to problem (10) and $\widehat{\boldsymbol{\theta}}$ an arbitrary solution to problem (9). By the optimality of $\widehat{\boldsymbol{\theta}}_{\ell_2}$ and $\widehat{\boldsymbol{\theta}}$, we have

$$J_n^{\ell_2}(\widehat{\boldsymbol{\theta}}_{\ell_2}; \lambda) \le J_n^{\ell_2}(\mathcal{T}_1(\widehat{\boldsymbol{\theta}}); \lambda), \qquad J_n(\widehat{\boldsymbol{\theta}}; \lambda) \le J_n(\widehat{\boldsymbol{\theta}}_{\ell_2}; \lambda), \tag{19}$$

where $J_n^{\ell_2}(\boldsymbol{\theta}; \lambda)$ is the objective function of (10). By the definition of $\mathcal{T}_1$, $J_n^{\ell_2}(\mathcal{T}_1(\boldsymbol{\theta}); \lambda) = J_n(\boldsymbol{\theta}; \lambda)$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}_m$. Moreover, $\widehat{\boldsymbol{\theta}}_{\ell_2} = \mathcal{T}_1(\widehat{\boldsymbol{\theta}}_{\ell_2})$. Combining these facts with (19) gives

$$J_n^{\ell_2}(\widehat{\boldsymbol{\theta}}_{\ell_2}; \lambda) \le J_n^{\ell_2}(\mathcal{T}_1(\widehat{\boldsymbol{\theta}}); \lambda) = J_n(\widehat{\boldsymbol{\theta}}; \lambda) \le J_n(\widehat{\boldsymbol{\theta}}_{\ell_2}; \lambda) = J_n^{\ell_2}(\widehat{\boldsymbol{\theta}}_{\ell_2}; \lambda),$$

which means that $\widehat{\boldsymbol{\theta}}_{\ell_2}$ is a minimizer of $J_n(\boldsymbol{\theta}; \lambda)$ and that $\mathcal{T}_1(\widehat{\boldsymbol{\theta}})$ a minimizer of $J_n^{\ell_2}(\boldsymbol{\theta}; \lambda)$, completing the proof. ■

Proof of Proposition 4 By direct differentiation, the gradient flow $d\boldsymbol{\theta}(t)/dt = -\nabla_{\boldsymbol{\theta}} J_n(\boldsymbol{\theta}(t); \lambda)$ for problem (9) can be written as

$$\frac{d}{dt} a_j(t) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - g(\mathbf{x}_i; \boldsymbol{\theta}(t))\big)^2 \sigma(\widetilde{\mathbf{x}}_i^T \mathbf{w}_j(t)) - \lambda \|\mathbf{w}_j(t)\|_2 \partial |a_j(t)|, \tag{20}$$

$$\frac{d}{dt} \mathbf{w}_j(t) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - g(\mathbf{x}_i; \boldsymbol{\theta}(t))\big)^2 a_j(t) \partial\sigma(\widetilde{\mathbf{x}}_i^T \mathbf{w}_j(t)) \widetilde{\mathbf{x}}_i - \lambda |a_j(t)| \partial \|w_j(t)\|_2, \tag{21}$$

for $j = 1, \ldots, m$, where ∂ denotes the subgradient. Using the identities $a\partial|a| = |a|$, $\mathbf{w}^T \partial \|\mathbf{w}\|_2 = \|\mathbf{w}\|_2^2$, and $z\partial\sigma(z) = \sigma(z)$, left multiplying (20) by $a_j(t)$ and (21) by $\mathbf{w}_j(t)^T$ gives

$$\frac{d}{dt} |a_j(t)|^2 = \frac{d}{dt} \|\mathbf{w}_j(t)\|_2^2, \quad j = 1, \ldots, m.$$

If the initialization is reparametrized by $\mathcal{T}_1$, that is, $|a_j(0)|^2 = \|\mathbf{w}_j(0)\|_2^2$ for all j, then we have, for all $t \geq 0$,

$$|a_j(t)|^2 = \|\mathbf{w}_j(t)\|_2^2, \quad j = 1, \dots, m. \tag{22}$$

Similarly, the gradient flow for problem (10) can be written as

$$\frac{d}{dt} a_j^{\ell_2}(t) = \frac{1}{n}\sum_{i=1}^{n} \big(y_i - g(\mathbf{x}_i; \boldsymbol{\theta}_{\ell_2}(t))\big)^2 \sigma(\widetilde{\mathbf{x}}_i^T \mathbf{w}_j^{\ell_2}(t)) - \lambda a_j^{\ell_2}(t), \tag{23}$$

$$\frac{d}{dt} \mathbf{w}_j^{\ell_2}(t) = \frac{1}{n}\sum_{i=1}^{n} \big(y_i - g(\mathbf{x}_i; \boldsymbol{\theta}_{\ell_2}(t))\big)^2 a_j^{\ell_2}(t) \partial\sigma(\widetilde{\mathbf{x}}_i^T \mathbf{w}_j^{\ell_2}(t))\widetilde{\mathbf{x}}_i - \lambda \mathbf{w}_j^{\ell_2}(t), \tag{24}$$

for $j = 1, \dots, m$. Using (22) we have $\|\mathbf{w}_j(t)\|_2 \partial |a_j(t)| = |a_j(t)| \partial |a_j(t)| = a_j(t)$ and $|a_j(t)| \partial \|\mathbf{w}_j(t)\|_2 = \|\mathbf{w}_j(t)\|_2 \partial \|\mathbf{w}_j(t)\|_2 = \mathbf{w}_j(t)$, in which case the gradient flows (20)–(21) and (23)–(24) are identical. Hence, their trajectories must coincide if initialized at the same point. ■

To prove Propositions 5 and 6, we first introduce the following lemma, which says that the ReLU function is linear over each cone Q_j.

Lemma 13 *If $(a_k, \mathbf{w}_k^T)^T$ and $(a_\ell, \mathbf{w}_\ell^T)^T$ lie in the interior of the same cone Q_j, then*

$$a_k \sigma(\mathbf{w}_k^T \widetilde{\mathbf{x}}_i) + a_\ell \sigma(\mathbf{w}_\ell^T \widetilde{\mathbf{x}}_i) = s_j \sigma(|a_k| \mathbf{w}_k^T \widetilde{\mathbf{x}}_i + |a_\ell| \mathbf{w}_\ell^T \widetilde{\mathbf{x}}_i)$$

for $i = 1, \dots, n$.

Proof By definition, all points $(a, \mathbf{w}^T)^T$ in the interior of Q_j satisfy $\operatorname{sgn}(a) = s_j$ and $I(\mathbf{w}^T \widetilde{\mathbf{x}}_i \geq 0) = (\mathbf{D}_j)_{ii}$, where $(\mathbf{D}_j)_{ii}$ is the ith diagonal entry of $\mathbf{D}_j$. Then

$$\begin{aligned} & a_k \sigma(\mathbf{w}_k^T \widetilde{\mathbf{x}}_i) + a_\ell \sigma(\mathbf{w}_\ell^T \widetilde{\mathbf{x}}_i) \\ & \quad = a_k \mathbf{w}_k^T \widetilde{\mathbf{x}}_i I(\mathbf{w}_k^T \widetilde{\mathbf{x}}_i \geq 0) + a_\ell \mathbf{w}_\ell^T \widetilde{\mathbf{x}}_i I(\mathbf{w}_\ell^T \widetilde{\mathbf{x}}_i \geq 0) \\ & \quad = s_j (\mathbf{D}_j)_{ii} (|a_k| \mathbf{w}_k^T \widetilde{\mathbf{x}}_i + |a_\ell| \mathbf{w}_\ell^T \widetilde{\mathbf{x}}_i) = s_j \sigma(|a_k| \mathbf{w}_k^T \widetilde{\mathbf{x}}_i + |a_\ell| \mathbf{w}_\ell^T \widetilde{\mathbf{x}}_i), \end{aligned}$$

which completes the proof. ■

Proof of Proposition 5 Suppose that $(\widehat{a}_k, \widehat{\mathbf{w}}_k^T)^T$ and $(\widehat{a}_\ell, \widehat{\mathbf{w}}_\ell^T)^T$ lie in the interior of Q_j but are not collinear. Define the new parameter $\widetilde{\boldsymbol{\theta}} = (\widetilde{a}_1, \dots, \widetilde{a}_m, \widetilde{\mathbf{w}}_1^T, \dots, \widetilde{\mathbf{w}}_m^T)^T$ with

$$\widetilde{a}_k = \widetilde{a}_\ell = s_j, \qquad \widetilde{\mathbf{w}}_k = \widetilde{\mathbf{w}}_\ell = \frac{1}{2}(|a_k| \mathbf{w}_k + |a_\ell| \mathbf{w}_\ell),$$

while keeping the other components unchanged. Then by Lemma 13 we have

$$\widehat{a}_k \sigma(\widehat{\mathbf{w}}_k^T \widetilde{\mathbf{x}}_i) + \widehat{a}_\ell \sigma(\widehat{\mathbf{w}}_\ell^T \widetilde{\mathbf{x}}_i) = \widetilde{a}_k \sigma(\widetilde{\mathbf{x}}_i^T \widetilde{\mathbf{w}}_\ell) + \widetilde{a}_\ell \sigma(\widetilde{\mathbf{x}}_i^T \widetilde{\mathbf{w}}_\ell),$$

and by the triangle inequality,

$$|\widetilde{a}_k| \|\widetilde{\mathbf{w}}_k\|_2 + |\widetilde{a}_\ell| \|\widetilde{\mathbf{w}}_\ell\|_2 = \big\| |\widehat{a}_k| \widehat{\mathbf{w}}_k + |\widehat{a}_\ell| \widehat{\mathbf{w}}_\ell \big\|_2 < |\widehat{a}_k| \|\widehat{\mathbf{w}}_k\|_2 + |\widehat{a}_\ell| \|\widehat{\mathbf{w}}_\ell\|_2.$$

This entails that $J_n(\widetilde{\boldsymbol{\theta}}; \lambda) < J_n(\widehat{\boldsymbol{\theta}}; \lambda)$, which contradicts the optimality of $\widehat{\boldsymbol{\theta}}$. ■

Proof of Proposition 6 By Lemma 13 and the definition of $\widetilde{\boldsymbol{\theta}} = \mathcal{T}_2(\boldsymbol{\theta})$ in (12), we have

$$\begin{aligned} g(\mathbf{x}_i;\boldsymbol{\theta}) &= \sum_{j=1}^{2p}\sum_{k\in S_j} a_k\sigma(\mathbf{w}_k^T\widetilde{\mathbf{x}}_i) = \sum_{j=1}^{2p} s_j\sigma\Big(\sum_{k\in S_j}|a_k|\mathbf{w}_k^T\widetilde{\mathbf{x}}_i\Big) \\ &= \sum_{j=1}^{2p}\sum_{k\in S_j}\widetilde{a}_k\sigma(\widetilde{\mathbf{w}}_k^T\widetilde{\mathbf{x}}_i) = g(\mathbf{x}_i;\widetilde{\boldsymbol{\theta}}) \end{aligned}$$

and

$$\begin{aligned} \|\mathbf{B}(\widetilde{\boldsymbol{\theta}})\|_{2,1} &= \sum_{j=1}^{2p}\|\boldsymbol{\beta}_j(\widetilde{\boldsymbol{\theta}})\|_2 = \sum_{j=1}^{2p}\Big\|\sum_{k\in S_j}|\widetilde{a}_k|\widetilde{\mathbf{w}}_k\Big\|_2 = \sum_{k=1}^{m}|\widetilde{a}_k|\|\widetilde{\mathbf{w}}_k\|_2 = \nu(\widetilde{\boldsymbol{\theta}}) \\ &= \sum_{j=1}^{2p}\Big\|\sum_{k\in S_j}|a_k|\mathbf{w}_k\Big\|_2 \le \sum_{k=1}^{m}|a_k|\|\mathbf{w}_k\|_2 = \nu(\boldsymbol{\theta}). \end{aligned}$$

To prove the second assertion, from (13) we have, for any $\boldsymbol{\theta}\in\boldsymbol{\Theta}_m$,

$$\frac{1}{2n}\Big\|\mathbf{y} - \sum_{k=1}^{m}a_k\sigma(\mathbf{X}\mathbf{w}_k)\Big\|_2^2 = \frac{1}{2n}\Big\|\mathbf{y} - \sum_{j=1}^{2p}\mathbf{D}_j\mathbf{X}\boldsymbol{\beta}_j(\boldsymbol{\theta})\Big\|_2^2. \tag{25}$$

Also, by the collinearity property of $\widehat{\boldsymbol{\theta}}$ from Proposition 5,

$$\|\mathbf{B}(\widehat{\boldsymbol{\theta}})\|_{2,1} = \sum_{j=1}^{2p}\Big\|\sum_{k\in S_j}|\widehat{a}_k|\widehat{\mathbf{w}}_k\Big\|_2 = \sum_{k=1}^{m}|\widehat{a}_k|\|\widehat{\mathbf{w}}_k\|_2 = \nu(\widehat{\boldsymbol{\theta}}). \tag{26}$$

Combining (25) and (26) yields the expression for $J_n(\widehat{\boldsymbol{\theta}};\lambda)$. ■

Appendix B. Proofs of Results in the Overparametrized Regime

In this section we provide the proofs of Theorems 7 and 8. We first introduce some notation. Define the class of two-layer ReLU networks with bounded scaled variation norm by $\mathcal{F}(m,F) = \{g(\cdot;\boldsymbol{\theta}) : \nu(\boldsymbol{\theta}) \le F\}$. For any $f^*\in\mathcal{G}_M$, let $g(\cdot;\boldsymbol{\theta}^*)$ denote the $L_\infty(\mathbb{B}^d)$-approximation of f^* in Theorem 2, where $\boldsymbol{\theta}_m^* = (a_1^*,\ldots,a_m^*,\mathbf{w}_1^{*T},\ldots,\mathbf{w}_m^{*T})^T$.

Proof of Theorem 7 By the optimality of $\widehat{\boldsymbol{\theta}}$, we have

$$\frac{1}{2n}\sum_{i=1}^{n}\big(g(\mathbf{x}_i;\widehat{\boldsymbol{\theta}}) - y_i\big)^2 + \lambda\nu(\widehat{\boldsymbol{\theta}}) \le \frac{1}{2n}\sum_{i=1}^{n}\big(g(\mathbf{x}_i;\boldsymbol{\theta}^*) - y_i\big)^2 + \lambda\nu(\boldsymbol{\theta}^*). \tag{27}$$

Rearranging terms gives

$$\begin{aligned} &\frac{1}{2}\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n^2 \\ &\quad\le \lambda\big(\nu(\boldsymbol{\theta}^*) - \nu(\widehat{\boldsymbol{\theta}})\big) + \frac{1}{2}\|g(\cdot;\boldsymbol{\theta}^*) - f^*\|_n^2 + \frac{1}{n}\Big|\sum_{i=1}^{n}\varepsilon_i\big(g(\mathbf{x}_i;\widehat{\boldsymbol{\theta}}) - g(\mathbf{x}_i;\boldsymbol{\theta}^*)\big)\Big| \\ &\quad\equiv T_1 + T_2 + T_3. \end{aligned} \tag{28}$$

For brevity, we write $\mathbf{B}^* = \mathbf{B}(\boldsymbol{\theta}^*) = (\boldsymbol{\beta}_1^*, \dots, \boldsymbol{\beta}_{2p}^*)$ and $\widehat{\mathbf{B}} = \mathbf{B}(\widehat{\boldsymbol{\theta}}) = (\widehat{\boldsymbol{\beta}}_1, \dots, \widehat{\boldsymbol{\beta}}_{2p})$. Since we need only evaluate $g(\cdot; \boldsymbol{\theta}^*)$ on the training sample, by Proposition 6 we can assume without loss of generality that $\nu(\boldsymbol{\theta}^*) = \|\mathbf{B}^*\|_{2,1}$. It also follows from Proposition 6 that $\nu(\widehat{\boldsymbol{\theta}}) = \|\widehat{\mathbf{B}}\|_{2,1}$. These facts, together with the triangle inequality, imply that

$$T_1 = \lambda(\|\mathbf{B}^*\|_{2,1} - \|\widehat{\mathbf{B}}\|_{2,1}) \le 2\lambda\|\mathbf{B}^*\|_{2,1} - \lambda\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{2,1}. \tag{29}$$

To bound T_2, applying Theorem 2 yields

$$T_2 = \frac{1}{2n}\sum_{i=1}^{n}\big(g(\mathbf{x}_i; \boldsymbol{\theta}^*) - f^*(\mathbf{x}_i)\big)^2 \le C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} \tag{30}$$

for some constant $C_1 > 0$. Also, by Hölder's inequality,

$$T_3 = \frac{1}{n}\left|\boldsymbol{\varepsilon}^T \sum_{j=1}^{2p} \mathbf{D}_j \mathbf{X}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*)\right| \le \frac{1}{\sqrt{n}} \max_{1\le j\le 2p} \|\mathbf{r}_j\|_2 \sum_{j=1}^{2p} \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_2, \tag{31}$$

where $\mathbf{r}_j = \mathbf{X}^T\mathbf{D}_j^T\boldsymbol{\varepsilon}/\sqrt{n}$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$. Combining (29)–(31), choosing $\lambda \ge 2n^{-1/2}\max_j \|\mathbf{r}_j\|_2$, and noting that $\nu(\boldsymbol{\theta}^*) \le 6\|f^*\|_{\mathcal{S}}$ by Theorem 2, we obtain

$$\begin{aligned} &\|g(\cdot; \widehat{\boldsymbol{\theta}}) - f^*\|_n^2 \\ &\quad \le 2C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + 4\lambda\nu(\boldsymbol{\theta}^*) + 2\left(\frac{1}{\sqrt{n}}\max_{1\le j\le 2p}\|\mathbf{r}_j\|_2 - \lambda\right)\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{2,1} \\ &\quad \le 2C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + 24\lambda\|f^*\|_{\mathcal{S}}. \end{aligned} \tag{32}$$

It remains to bound $n^{-1/2}\max_j \|\mathbf{r}_j\|_2$. Let $\mathbf{H}_j = \mathbf{D}_j\mathbf{X}\mathbf{X}^T\mathbf{D}_j/n$, so that $\mathbf{r}_j^T\mathbf{r}_j = \boldsymbol{\varepsilon}^T\mathbf{H}_j\boldsymbol{\varepsilon}$. By the definition of $\mathbf{D}_j$ and the assumption that $\max_i \|\mathbf{x}_i\|_2 \le 1$, we have

$$\|\mathbf{H}_j\|_2 \le \operatorname{tr}(\mathbf{H}_j) = n^{-1}\operatorname{tr}(\mathbf{X}^T\mathbf{X}) \le 2.$$

Applying a tail bound for quadratic forms of sub-Gaussian vectors (Hsu et al., 2012) gives

$$\mathbb{P}(\|\mathbf{r}_j\|_2^2 \ge 2\sigma_\varepsilon^2 + 4\sigma_\varepsilon^2\sqrt{t} + 4\sigma_\varepsilon^2 t) \le e^{-t}.$$

Hence, by the union bound, $\mathbb{P}(\max_j \|\mathbf{r}_j\|_2^2 \ge 2\sigma_\varepsilon^2 + 4\sigma_\varepsilon^2\sqrt{t} + 4\sigma_\varepsilon^2 t) \le 2pe^{-t}$. Recall from (11) that $p \le 2n^d$. Choosing $t = (4+d)\log n > 1$ for $n \ge 2$ yields

$$\max_{1\le j\le 2p} \|\mathbf{r}_j\|_2^2 < 2\sigma_\varepsilon^2 + 4\sigma_\varepsilon^2\sqrt{t} + 4\sigma_\varepsilon^2 t < 10\sigma_\varepsilon^2 t < 16\sigma_\varepsilon^2(4+d)\log n \tag{33}$$

with probability at least $1 - 4n^{-4}$. Thus, for $\lambda \ge 2n^{-1/2}\max_j \|\mathbf{r}_j\|_2$ to hold with the same probability, it suffices to set $\lambda = 8\sigma_\varepsilon\sqrt{(4+d)\log n/n}$. To complete the proof of (14), substituting the value of λ into (32) gives

$$\|g(\cdot; \widehat{\boldsymbol{\theta}}) - f^*\|_n^2 \le 2C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + 96(\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2)\sqrt{\frac{(4+d)\log n}{n}},$$

where we have used the inequality $2\sigma_\varepsilon\|f^*\|_{\mathcal{S}} \le \sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2$.

To prove the bound (15), define the event $E_1 = \{\lambda \geq 2n^{-1/2} \max_j \|\mathbf{r}_j\|_2\}$ with $\lambda = 8\sigma_\varepsilon \sqrt{(4+d)\log n/n}$. It follows from (32) that

$$\mathbb{E}\{\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n^2 I(E_1)\} \leq 24\lambda\|f^*\|_{\mathcal{S}} + 2C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d}.$$

It remains to bound $\mathbb{E}\{\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n^2 I(E_1^c)\}$. Recall the definition of T_3 in (28). By the Cauchy–Schwarz inequality, we have

$$\begin{aligned} T_3 &\leq \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2\right)^{1/2} \left\{\frac{1}{n}\sum_{i=1}^n \big(g(\mathbf{x}_i;\widehat{\boldsymbol{\theta}}) - g(\mathbf{x}_i;\boldsymbol{\theta}^*)\big)^2\right\}^{1/2} \\ &\leq \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2\right)^{1/2} (\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n + \|g(\cdot;\boldsymbol{\theta}^*) - f^*\|_n) \\ &\leq \frac{2}{n}\sum_{i=1}^n \varepsilon_i^2 + \frac{1}{4}\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n^2 + \frac{1}{4}\|g(\cdot;\boldsymbol{\theta}^*) - f^*\|_n^2. \end{aligned}$$

Rearranging terms, (28) becomes

$$\frac{1}{4}\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n^2 \leq \lambda\nu(\boldsymbol{\theta}^*) + \frac{3}{4}\|g(\cdot;\boldsymbol{\theta}^*) - f^*\|_n^2 + \frac{2}{n}\sum_{i=1}^n \varepsilon_i^2.$$

Taking expectation gives

$$\begin{aligned} &\mathbb{E}\{\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n^2 I(E_1^c)\} \\ &\quad\leq \{4\lambda\nu(\boldsymbol{\theta}^*) + 3\|g(\cdot;\boldsymbol{\theta}^*) - f^*\|_n^2\}\mathbb{P}(E_1^c) + \frac{8}{n}\mathbb{E}\left\{\sum_{i=1}^n \varepsilon_i^2 I(E_1^c)\right\} \\ &\quad\equiv R_1 + R_2. \end{aligned}$$

It follows from (33) that $\mathbb{P}(E_1^c) \leq 4n^{-4}$. Setting $\lambda = 8\sigma_\varepsilon\sqrt{(4+d)\log n/n}$ and applying Theorem 2 yields

$$R_1 \leq C_2\left(\|f^*\|_{\mathcal{S}}\sigma_\varepsilon\sqrt{\frac{d\log n}{n}} + \|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d}\right)n^{-4}$$

for some constant $C_2 > 0$. By the Cauchy–Schwarz inequality, we have $R_2 \leq 8(\mathbb{E}\varepsilon^4)^{1/2}n^{-2}$. We then conclude that $\mathbb{E}\{\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n^2 I(E_1^c)\}$ is of smaller order than $\mathbb{E}\{\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n^2 I(E_1)\}$, and consequently

$$\mathbb{E}\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n^2 \leq C_3\left\{\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2)\sqrt{\frac{d\log n}{n}}\right\}$$

for some constant $C_3 > 0$. ■

To prove Theorem 8, we need the following maximal inequality, whose proof can be found in Appendix F.

Lemma 14 *Assume that Condition (C2) holds. Let* $\mathcal{F}^*(m,1) = \{f - f^* \colon f \in \mathcal{F}(m,1), f^*$ *is fixed with* $\|f^*\|_{\mathcal{S}} \le 1\}$ *and* $Z_n = \sup_{f \in \mathcal{F}^*(m,1)} \big| \|f\|_n^2 - \|f\|_2^2 \big|$*. Then* $\mathbb{E} Z_n \le C_{\mathcal{F}} n^{-1/2}$ *for some constant* $C_{\mathcal{F}} > 0$ *depending only on* d*. Moreover, for any* $t \ge 0$*,*

$$\mathbb{P}\bigg(Z_n \ge \frac{C_{\mathcal{F}}}{\sqrt{n}} + t \bigg) \le \exp\bigg(-\frac{nt^2}{C_1 + C_2 t} \bigg) \tag{34}$$

for some constants $C_1, C_2 > 0$*.*

Proof of Theorem 8 Let $\widehat{\Delta}(\cdot) = g(\cdot; \widehat{\boldsymbol{\theta}}) - f^*(\cdot)$. By the proof of (14) in Theorem 7 and, in particular, (32), if we choose $\lambda \ge 8\sigma_\varepsilon \sqrt{(4+d)\log n/n}$, then, with probability at least $1 - 4n^{-4}$,

$$0 \le \|\widehat{\Delta}\|_n^2 \le 2C_1 \|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + 4\lambda \nu(\boldsymbol{\theta}^*) - \lambda(\|\widehat{\mathbf{B}}\|_{2,1} - \|\mathbf{B}^*\|_{2,1})$$

for some constant $C_1 > 0$. Since $\|\mathbf{B}^*\|_{2,1} = \nu(\boldsymbol{\theta}^*) \le 6\|f^*\|_{\mathcal{S}}$ and $\|\widehat{\mathbf{B}}\|_{2,1} = \nu(\widehat{\boldsymbol{\theta}})$, we further obtain

$$\nu(\widehat{\boldsymbol{\theta}}) \le 30\|f^*\|_{\mathcal{S}} + 2C_1 \lambda^{-1} \|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d}.$$

If we choose $\lambda \ge C_1 \|f^*\|_{\mathcal{S}} m^{-(d+3)/d}$, then

$$\nu(\widehat{\boldsymbol{\theta}}) \le 30\|f^*\|_{\mathcal{S}} + 2\|f^*\|_{\mathcal{S}} = 32\|f^*\|_{\mathcal{S}}. \tag{35}$$

By the positive homogeneity of ReLU, the scaled variation norm of $g(\cdot;\widehat{\boldsymbol{\theta}})/\nu(\widehat{\boldsymbol{\theta}})$ is exactly 1. Also, by definition, the $\mathcal{S}$-norm of $f^*/(32\|f^*\|_{\mathcal{S}})$ is smaller than 1. Thus, the event

$$\frac{\widehat{\Delta}}{32\|f^*\|_{\mathcal{S}}} = \frac{g(\cdot;\widehat{\boldsymbol{\theta}})}{32\|f^*\|_{\mathcal{S}}} - \frac{f^*}{32\|f^*\|_{\mathcal{S}}} \in \mathcal{F}^*(m,1) \tag{36}$$

occurs with probability at least $1 - 4n^{-4}$.

Now, conditioning on the event $\{\widehat{\Delta}/(32\|f^*\|_{\mathcal{S}}) \in \mathcal{F}^*(m,1)\}$, applying Lemma 14 with $t = \sqrt{d \log n/n}$ yields

$$\|\widehat{\Delta}\|_2^2 \le \|\widehat{\Delta}\|_n^2 + 1024\|f^*\|_{\mathcal{S}}^2 \frac{C_{\mathcal{F}}}{\sqrt{n}} + 1024\|f^*\|_{\mathcal{S}}^2 \sqrt{\frac{d\log n}{n}}$$

with probability at least $1 - O(n^{-C_3})$ for some constant $C_3 > 0$. Also, it follows from Theorem 7 that, with probability at least $1 - 4n^{-4}$,

$$\|\widehat{\Delta}\|_n^2 \le C_4 \bigg\{ \|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2) \sqrt{\frac{d \log n}{n}} \bigg\}$$

for some constant $C_4 > 0$. Combining these pieces, we conclude that

$$\|\widehat{\Delta}\|_2^2 \le C_5 \bigg\{ \|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2) \sqrt{\frac{d \log n}{n}} \bigg\}$$

with probability at least $1 - O(n^{-C_6})$ for some constants $C_5, C_6 > 0$, completing the proof of (16).

In what follows, we prove the bound (17). Note that $\mathbb{E}\|\widehat{\Delta}\|_2^2 = R_1 + R_2$, where

$$R_1 = \mathbb{E}\{\|\widehat{\Delta}\|_2^2 I(E_1)\}, \quad R_2 = \mathbb{E}\{\|\widehat{\Delta}\|_2^2 I(E_1^c)\}.$$

Under the event E_1, we have proved in (35) that $\nu(\widehat{\boldsymbol{\theta}}) \le 32\|f^*\|_{\mathcal{S}}$; that is, $I(E_1) \le I(\nu(\widehat{\boldsymbol{\theta}}) \le 32\|f^*\|_{\mathcal{S}})$. It follows from (36), Lemma 14, and Theorem 7 that

$$\begin{aligned} R_1 &\le \mathbb{E}\{\|\widehat{\Delta}\|_2^2 I(\nu(\widehat{\boldsymbol{\theta}}) \le 32\|f^*\|_{\mathcal{S}})\} \\ &\le \mathbb{E}\|\widehat{\Delta}\|_n^2 + 1024\|f^*\|_{\mathcal{S}}^2 \mathbb{E} \sup_{f \in \mathcal{F}^*(m,1)} \big| \|f\|_n^2 - \|f\|_2^2 \big| \\ &\le \mathbb{E}\|\widehat{\Delta}\|_n^2 + \frac{1024}{\sqrt{n}} C_{\mathcal{F}} \|f^*\|_{\mathcal{S}}^2 \\ &\le C_7 \left\{ \|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2) \sqrt{\frac{d \log n}{n}} \right\} \end{aligned} \tag{37}$$

for some constant $C_7 > 0$.

It remains to bound R_2. Using (27), (30), and our choice that $\lambda \ge C_1 \|f^*\|_{\mathcal{S}} m^{-(d+3)/d}$, we obtain

$$\begin{aligned} \nu(\widehat{\boldsymbol{\theta}}) &\le \nu(\boldsymbol{\theta}^*) + (2n\lambda)^{-1} \sum_{i=1}^n \big(g(\mathbf{x}_i; \boldsymbol{\theta}^*) - y_i\big)^2 \\ &\le \nu(\boldsymbol{\theta}^*) + \lambda^{-1} \|g(\cdot; \boldsymbol{\theta}^*) - f^*\|_n^2 + (n\lambda)^{-1} \sum_{i=1}^n \varepsilon_i^2 \\ &\le 6\|f^*\|_{\mathcal{S}} + C_1 \lambda^{-1} \|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (n\lambda)^{-1} \sum_{i=1}^n \varepsilon_i^2 \\ &\le 7\|f^*\|_{\mathcal{S}} + (n\lambda)^{-1} \sum_{i=1}^n \varepsilon_i^2 \end{aligned} \tag{38}$$

Note that $|g(\mathbf{x}; \widehat{\boldsymbol{\theta}})| \le \nu(\widehat{\boldsymbol{\theta}})$ and $|f^*(\mathbf{x})| \le \|f^*\|_{\mathcal{S}}$ for any $\mathbf{x} \in \mathbb{B}^d$. Thus,

$$R_2 \le \mathbb{E}\left\{ \sup_{\mathbf{x} \in \mathbb{B}^d} (2|g(\mathbf{x}; \widehat{\boldsymbol{\theta}})|^2 + 2|f^*(\mathbf{x})|^2) I(E_1^c) \right\} \le \mathbb{E}\{(2\nu^2(\widehat{\boldsymbol{\theta}}) + 2\|f^*\|_{\mathcal{S}}^2) I(E_1^c)\}.$$

By the Cauchy–Schwarz inequality and the fact that $\mathbb{P}(E_1^c) \le 4n^{-4}$, we further obtain

$$\begin{aligned} \mathbb{E}\{(2\nu^2(\widehat{\boldsymbol{\theta}}) + 2\|f^*\|_{\mathcal{S}}^2) I(E_1^c)\} &\le 2\sqrt{\mathbb{E}\{(\nu^2(\widehat{\boldsymbol{\theta}}) + \|f^*\|_{\mathcal{S}}^2)^2\}} \sqrt{\mathbb{P}(E_1^c)} \\ &\le 4n^{-2} \sqrt{2\mathbb{E}\nu^4(\widehat{\boldsymbol{\theta}}) + 2\|f^*\|_{\mathcal{S}}^4}. \end{aligned}$$

Also, it follows from (38) that

$$\nu^4(\widehat{\boldsymbol{\theta}}) \le \left\{ 98\|f^*\|_{\mathcal{S}}^2 + 2(n\lambda)^{-2} \left(\sum_{i=1}^n \varepsilon_i^2 \right)^2 \right\}^2 \le 19208\|f^*\|_{\mathcal{S}}^4 + 8(n\lambda)^{-4} \left(\sum_{i=1}^n \varepsilon_i^2 \right)^4.$$

Since ε_i are independent Gaussian variables, we have $\mathbb{E}(n^{-1}\sum_{i=1}^n \varepsilon_i^2)^4 < \infty$. Combining these pieces yields

$$R_2 \le C_8 n^{-2}(\|f^*\|_{\mathcal{S}}^2 + \lambda^{-2}) \tag{39}$$

for some constant $C_8 > 0$. Finally, note that $n^{-2}\lambda^{-2} = o(n^{-1/2}\lambda)$ under our choice of λ. Combining (37) and (39), we conclude that

$$\mathbb{E}\|\widehat{\Delta}\|_2^2 \le C_9\left\{\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2)\sqrt{\frac{d\log n}{n}}\right\}$$

for some constant $C_9 > 0$. ■

Appendix C. Proofs of Results in the Underparametrized Regime

In this section we present the proof of Theorem 9. Let $\mathcal{N}_m(\delta) \equiv \mathcal{N}(\delta, \mathcal{F}(m,1), \|\cdot\|_n)$ be the δ-covering number of $\mathcal{F}(m,1)$ with respect to $\|\cdot\|_n$, and define $\mathcal{F}^*(m,1) = \{f - f^* \colon \|f^*\|_{\mathcal{S}} \le 1, f \in \mathcal{F}(m,1)\}$.

We first bound the empirical error of the regularized network estimator in the underparametrized regime. To deal with scaled variation regularization, we need to analyze the supremum of an empirical process,

$$V_{m,\delta}(\boldsymbol{\varepsilon}) \equiv \sup_{f\in\mathcal{F}(m,1)}\left\{\frac{n^{-1}\left|\sum_{i=1}^n \varepsilon_i f(\mathbf{x}_i)\right| - \delta\sqrt{n^{-1}\sum_{i=1}^n \varepsilon_i^2}}{\|f\|_n + \delta}\right\},$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$.

Lemma 15 *The function $V_{m,\delta}(\cdot)\colon \mathbb{R}^n \to \mathbb{R}$ is $n^{-1/2}$-Lipschitz, and*

$$\mathbb{E}V_{m,\delta}(\boldsymbol{\varepsilon}) \le \sigma_\varepsilon\sqrt{\frac{2\log(2\mathcal{N}_m(\delta))}{n}}.$$

Proof We first show that $V_{m,\delta}(\boldsymbol{\varepsilon})$ is Lipschitz continuous with respect to the Euclidean norm. For any two vectors $\boldsymbol{\varepsilon}^{(1)} = (\varepsilon_1^{(1)}, \dots, \varepsilon_n^{(1)})^T$ and $\boldsymbol{\varepsilon}^{(2)} = (\varepsilon_1^{(2)}, \dots, \varepsilon_n^{(2)})^T$, the inequality $|\sup_{a\in\mathcal{A}} F(a) - \sup_{a\in\mathcal{A}} G(a)| \le \sup_{a\in\mathcal{A}}|F(a) - G(a)|$ implies that

$$\begin{aligned}
&|V_{m,\delta}(\boldsymbol{\varepsilon}^{(1)}) - V_{m,\delta}(\boldsymbol{\varepsilon}^{(2)})| \\
&\quad\le \sup_{f\in\mathcal{F}(m,1)}\left\{\frac{|n^{-1}\sum_{i=1}^n(\varepsilon_i^{(1)} - \varepsilon_i^{(2)})f(\mathbf{x}_i)| + \delta n^{-1/2}\|\boldsymbol{\varepsilon}^{(1)} - \boldsymbol{\varepsilon}^{(2)}\|_2}{\|f\|_n + \delta}\right\} \\
&\quad\le \sup_{f\in\mathcal{F}(m,1)}\left\{\frac{n^{-1/2}\|\boldsymbol{\varepsilon}^{(1)} - \boldsymbol{\varepsilon}^{(2)}\|_2\|f\|_n + \delta n^{-1/2}\|\boldsymbol{\varepsilon}^{(1)} - \boldsymbol{\varepsilon}^{(2)}\|_2}{\|f\|_n + \delta}\right\} = \frac{1}{\sqrt{n}}\|\boldsymbol{\varepsilon}^{(1)} - \boldsymbol{\varepsilon}^{(2)}\|_2,
\end{aligned}$$

which gives the desired result.

Next, we bound $\mathbb{E}V_{m,\delta}(\boldsymbol{\varepsilon})$. Let $\{f_j\}_{j=1}^{\mathcal{N}_m(\delta)}$ be a minimal δ-covering of $\mathcal{F}(m,1)$ with respect to $\|\cdot\|_n$. By definition, for any $f \in \mathcal{F}(m,1)$, there exists some j^* such that $\|f_{j^*} - f\|_n \le \delta$.

By the triangle inequality, we obtain

$$\begin{aligned}\left|\sum_{i=1}^n \varepsilon_i f(\mathbf{x}_i)\right| &\le \left|\sum_{i=1}^n \varepsilon_i f_{j^*}(\mathbf{x}_i)\right| + \left|\sum_{i=1}^n \varepsilon_i \big(f(\mathbf{x}_i) - f_{j^*}(\mathbf{x}_i)\big)\right| \\ &\le \left|\sum_{i=1}^n \varepsilon_i \frac{f_{j^*}(\mathbf{x}_i)}{\|f_{j^*}\|_n}\right| \|f_{j^*}\|_n + \delta\bigg(n\sum_{i=1}^n \varepsilon_i^2\bigg)^{1/2} \\ &\le \max_{1\le j\le \mathcal{N}_m(\delta)} \left|\sum_{i=1}^n \varepsilon_i \frac{f_j(\mathbf{x}_i)}{\|f_j\|_n}\right| (\|f\|_n + \delta) + \delta\bigg(n\sum_{i=1}^n \varepsilon_i^2\bigg)^{1/2}.\end{aligned}$$

After some algebra, we have

$$(\|f\|_n + \delta)^{-1}\left\{\left|\sum_{i=1}^n \varepsilon_i f(\mathbf{x}_i)\right| - \delta\bigg(n\sum_{i=1}^n \varepsilon_i^2\bigg)^{1/2}\right\} \le \max_{1\le j\le \mathcal{N}_m(\delta)} \left|\sum_{i=1}^n \varepsilon_i \frac{f_j(\mathbf{x}_i)}{\|f_j\|_n}\right|$$

for all $f \in \mathcal{F}(m, 1)$, or equivalently

$$V_{m,\delta}(\boldsymbol{\varepsilon}) \le \frac{1}{\sqrt{n}} \max_{1\le j\le \mathcal{N}_m(\delta)} \frac{1}{\sqrt{n}} \left|\sum_{i=1}^n \varepsilon_i \frac{f_j(\mathbf{x}_i)}{\|f_j\|_n}\right|.$$

Since ε_i are independent $N(0, \sigma_\varepsilon^2)$, $n^{-1/2}\sum_{i=1}^n \varepsilon_i f_j(\mathbf{x}_i)/\|f_j\|_n$ are also $N(0, \sigma_\varepsilon^2)$. It follows from the maximal inequality for sub-Gaussian variables (Boucheron et al., 2013, Theorem 2.5) that

$$\begin{aligned}\mathbb{E}V_{m,\delta}(\boldsymbol{\varepsilon}) &\le \frac{1}{\sqrt{n}} \max_{1\le j\le \mathcal{N}_m(\delta)} \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i \frac{f_j(\mathbf{x}_i)}{\|f_j\|_n}, -\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i \frac{f_j(\mathbf{x}_i)}{\|f_j\|_n}\right\} \\ &\le \sigma_\varepsilon \sqrt{\frac{2\log(2\mathcal{N}_m(\delta))}{n}},\end{aligned}$$

which completes the proof. ■

Next, we apply Lemma 15 to bound the empirical error $\|g(\cdot; \widehat{\boldsymbol{\theta}}) - f^*\|_n^2$ in the underparametrized regime.

Theorem 16 *Under Conditions (C1)–(C3), the regularized network estimator* $g(\cdot; \widehat{\boldsymbol{\theta}})$ *with* $\lambda = C_1 \max(\|f^*\|_{\mathcal{S}} m^{-(d+3)/d}, \sigma_\varepsilon md\log n/n)$ *satisfies*

$$\|g(\cdot; \widehat{\boldsymbol{\theta}}) - f^*\|_n^2 \le C\bigg\{\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2)\frac{md\log n}{n}\bigg\} \tag{40}$$

and $\nu(\widehat{\boldsymbol{\theta}}) \le C_2(\|f^*\|_{\mathcal{S}} + \sigma_\varepsilon)$ *with probability at least* $1 - O(n^{-C_3})$ *for some constants* $C_1, C_2, C_3, C > 0$.

Proof Let $\delta_n = \min(md\log n/n, 1)$. As in the proof of Theorem 7, we will bound T_1, T_2, and T_3 in (28). Let $\widetilde{\Delta}(\cdot) = g(\cdot; \widehat{\boldsymbol{\theta}}) - g(\cdot; \boldsymbol{\theta}^*)$. Note that $\widetilde{\Delta}(\cdot)$ is a two-layer ReLU network of width at most $2m$. With a slight abuse of notation, we write $\nu(\widetilde{\Delta}) = \|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{2,1}$ for the scaled variation norm of $\widetilde{\Delta}$. From (29) and (30), we have

$$T_1 \leq 2\lambda\nu(\boldsymbol{\theta}^*) - \lambda\nu(\widetilde{\Delta}), \quad T_2 \leq C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} \tag{41}$$

for some constant $C_1 > 0$. Define $\widehat{\sigma}_\varepsilon = \sqrt{n^{-1}\sum_{i=1}^n \varepsilon_i^2}$. For T_3, since $\widetilde{\Delta}/\nu(\widetilde{\Delta}) \in \mathcal{F}(2m, 1)$, we have

$$\frac{T_3 - \delta_n\nu(\widetilde{\Delta})\widehat{\sigma}_\varepsilon}{\|\widetilde{\Delta}\|_n + \delta_n\nu(\widetilde{\Delta})} = \frac{n^{-1}\left|\sum_{i=1}^n \varepsilon_i\widetilde{\Delta}(\mathbf{x}_i)/\nu(\widetilde{\Delta})\right| - \delta_n\widehat{\sigma}_\varepsilon}{\|\widetilde{\Delta}/\nu(\widetilde{\Delta})\|_n + \delta_n} \leq V_{2m,\delta_n}(\boldsymbol{\varepsilon}).$$

Noting that $V_{2m,\delta_n}(\boldsymbol{\varepsilon})$ is $n^{-1/2}$-Lipschitz by Lemma 15 and applying Theorem 2.26 of Wainwright (2019) yields

$$\mathbb{P}(|V_{2m,\delta_n}(\boldsymbol{\varepsilon}) - \mathbb{E}V_{2m,\delta_n}(\boldsymbol{\varepsilon})| \geq t) \leq 2\exp\left(-\frac{nt^2}{2}\right).$$

Also by Lemma 15, $\mathbb{E}V_{2m,\delta_n}(\boldsymbol{\varepsilon}) \leq 2\sigma_\varepsilon\sqrt{\log\mathcal{N}_{2m}(\delta_n)/n}$. Choosing $t = 2\sigma_\varepsilon\sqrt{\log\widetilde{p}/n}$ for some $\widetilde{p} \geq \mathcal{N}_{2m}(\delta_n)$ to be specified later, we have, with probability at least $1 - 2\widetilde{p}^{-2\sigma_\varepsilon^2}$,

$$V_{2m,\delta_n}(\boldsymbol{\varepsilon}) \leq 4\sigma_\varepsilon\sqrt{\frac{\log\widetilde{p}}{n}}.$$

Similarly,

$$\mathbb{P}(\widehat{\sigma}_\varepsilon \geq \sigma_\varepsilon + t) \leq \exp\left(-\frac{nt^2}{2}\right)$$

since $n^{-1/2}\|\boldsymbol{\varepsilon}\|_2$ is $n^{-1/2}$-Lipschitz and $n^{-1/2}\mathbb{E}\|\boldsymbol{\varepsilon}\|_2 \leq \sqrt{n^{-1}\mathbb{E}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}} = \sigma_\varepsilon$. Choosing $t = \sigma_\varepsilon$, we have, with probability at least $1 - \exp(-n\sigma_\varepsilon^2/2)$,

$$\widehat{\sigma}_\varepsilon \leq 2\sigma_\varepsilon.$$

Combining these pieces gives

$$T_3 \leq 4\sigma_\varepsilon\sqrt{\frac{\log\widetilde{p}}{n}}\big(\|\widetilde{\Delta}\|_n + \delta_n\nu(\widetilde{\Delta})\big) + 2\sigma_\varepsilon\delta_n\nu(\widetilde{\Delta}) \tag{42}$$

with probability at least $1 - 2\widetilde{p}^{-2\sigma_\varepsilon^2} - \exp(-\sigma_\varepsilon^2 n/2)$. Further combining (28), (41), and (42) yields

$$\begin{aligned}\frac{1}{2}\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n^2 \leq C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + 4\sigma_\varepsilon\sqrt{\frac{\log\widetilde{p}}{n}}\|\widetilde{\Delta}\|_n \\ + \left\{\left(2\sqrt{\frac{\log\widetilde{p}}{n}} + 1\right)2\sigma_\varepsilon\delta_n - \lambda\right\}\nu(\widetilde{\Delta}) + 2\lambda\nu(\boldsymbol{\theta}^*).\end{aligned} \tag{43}$$

Choosing $\lambda \geq (2\sqrt{\log\widetilde{p}/n} + 1)4\sigma_\varepsilon\delta_n$, we have

$$\begin{aligned}\frac{1}{2}\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n^2 \leq C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + 2\lambda\nu(\boldsymbol{\theta}^*) \\ + 4\sigma_\varepsilon\sqrt{\frac{\log\widetilde{p}}{n}}(\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n + \|g(\cdot;\boldsymbol{\theta}^*) - f^*\|_n),\end{aligned} \tag{44}$$

where we have used the triangle inequality to bound $\|\widetilde{\Delta}\|_n$. Note that

$$4\sigma_\varepsilon\sqrt{\frac{\log\widetilde{p}}{n}}\|g(\cdot;\widehat{\boldsymbol{\theta}})-f^*\|_n \leq 16\sigma_\varepsilon^2\frac{\log\widetilde{p}}{n}+\frac{1}{4}\|g(\cdot;\widehat{\boldsymbol{\theta}})-f^*\|_n^2, \tag{45}$$

$$4\sigma_\varepsilon\sqrt{\frac{\log\widetilde{p}}{n}}\|g(\cdot;\boldsymbol{\theta}^*)-f^*\|_n \leq 16\sigma_\varepsilon^2\frac{\log\widetilde{p}}{n}+\frac{1}{4}\|g(\cdot;\boldsymbol{\theta}^*)-f^*\|_n^2. \tag{46}$$

Substituting into (44), using $\nu(\boldsymbol{\theta}^*)\leq 6\|f^*\|_{\mathcal{S}}$, and rearranging, we obtain

$$\|g(\cdot;\widehat{\boldsymbol{\theta}})-f^*\|_n^2 \leq 6C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d}+128\sigma_\varepsilon^2\frac{\log\widetilde{p}}{n}+48\lambda\|f^*\|_{\mathcal{S}} \tag{47}$$

with probability at least $1-2\widetilde{p}^{-2\sigma_\varepsilon^2}-\exp(-\sigma_\varepsilon^2 n/2)$.

It remains to bound $\mathcal{N}_{2m}(\delta_n)\equiv\mathcal{N}(\delta_n,\mathcal{F}(2m,1),\|\cdot\|_n)$. Since a δ_n-covering of $\mathcal{F}(2m,1)$ with respect to $\|\cdot\|_\infty$ is always a δ_n-covering with respect to $\|\cdot\|_n$, by Lemma 24 we have

$$\log\mathcal{N}_{2m}(\delta_n)\leq\log\mathcal{N}(\delta_n,\mathcal{F}(2m,1),\|\cdot\|_\infty)\leq 2m(d+1)\log(1+2\sqrt{2}\delta_n^{-1}).$$

Hence, by the definition of δ_n,

$$\log\mathcal{N}_{2m}(\delta_n)\leq C_4 md\log n \tag{48}$$

for some constant $C_4>0$. Now take $\widetilde{p}=n^{C_4md}$. Then, for $\lambda\geq(2\sqrt{\log\widetilde{p}/n}+1)4\sigma_\varepsilon\delta_n$ to hold, it suffices to choose $\lambda\geq C_5\sigma_\varepsilon md\log n/n$ for some constant $C_5>0$. Plugging the values of $\widetilde{p}$ and λ into (47), we see that (40) holds with probability at least $1-2\widetilde{p}^{-2\sigma_\varepsilon^2}-\exp(-\sigma_\varepsilon^2 n/2)=1-O(n^{-C_3})$ for some constant $C_3>0$.

Next, we prove the bound on $\nu(\widehat{\boldsymbol{\theta}})$. Note that $(2\sqrt{\log\widetilde{p}/n}+1)2\sigma_\varepsilon\delta_n\leq\lambda/2$ under our choice of λ. Then (43) becomes

$$\begin{aligned}&\lambda\nu(\widetilde{\Delta})+\|g(\cdot;\widehat{\boldsymbol{\theta}})-f^*\|_n^2\\ &\quad\leq 2C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d}+8\sigma_\varepsilon\sqrt{\frac{\log\widetilde{p}}{n}}\|\widetilde{\Delta}\|_n+4\lambda\nu(\boldsymbol{\theta}^*).\end{aligned}$$

Using (45), (46), and $\nu(\boldsymbol{\theta}^*)\leq 6\|f^*\|_{\mathcal{S}}$, we have

$$\nu(\widetilde{\Delta})\leq 3C_1\lambda^{-1}\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d}+64\lambda^{-1}\sigma_\varepsilon^2\frac{\log\widetilde{p}}{n}+24\|f^*\|_{\mathcal{S}}.$$

Plugging in the values of $\widetilde{p}$ and λ, we obtain

$$\nu(\widetilde{\Delta})\leq 27\|f^*\|_{\mathcal{S}}+\frac{64C_4}{C_1}\sigma_\varepsilon,$$

and hence

$$\nu(\widehat{\boldsymbol{\theta}})\leq\nu(\widetilde{\Delta})+\nu(\boldsymbol{\theta}^*)\leq C_2(\|f^*\|_{\mathcal{S}}+\sigma_\varepsilon)$$

where $C_2=\max(33,64C_4/C_1)$. ■

Before proving Theorem 9, we present a useful lemma, whose proof can be found in Appendix F.

Lemma 17 *For any* $0 < \gamma \leq 1$, *let* $\mathcal{B}_{\mathcal{F}}(\gamma) = \{f \in \mathcal{F}^*(m,1) \colon \|f\|_2 \leq \gamma\}$ *denote the* $L_2(\mu)$*-ball of radius* γ *in* $\mathcal{F}^*(m,1)$*. Let* $Z_n(\gamma) = \sup_{f \in \mathcal{B}_{\mathcal{F}}(\gamma)} \big| \|f\|_n^2 - \|f\|_2^2 \big|$*. Then*

$$\mathbb{E} Z_n(\gamma) \leq C\bigg(\gamma\sqrt{\frac{md\log n}{n}} + \frac{md\log n}{n}\bigg)$$

for some constant $C > 0$.

We are now in a position to prove Theorem 9.

Proof of Theorem 9 Let $\widehat{\Delta}(\cdot) = g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*(\cdot)$ and $\gamma_n = \sqrt{md\log n/n}$. Define the events $E_2(\gamma) = \{Z_n(\gamma) \leq c_2\gamma^2\}$, $E_3 = \{Z_n(\|f\|_2) \leq c_3\gamma_n\|f\|_2$ for all $f \in \mathcal{F}^*(m,1)$ with $\|f\|_2 \geq \gamma_n\}$, and

$$E_4 = \big\{\big|\|f\|_n - \|f\|_2\big| \leq c_4\gamma_n \text{ for all } f \in \mathcal{F}^*(m,1)\big\},$$

where $c_2, c_3, c_4 > 0$ are universal constants to be specified later.

We claim that $E_2(\gamma_n) \cap E_3 \subset E_4$. Indeed, if $f \in \mathcal{F}^*(m,1)$ with $\|f\|_2 \leq \gamma_n$, then conditioning on $E_2(\gamma_n)$ we have

$$\|f\|_n \leq \sqrt{Z_n(\gamma_n) + \|f\|_2^2} \leq \sqrt{c_2\gamma_n^2 + \gamma_n^2} = \sqrt{c_2+1}\gamma_n,$$

in which case E_4 occurs with $c_4 \geq \sqrt{c_2+1}$. On the other hand, if $f \in \mathcal{F}^*(m,1)$ with $\|f\|_2 > \gamma_n$, then conditioning on E_3 we have

$$\big|\|f\|_n - \|f\|_2\big| = \frac{\big|\|f\|_n^2 - \|f\|_2^2\big|}{\|f\|_n + \|f\|_2} \leq \frac{c_3\gamma_n\|f\|_2}{\|f\|_2} = c_3\gamma_n,$$

and thus E_4 occurs with $c_4 \geq c_3$. Combining these two cases, we see that $E_2(\gamma_n) \cap E_3 \subset E_4$ with $c_4 = \max(\sqrt{c_2+1}, c_3)$.

Note that $\widehat{\Delta}/\max(\nu(\widehat{\boldsymbol{\theta}}), \|f^*\|_{\mathcal{S}}) \in \mathcal{F}^*(m,1)$. Also, it follows from Theorem 16 that $\nu(\widehat{\boldsymbol{\theta}}) \leq C_2(\|f^*\|_{\mathcal{S}} + \sigma_\varepsilon)$ with probability at least $1 - O(n^{-C_3})$, where $C_2 \geq 33$. Then, conditioning on E_4 we have

$$\big|\|\widehat{\Delta}\|_n - \|\widehat{\Delta}\|_2\big| \leq \max(\nu(\widehat{\boldsymbol{\theta}}), \|f^*\|_{\mathcal{S}})c_4\gamma_n \leq C_2(\|f^*\|_{\mathcal{S}} + \sigma_\varepsilon)c_4\gamma_n.$$

This, together with (40) in Theorem 16, yields the desired error bound.

It remains to bound the probability of the event E_4, or those of $E_2(\gamma_n)$ and E_3. To bound the former, note from Lemma 17 that

$$\mathbb{E} Z_n(\gamma) \leq C\bigg(\gamma\sqrt{\frac{md\log n}{n}} + \frac{md\log n}{n}\bigg) \leq 2C\gamma_n\gamma \tag{49}$$

for any $\gamma_n \leq \gamma \leq 1$. Applying Bousquet's inequality for suprema of empirical processes (Boucheron et al., 2013, Theorem 12.5) yields, for any $t \geq 0$,

$$\mathbb{P}\{Z_n(\gamma) - \mathbb{E}Z_n(\gamma) \geq t\} \leq \exp\bigg(-\frac{nt^2}{2(K_n + Ut/3)}\bigg), \tag{50}$$

where

$$U = \sup_{f \in \mathcal{F}^*(m,1)} \sup_{\mathbf{x} \in \mathbb{B}^d} |f(\mathbf{x})|^2, \quad K_n = 2U\mathbb{E}Z_n(\gamma) + \eta^2,$$

and

$$\eta^2 = \sup_{f \in \mathcal{B}_F(\gamma)} \mathrm{Var}(|f(\mathbf{x})|^2).$$

Note that

$$U \leq 2 \sup_{f \in \mathcal{F}(m,1)} \sup_{\mathbf{x} \in \mathbb{B}^d} |f(\mathbf{x})|^2 + 2 \sup_{\mathbf{x} \in \mathbb{B}^d} |f^*(\mathbf{x})|^2 \leq 4$$

and

$$\eta^2 \leq \sup_{f \in \mathcal{B}_{\mathcal{F}}(\gamma)} \mathbb{E}|f(\mathbf{x})|^4 \leq U \sup_{f \in \mathcal{B}_{\mathcal{F}}(\gamma)} \mathbb{E}|f(\mathbf{x})|^2 \leq 4\gamma^2.$$

Taking $\gamma = \gamma_n$ and $t = \gamma_n^2$ in (49) and (50), we have, with probability at least $1 - O(n^{-C_4})$ for some constant $C_4 > 0$,

$$Z(\gamma_n) \leq \mathbb{E}Z_n(\gamma_n) + \gamma_n^2 \leq (2C+1)\gamma_n^2,$$

that is, $E_2(\gamma_n)$ occurs with $c_2 = 2C + 1$.

Similarly, taking $\gamma = s_n$ and $t = \gamma_n s_n$ with $\gamma_n \leq s_n \leq 1$ in (49) and (50) yields

$$\mathbb{P}\{Z(s_n) \leq c_2 \gamma_n s_n\} \geq 1 - O(n^{-C_4}). \tag{51}$$

Note that (51) applies only to fixed s_n. For our purpose, we need to extend it to random choices of f that include $\widehat{\Delta}$. We do so by using a peeling argument. Since $\|f\|_2 \leq \sqrt{U} \leq 2$ for any $f \in \mathcal{F}^*(m,1)$, it suffices to consider the case $\gamma_n \leq \|f\|_2 \leq 2$; otherwise the desired result follows immediately. We cover the set $\{f \in \mathcal{F}^*(m,1) : \gamma_n \leq \|f\|_2 \leq 2\}$ by finitely many spherical shells $\mathcal{F}_j = \{f \in \mathcal{F}^*(m,1) : 2^{j-1}\gamma_n \leq \|f\|_2 \leq 2^j\gamma_n\}$, $j = 1, \ldots, Q$, where $Q = \lceil \log_2(2/\gamma_n) \rceil = O(\log n)$. If $Z_n(\|f\|_2) > c_3\gamma_n\|f\|_2$ for some $f \in \mathcal{F}_j$, then

$$Z_n(2^j\gamma_n) \geq Z_n(\|f\|_2) > c_3\gamma_n\|f\|_2 \geq c_3 2^{j-1}\gamma_n^2 = \frac{c_3}{2} 2^j \gamma_n^2.$$

Let $c_3 = 2c_2$. By the union bound and (51) with $s_n = 2^j\gamma_n$, we have

$$\begin{aligned}
\mathbb{P}(E_3^c) &\leq \sum_{j=1}^{Q} \mathbb{P}\{Z_n(\|f\|_2) > c_3\gamma_n\|f\|_2 \text{ for some } f \in \mathcal{F}_j\} \\
&\leq \sum_{j=1}^{Q} \mathbb{P}\{Z_n(2^j\gamma_n) > c_2 2^j \gamma_n^2\} \\
&\leq QO(n^{-C_4}) = O(n^{-C_4}\log n).
\end{aligned}$$

Combining the bounds for E_2 and E_3, we conclude that E_4 occurs with probability at least $1 - O(n^{-C_5})$ for some constant $C_5 > 0$, thereby completing the proof. ■

Appendix D. Mathematical Details of the Target Function Space

In this section we provide the omitted proofs regarding our target function space and the proof of Theorem 2. We first review the concepts of signed measures and total variation norm.

D.1 Signed Measures

Let $(D, \mathcal{B}(D))$ be a measurable space, where $D \subset \mathbb{R}^d$ and $\mathcal{B}(D)$ is the Borel σ-algebra on D. A finite signed measure μ is a set function $\mu \colon \mathcal{B}(D) \to \mathbb{R}$ such that $\mu(\emptyset) = 0$ and μ is σ-additive. Denote by $\mathcal{M}(D)$ the set of finite signed measures on $(D, \mathcal{B}(D))$. The Jordan decomposition theorem states that any finite signed measure $\mu \in \mathcal{M}(D)$ has a decomposition

$$\mu = \mu_+ - \mu_-,$$

where μ_+ and μ_- are mutually singular positive measures. Then the total variation of μ is defined by $|\mu| = \mu_+ + \mu_-$, and the total variation norm defined by $|\mu|(D) = \int_D d|\mu|$.

D.2 Proofs for Section 2

In the subsection, we formally state and prove the properties of the target function space $\mathcal{G}$ claimed in Section 2 and give the proof of Theorem 2.

For a two-layer ReLU network with parameter $\boldsymbol{\theta}$, denote by $C(\boldsymbol{\theta})$ the squared ℓ_2-norm of the network weights excluding the bias term, that is,

$$C(\boldsymbol{\theta}) = \frac{1}{2}\sum_{k=1}^{m}(\|\mathbf{v}_k\|_2^2 + |a_k|^2).$$

The *universal approximation ability* of neural networks refers to the property that any continuous function $f \colon \mathbb{R}^d \to \mathbb{R}$ can be approximated arbitrarily well by a neural network $g(\cdot; \boldsymbol{\theta})$ on a compact set $K \subset \mathbb{R}^d$ such that

$$\sup_{\mathbf{x} \in K} |f(\mathbf{x}) - g(\mathbf{x}; \boldsymbol{\theta})| \leq \varepsilon$$

for any $\varepsilon > 0$. Given any continuous function f, we are interested in the *representational cost* $\overline{R}(f)$ for approximating f using finite-width ReLU networks, which is defined by

$$\overline{R}(f) = \liminf_{\varepsilon \to 0} \left\{ C(\boldsymbol{\theta}) : \sup_{\|\mathbf{x}\|_2 \leq 1/\varepsilon} |g(\mathbf{x}; \boldsymbol{\theta}) - f(\mathbf{x})| \leq \varepsilon \text{ and } g(\mathbf{0}; \boldsymbol{\theta}) = f(\mathbf{0}) \right\}.$$

Ongie et al. (2020) proved that the representational cost $\overline{R}(f)$ is finite if and only if f is an infinite-width two-layer ReLU network with skip connections, which we restate below.

Lemma 18 (Lemma 10 and Theorem 2 of Ongie et al. (2020)) *Let f be a Lipschitz function defined on $\mathbb{R}^d$. There exists a seminorm $\|\cdot\|_{\mathcal{R}}$ such that $\|f\|_{\mathcal{R}}$ is finite if and only if*

$$f(\mathbf{x}) = \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \big(\sigma(\mathbf{v}^T\mathbf{x} + b) - \sigma(b)\big)\, d\alpha(\mathbf{w}) + \boldsymbol{\beta}^T\mathbf{x} + c$$

for some unique even signed measure $\alpha \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})$ and unique $\boldsymbol{\beta} \in \mathbb{R}^d$, $c \in \mathbb{R}$. Moreover, $\overline{R}(f)$ is finite if and only if $\|f\|_{\mathcal{R}}$ is finite, in which case $\|f\|_{\mathcal{R}} = |\alpha|(\mathbb{S}^{d-1} \times \mathbb{R}) \leq \overline{R}(f)$.

Ongie et al. (2020) used $\|\cdot\|_{\mathcal{R}}$ to characterize the representational cost $\overline{R}(\cdot)$, which was in turn linked to infinite-width two-layer ReLU networks. However, $\|\cdot\|_{\mathcal{R}}$ is not a norm on the function class $\{f : \overline{R}(f) < \infty\}$. We now extend Lemma 18 to provide another equivalent characterization of the representational cost, which relates $\overline{R}(\cdot)$ to a norm and motivates our definition of the $\mathcal{S}$-norm. Recall that

$$\mathcal{M}_2(\mathbb{R}^{d+1}) = \left\{\alpha \in \mathcal{M}(\mathbb{R}^{d+1}) : \int_{\mathbb{R}^{d+1}} \|\mathbf{v}\|_2 \, d|\alpha|(\mathbf{w}) < \infty\right\}.$$

Proposition 19 *Let f be a Lipschitz function defined on $\mathbb{R}^d$. Then $\|f\|_{\mathcal{R}}$ is finite if and only if*

$$f(\mathbf{x}) = \int_{\mathbb{R}^{d+1}} \big(\sigma(\mathbf{v}^T\mathbf{x} + b) - \sigma(b)\big)\, d\alpha(\mathbf{w}) + c \tag{52}$$

for some unique signed measure $\alpha \in \mathcal{M}_2(\mathbb{R}^{d+1})$ and unique $c \in \mathbb{R}$. Moreover, $\overline{R}(f)$ is finite if and only if $\int_{\mathbb{R}^{d+1}} \|\mathbf{v}\|_2 \, d|\alpha|(\mathbf{w})$ is finite.

Proof First, suppose that $\|f\|_{\mathcal{R}}$ is finite. By Lemma 18, there exists a unique even signed measure $\alpha \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})$ and unique $\boldsymbol{\beta} \in \mathbb{R}^d$, $c \in \mathbb{R}$ such that

$$f(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \big(\sigma(\mathbf{v}^T\mathbf{x} + b) - \sigma(b)\big)\, d\alpha(\mathbf{w}) + \boldsymbol{\beta}^T\mathbf{x} + c.$$

Note that $\boldsymbol{\beta}^T\mathbf{x} = \sigma(\boldsymbol{\beta}^T\mathbf{x}) - \sigma(-\boldsymbol{\beta}^T\mathbf{x})$ and define

$$\alpha_{\boldsymbol{\beta}}(\mathbf{w}) = \|\boldsymbol{\beta}\|_2 I(\boldsymbol{\beta} \neq \mathbf{0})\{I(\mathbf{v} = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|_2, b = 0) - I(\mathbf{v} = -\boldsymbol{\beta}/\|\boldsymbol{\beta}\|_2, b = 0)\}.$$

Then we can write $\boldsymbol{\beta}^T\mathbf{x} = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \big(\sigma(\mathbf{v}^T\mathbf{x}+b) - \sigma(b)\big)\, d\alpha_{\boldsymbol{\beta}}(\mathbf{w})$ and absorb it into the integral representation:

$$f(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \big(\sigma(\mathbf{v}^T\mathbf{x} + b) - \sigma(b)\big)\, d\widetilde{\alpha}(\mathbf{w}) + c,$$

where $\widetilde{\alpha} = \alpha + \alpha_{\boldsymbol{\beta}}$. Moreover, $|\widetilde{\alpha}|(\mathbb{S}^{d-1} \times \mathbb{R}) \leq |\alpha|(\mathbb{S}^{d-1} \times \mathbb{R}) + 2\|\boldsymbol{\beta}\|_2 < \infty$ and hence $\widetilde{\alpha} \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})$. A trivial extension of $\widetilde{\alpha}$ then yields a signed measure in $\mathcal{M}_2(\mathbb{R}^{d+1})$. The uniqueness of $\widetilde{\alpha}$ is implied by the uniqueness of α and $\boldsymbol{\beta}$.

Conversely, suppose that there exists some $\alpha \in \mathcal{M}_2(\mathbb{R}^{d+1})$ and $c \in \mathbb{R}$ such that the integral representation (52) holds. Define the normalization map $T\colon (\mathbb{R}^d \setminus \{\mathbf{0}\}) \times \mathbb{R} \to \mathbb{S}^{d-1} \times \mathbb{R}$ by $T(\mathbf{w}) = (\mathbf{v}^T, b)^T/\|\mathbf{v}\|_2$. Denote by $T_*\alpha$ the pushforward measure of α under T. By the positive homogeneity of ReLU and a change of variables, we have

$$\begin{aligned}
&\int_{\mathbb{R}^{d+1}} \big(\sigma(\mathbf{v}^T\mathbf{x} + b) - \sigma(b)\big)\, d\alpha(\mathbf{w}) \\
&\quad= \int_{\mathbb{R}^{d+1}} \big(\sigma(\mathbf{v}^T\mathbf{x}/\|\mathbf{v}\|_2 + b/\|\mathbf{v}\|_2) - \sigma(b/\|\mathbf{v}\|_2)\big)\|\mathbf{v}\|_2 I(\mathbf{v} \neq \mathbf{0})\, d\alpha(\mathbf{w}) \\
&\quad= \int_{\mathbb{R}^{d+1}} (\widetilde{\sigma}_{\mathbf{x}} \circ T)(\mathbf{w})\|\mathbf{v}\|_2 I(\mathbf{v} \neq \mathbf{0})\, d\alpha(\mathbf{w}) \\
&\quad= \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \widetilde{\sigma}_{\mathbf{x}}(\mathbf{w})\, d(T_*\alpha)(\mathbf{w}),
\end{aligned} \tag{53}$$

where $\widetilde{\sigma}_{\mathbf{x}}(\mathbf{w}) = \sigma(\mathbf{v}^T\mathbf{x} + b) - \sigma(b)$. Now decompose $T_*\alpha$ into an even and an odd part:

$$T_*\alpha = (T_*\alpha)_{\text{even}} + (T_*\alpha)_{\text{odd}}, \tag{54}$$

where $(T_*\alpha)_{\text{even}} = (T_*\alpha(A) + T_*\alpha(-A))/2$ and $(T_*\alpha)_{\text{odd}} = (T_*\alpha(A) - T_*\alpha(-A))/2$ for all $A \in \mathcal{B}(\mathbb{S}^{d-1} \times \mathbb{R})$. Then

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \widetilde{\sigma}_{\mathbf{x}}(\mathbf{w})\, d(T_*\alpha)_{\text{odd}}(\mathbf{w}) \\
&\quad= \frac{1}{2}\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \big(\widetilde{\sigma}_{\mathbf{x}}(\mathbf{w}) - \widetilde{\sigma}_{\mathbf{x}}(-\mathbf{w})\big)\, d(T_*\alpha)_{\text{odd}}(\mathbf{w}) \\
&\quad= \frac{1}{2}\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \big(\sigma(\mathbf{v}^T\mathbf{x} + b) - \sigma(b) - \sigma(-\mathbf{v}^T\mathbf{x} - b) + \sigma(-b)\big)\, d(T_*\alpha)_{\text{odd}}(\mathbf{w}) \qquad (55)\\
&\quad= \frac{1}{2}\int_{\mathbb{S}^{d-1}\times\mathbb{R}} (\mathbf{v}^T\mathbf{x} + b - b)\, d(T_*\alpha)_{\text{odd}}(\mathbf{w}) \\
&\quad= \frac{1}{2}\mathbf{x}^T \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \mathbf{v}\, d(T_*\alpha)_{\text{odd}}(\mathbf{w}) = \mathbf{x}^T\boldsymbol{\beta},
\end{aligned}$$

where $\boldsymbol{\beta} = \frac{1}{2}\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \mathbf{v}\, d(T_*\alpha)_{\text{odd}}(\mathbf{w})$ is well defined since

$$\|\boldsymbol{\beta}\|_2 \le \frac{1}{2}\int_{\mathbb{S}^{d-1}\times\mathbb{R}} d|T_*\alpha|(\mathbf{w}) \le \frac{1}{2}\int_{\mathbb{R}^{d+1}} \|\mathbf{v}\|_2\, d|\alpha|(\mathbf{w}) < \infty.$$

Combining (53)–(55), we obtain the representation

$$f(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \big(\sigma(\mathbf{x}^T\mathbf{v} + b) - \sigma(b)\big)\, d(T_*\alpha)_{\text{even}}(\mathbf{w}) + \boldsymbol{\beta}^T\mathbf{x} + c,$$

where $(T_*\alpha)_{\text{even}} \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})$ since

$$\int_{\mathbb{S}^{d-1}\times\mathbb{R}} d|(T_*\alpha)_{\text{even}}|(\mathbf{w}) \le \int_{\mathbb{R}^{d+1}} \|\mathbf{v}\|_2\, d|\alpha|(\mathbf{w}) < \infty.$$

By Lemma 18, we see that $\|f\|_{\mathcal{R}}$ is finite. This completes the proof of the first claim.

The second claim follows directly from Lemma 18 and the first claim. ■

In the following proposition, we show that functions in $\mathcal{G}(\mathbb{B}^d)$ have a cleaner integral representation.

Proposition 20 *For any $f \in \mathcal{G}(\mathbb{B}^d)$, there exists a signed measure $\widetilde{\alpha} \in \mathcal{M}(\mathbb{S}^{d-1} \times [-1, 1])$ and $c \in \mathbb{R}$ such that*

$$f(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} \sigma(\mathbf{v}^T\mathbf{x} + b)\, d\widetilde{\alpha}(\mathbf{w}) + c, \quad \mathbf{x} \in \mathbb{B}^d$$

and $|\widetilde{\alpha}|(\mathbb{S}^{d-1} \times [-1, 1]) \le 2\|f\|_{\mathcal{S}}$.

Proof Note that for any $\mathbf{x} \in \mathbb{B}^d$,

$$\begin{aligned}\sigma(\mathbf{v}^T\mathbf{x}+b)-\sigma(b) &= \big(\sigma(\mathbf{v}^T\mathbf{x}+b)-\sigma(b)\big)I(|b|\leq\|\mathbf{v}\|_2)\\ &\quad+\big(\sigma(\mathbf{v}^T\mathbf{x}+b)-\sigma(b)\big)I(b>\|\mathbf{v}\|_2)\\ &\quad+\big(\sigma(\mathbf{v}^T\mathbf{x}+b)-\sigma(b)\big)I(b<-\|\mathbf{v}\|_2)\\ &= \big(\sigma(\mathbf{v}^T\mathbf{x}+b)-\sigma(b)\big)I(|b|\leq\|\mathbf{v}\|_2)+\mathbf{v}^T\mathbf{x}I(b>\|\mathbf{v}\|_2).\end{aligned}$$

Using (53), the integral representation of $f \in \mathcal{G}(\mathbb{B}^d)$ can be written

$$\begin{aligned}f(\mathbf{x}) &= \int_{\mathbb{R}^{d+1}}\big(\sigma(\mathbf{v}^T\mathbf{x}+b)-\sigma(b)\big)\,d\alpha(\mathbf{w})\\ &= \int_{\mathbb{S}^{d-1}\times\mathbb{R}}\big(\sigma(\mathbf{x}^T\mathbf{v}+b)-\sigma(b)\big)\,d(T_*\alpha)(\mathbf{w})\\ &= \int_{\mathbb{S}^{d-1}\times\mathbb{R}}\sigma(\mathbf{v}^T\mathbf{x}+b)I(|b|\leq\|\mathbf{v}\|_2)\,d(T_*\alpha)(\mathbf{w})\\ &\quad-\int_{\mathbb{S}^{d-1}\times\mathbb{R}}\sigma(b)I(|b|\leq\|\mathbf{v}\|_2)\,d(T_*\alpha)(\mathbf{w})\\ &\quad+\mathbf{x}^T\int_{\mathbb{S}^{d-1}\times\mathbb{R}}\mathbf{v}I(b>\|\mathbf{v}\|_2)\,d(T_*\alpha)(\mathbf{w})\\ &= \int_{\mathbb{S}^{d-1}\times[-1,1]}\sigma(\mathbf{v}^T\mathbf{x}+b)\,d\alpha_1(\mathbf{w})+\mathbf{x}^T\boldsymbol{\beta}+c,\end{aligned}$$

where $d\alpha_1(\mathbf{w}) = I(|b|\leq\|\mathbf{v}\|_2)\,d(T_*\alpha)(\mathbf{w})$, $\boldsymbol{\beta} = \int_{\mathbb{S}^{d-1}\times\mathbb{R}}\mathbf{v}I(b>\|\mathbf{v}\|_2)\,d(T_*\alpha)(\mathbf{w})$, and $c = -\int_{\mathbb{S}^{d-1}\times\mathbb{R}}\sigma(b)I(|b|\leq\|\mathbf{v}\|_2)\,d(T_*\alpha)(\mathbf{w})$. Define

$$\alpha_{\boldsymbol{\beta}}(\mathbf{w}) = \|\boldsymbol{\beta}\|_2 I(\boldsymbol{\beta}\neq\mathbf{0})\{I(\mathbf{v}=\boldsymbol{\beta}/\|\boldsymbol{\beta}\|_2, b=0)-I(\mathbf{v}=-\boldsymbol{\beta}/\|\boldsymbol{\beta}\|_2, b=0)\},$$

and we can further write

$$f(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]}\sigma(\mathbf{v}^T\mathbf{x}+b)\,d\widetilde{\alpha}(\mathbf{w})+c,$$

where $\widetilde{\alpha} = \alpha_1+\alpha_{\boldsymbol{\beta}}$. Finally, note that

$$\begin{aligned}|\widetilde{\alpha}|(\mathbb{S}^{d-1}\times[-1,1]) &\leq |\alpha_1|(\mathbb{S}^{d-1}\times[-1,1])+2\|\boldsymbol{\beta}\|_2\\ &\leq \int_{\mathbb{S}^{d-1}\times\mathbb{R}}I(|b|\leq\|\mathbf{v}\|_2)\,d|T_*\alpha|(\mathbf{w})+2\int_{\mathbb{S}^{d-1}\times\mathbb{R}}I(b>\|\mathbf{v}\|_2)\,d|T_*\alpha|(\mathbf{w})\\ &\leq 2\int_{\mathbb{S}^{d-1}\times\mathbb{R}}d|T_*\alpha|(\mathbf{w})\leq 2\|f\|_{\mathcal{S}},\end{aligned}$$

completing the proof. ■

We are now ready to give the proof of Theorem 2, which is based on Proposition 1 of Bach (2017). While the case $d < 4$ and weight control were not addressed by Bach (2017), we refer to Siegel (2025) for a more complete statement.

Proof of Theorem 2 Fix any $f \in \mathcal{G}(\mathbb{B}^d)$. By Proposition 20, there exists a signed measure $\alpha \in \mathcal{M}(\mathbb{S}^{d-1} \times [-1,1])$ and $c \in \mathbb{R}$ such that

$$f(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} \sigma(\mathbf{v}^T\mathbf{x} + b)\, d\alpha(\mathbf{w}) + c, \quad \mathbf{x} \in \mathbb{B}^d.$$

Moreover, $|\alpha|(\mathbb{S}^{d-1} \times [-1,1]) \leq 2\|f\|_{\mathcal{S}}$ and

$$|c| = \left|\int_{\mathbb{S}^{d-1}\times\mathbb{R}} \sigma(b) I(|b| \leq \|\mathbf{v}\|_2)\, d(T_*\alpha)(\mathbf{w})\right| \leq \int_{\mathbb{S}^{d-1}\times\mathbb{R}} d|T_*\alpha|(\mathbf{w}) \leq \|f\|_{\mathcal{S}}.$$

Define the normalization map $\widetilde{T}\colon \mathbb{S}^{d-1} \times [-1,1] \to \mathbb{S}^d$ by $\widetilde{T}(\mathbf{w}) = \mathbf{w}/\|\mathbf{w}\|_2$. Denote by $\widetilde{T}_*\alpha$ the pushforward measure of α under $\widetilde{T}$. By the positive homogeneity of ReLU and a change of variables, we can write

$$\begin{aligned}\int_{\mathbb{S}^{d-1}\times[-1,1]} \sigma(\mathbf{v}^T\mathbf{x} + b)\, d\alpha(\mathbf{w}) &= \int_{\mathbb{S}^{d-1}\times[-1,1]} (\sigma_{\mathbf{x}} \circ \widetilde{T})(\mathbf{w}) \|\mathbf{w}\|_2\, d\alpha(\mathbf{w}) \\ &= \int_{\mathbb{S}^d} \sigma_{\mathbf{x}}(\mathbf{w})\, d(\widetilde{T}_*\alpha)(\mathbf{w}),\end{aligned} \tag{56}$$

where $\sigma_{\mathbf{x}}(\mathbf{w}) = \sigma(\mathbf{v}^T\mathbf{x} + b)$. Let $\widetilde{\mathbf{x}} = (\mathbf{x}^T, 1)^T$ and $\mathbf{z} = \widetilde{\mathbf{x}}/\|\widetilde{\mathbf{x}}\|_2$. Then $\mathbf{z} \in \mathbb{S}^d$ and we can further write

$$\int_{\mathbb{S}^d} \sigma_{\mathbf{x}}(\mathbf{w})\, d(\widetilde{T}_*\alpha)(\mathbf{w}) = \|\widetilde{\mathbf{x}}\|_2 \int_{\mathbb{S}^d} \sigma(\mathbf{w}^T\mathbf{z})\, d(\widetilde{T}_*\alpha)(\mathbf{w}). \tag{57}$$

Define the function $\widetilde{f}(\mathbf{z}) = \int_{\mathbb{S}^d} \sigma(\mathbf{w}^T\mathbf{z})\, d(\widetilde{T}_*\alpha)(\mathbf{w})$. Note that the γ_1-norm of $\widetilde{f}$ as defined in Bach (2017) is bounded by

$$\int_{\mathbb{S}^d} d|\widetilde{T}_*\alpha|(\mathbf{w}) \leq \int_{\mathbb{S}^{d-1}\times[-1,1]} \|\mathbf{w}\|_2\, d|\alpha|(\mathbf{w}) \leq 2\int_{\mathbb{S}^{d-1}\times[-1,1]} d|\alpha|(\mathbf{w}) \leq 4\|f\|_{\mathcal{S}}.$$

Applying Proposition 1 of Bach (2017) to $\widetilde{f}$, we see that there exist $a_1, \ldots, a_m \in \mathbb{R}$ and $\mathbf{w}_1, \ldots, \mathbf{w}_m \in \mathbb{S}^d$ such that

$$\left|\int_{\mathbb{S}^d} \sigma(\mathbf{w}^T\mathbf{z})\, d(\widetilde{T}_*\alpha)(\mathbf{w}) - \sum_{k=1}^m a_k \sigma(\mathbf{w}_k^T\mathbf{z})\right| \leq C_1 \|f\|_{\mathcal{S}} m^{-(d+3)/(2d)} \tag{58}$$

for some constant $C_1 > 0$ depending only on d, and $\sum_{k=1}^m |a_k| \leq |\widetilde{T}_*\alpha|(\mathbb{S}^d) \leq 4\|f\|_{\mathcal{S}}$. Combining (56)–(58), we have for any $\mathbf{x} \in \mathbb{B}^d$,

$$\begin{aligned}&\left|\int_{\mathbb{S}^{d-1}\times[-1,1]} \sigma(\mathbf{v}^T\mathbf{x} + b)\, d\alpha(\mathbf{w}) - \sum_{k=1}^m a_k \sigma(\mathbf{w}_k^T\widetilde{\mathbf{x}})\right| \\ &\quad = \|\widetilde{\mathbf{x}}\|_2 \left|\int_{\mathbb{S}^d} \sigma(\mathbf{w}^T\mathbf{z})\, d(\widetilde{T}_*\alpha)(\mathbf{w}) - \sum_{k=1}^m a_k \sigma(\mathbf{w}_k^T\mathbf{z})\right| \leq 2C_1 \|f\|_{\mathcal{S}} m^{-(d+3)/(2d)}.\end{aligned}$$

By adding two extra terms $\sigma(c) - \sigma(-c)$ to represent the constant term, we obtain a network of width $m+2$ with the scaled variation norm $\nu(\boldsymbol{\theta}) = \sum_{k=1}^m |a_k| + 2|c| \leq 6\|f\|_{\mathcal{S}}$. Replacing $m+2$ by m and adjusting the constant accordingly, the desired result follows. ■

Lemmas 21 and 22 verify two less immediate examples of functions in $\mathcal{G}$ given in Section 2.2. These examples will also be used in the proofs of Theorem 11 and Proposition 12.

Lemma 21 *For some probability measure ρ on $\mathbb{S}^{d-1} \times [-1,1]$, the RKHS associated with the kernel function $H_\rho(\mathbf{x}, \mathbf{z}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} \sigma(\mathbf{v}^T\mathbf{x} + b)\sigma(\mathbf{v}^T\mathbf{z} + b)\, d\rho(\mathbf{w})$ is*

$$\mathcal{H}_\rho = \left\{ \mathbf{x} \mapsto \int_{\mathbb{S}^{d-1}\times[-1,1]} a(\mathbf{w})\sigma(\mathbf{v}^T\mathbf{x} + b)\, d\rho(\mathbf{w}) : \int_{\mathbb{S}^{d-1}\times[-1,1]} |a(\mathbf{w})|^2\, d\rho(\mathbf{w}) < \infty \right\}$$

equipped with the inner product

$$\langle f_1, f_2 \rangle_{\mathcal{H}_\rho} = \int_{\mathbb{S}^{d-1}\times[-1,1]} a_1(\mathbf{w})a_2(\mathbf{w})\, d\rho(\mathbf{w}), \tag{59}$$

where $f_j(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} a_j(\mathbf{w})\sigma(\mathbf{v}^T\mathbf{x} + b)\, d\rho(\mathbf{w})$, $j = 1, 2$.

Proof With the inner product defined in (59), it is easy to verify, for any $f_j \in \mathcal{H}_\rho$ with $f_j(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} a_j(\mathbf{w})\sigma(\mathbf{v}^T\mathbf{x} + b)\, d\rho(\mathbf{w})$, $j = 1, 2, 3$, that (i) $\langle f_1, f_2 \rangle_{\mathcal{H}_\rho} = \langle f_2, f_1 \rangle_{\mathcal{H}_\rho}$, (ii) $\langle f_1, f_1 \rangle_{\mathcal{H}_\rho} \geq 0$ with $\langle f_1, f_1 \rangle_{\mathcal{H}_\rho} = 0$ if and only if $a_1(\mathbf{w}) = 0$ ρ-almost surely or $f_1 = 0$, and (iii) $\langle f_1 + cf_2, f_3 \rangle_{\mathcal{H}_\rho} = \langle f_1, f_3 \rangle_{\mathcal{H}_\rho} + c\langle f_2, f_3 \rangle_{\mathcal{H}_\rho}$ for any $c \in \mathbb{R}$. Thus, $\langle \cdot, \cdot \rangle_{\mathcal{H}_\rho}$ is a valid inner product on $\mathcal{H}_\rho$. Also, it satisfies the reproducing property. Indeed, for any $f \in \mathcal{H}_\rho$ with $f(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} a(\mathbf{w})\sigma(\mathbf{v}^T\mathbf{x} + b)\, d\rho(\mathbf{w})$ and $\mathbf{x} \in \mathbb{R}^d$, we have

$$\langle f, H_\rho(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_\rho} = \int_{\mathbb{S}^{d-1}\times[-1,1]} a(\mathbf{w})\sigma(\mathbf{v}^T\mathbf{x} + b)\, d\rho(\mathbf{w}) = f(\mathbf{x}).$$

It remains to show that $\mathcal{H}_\rho$ is complete. Denote by $\|\cdot\|_{\mathcal{H}_\rho} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}_\rho}}$ the norm induced by the inner product. If $\{f_n\}_{n=1}^\infty$ is a Cauchy sequence in $\mathcal{H}_\rho$ with $f_n(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} a_n(\mathbf{w})\sigma(\mathbf{v}^T\mathbf{x} + b)\, d\rho(\mathbf{w})$, then

$$\|f_n - f_m\|_{\mathcal{H}_\rho}^2 = \int_{\mathbb{S}^{d-1}\times[-1,1]} |a_n(\mathbf{w}) - a_m(\mathbf{w})|^2\, d\rho(\mathbf{w}) \to \infty$$

as $n, m \to \infty$. This implies that $\{a_n\}_{n=1}^\infty$ is a Cauchy sequence in $L_2(\mathbb{S}^{d-1} \times [-1,1], \rho)$. Since $L_2(\mathbb{S}^{d-1} \times [-1,1], \rho)$ is complete, there exists some $a \in L_2(\mathbb{S}^{d-1} \times [-1,1], \rho)$ such that $\|a_n - a\|_{L_2(\mathbb{S}^{d-1}\times[-1,1],\rho)} \to 0$ as $n \to \infty$. Define $f(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} a(\mathbf{w})\sigma(\mathbf{v}^T\mathbf{x} + b)\, d\rho(\mathbf{w})$. Then $f \in \mathcal{H}_\rho$ and

$$\|f_n - f\|_{\mathcal{H}_\rho}^2 = \int_{\mathbb{S}^{d-1}\times[-1,1]} |a_n(\mathbf{w}) - a(\mathbf{w})|^2\, d\rho(\mathbf{w}) \to 0$$

as $n \to \infty$. This shows that $\mathcal{H}_\rho$ is complete and finishes the proof. ■

Lemma 22 *Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be twice differentiable with $\phi'' \in L_1(\mathbb{R})$, and fix some $\mathbf{v}_0 \in \mathbb{S}^{d-1}$. If $f(\mathbf{x}) = \phi(\mathbf{v}_0^T\mathbf{x})$, then $f \in \mathcal{G}$ and*

$$\|f\|_{\mathcal{S}} = 2|\phi'(0)| + \int_{-\infty}^{\infty} |\phi''(t)|\, dt < \infty.$$

Proof By Taylor's expansion, we have

$$\phi(y) = \phi(0) + \phi'(0)y + \int_0^y (y-t)\phi''(t)\,dt.$$

Using the identity $y - t = \sigma(y-t) - \sigma(t-y)$, we can split the integral into two parts:

$$\begin{aligned}\int_0^y (y-t)\phi''(t)\,dt &= \int_0^y \sigma(y-t)\phi''(t)\,dt - \int_0^y \sigma(t-y)\phi''(t)\,dt \\ &= \int_0^\infty \sigma(y-t)\phi''(t)\,dt + \int_{-\infty}^0 \sigma(t-y)\phi''(t)\,dt.\end{aligned}$$

Substituting $y = \mathbf{v}_0^T\mathbf{x}$ and letting

$$d\alpha_f(\mathbf{w}) = \{\delta_{\mathbf{v}_0}(d\mathbf{v}) - \delta_{-\mathbf{v}_0}(d\mathbf{v})\}\phi'(0)\delta_0(db) + \{\delta_{\mathbf{v}_0}(d\mathbf{v})\phi''(b) + \delta_{-\mathbf{v}_0}(d\mathbf{v})\phi''(-b)\}I(b>0)\,db,$$

we obtain the representation

$$f(\mathbf{x}) = \phi(\mathbf{v}_0^T\mathbf{x}) = \phi(0) + \int_{\mathbb{R}^{d+1}} \big(\sigma(\mathbf{v}^T\mathbf{x} - b) - \sigma(-b)\big)\,d\alpha_f(\mathbf{w}).$$

Thus, by definition,

$$\begin{aligned}\|f\|_{\mathcal{S}} &= \int_{\mathbb{R}^{d+1}} \|\mathbf{v}\|_2\,d|\alpha_f|(\mathbf{w}) \\ &= 2|\phi'(0)| + \int_0^\infty |\phi''(b)|\,db + \int_0^\infty |\phi''(-b)|\,db \\ &= 2|\phi'(0)| + \int_{-\infty}^\infty |\phi''(b)|\,db < \infty,\end{aligned}$$

which implies $f \in \mathcal{G}$. ∎

Appendix E. Proofs of Lower Bounds

In this section we prove Theorem 11 and Proposition 12.

Proof of Theorem 11 The idea of the proof is to reduce the problem to estimation over a high-dimensional ℓ_1-ball and then apply the Yang–Barron version of Fano's method (Yang and Barron, 1999; Wainwright, 2019).

We choose ρ to be the uniform distribution on $\mathbb{S}^{d-1}$ and consider the kernel function $H_\rho(\mathbf{x},\mathbf{z}) = \int_{\mathbb{S}^{d-1}} \sigma(\mathbf{v}^T\mathbf{x})\sigma(\mathbf{v}^T\mathbf{z})\,d\rho(\mathbf{v})$. By the computations in Bach (2017, Appendix D.2), there exists a sequence of eigenfunctions $\phi_1, \phi_2, \dots$ that are orthonormal in $L_2(\mathbb{S}^{d-1})$ with the corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ such that

$$\int_{\mathbb{S}^{d-1}} H_\rho(\mathbf{x},\mathbf{z})\phi_j(\mathbf{z})\,d\rho(\mathbf{z}) = \lambda_j\phi_j(\mathbf{x}) \tag{60}$$

for all j. Let $a_j(\mathbf{v}) = \lambda_j^{-1} \int_{\mathbb{S}^{d-1}} \sigma(\mathbf{v}^T\mathbf{z})\phi_j(\mathbf{z})\, d\rho(\mathbf{z})$. Note from (60) that

$$\begin{aligned}\phi_j(\mathbf{x}) &= \frac{1}{\lambda_j}\int_{\mathbb{S}^{d-1}} H_\rho(\mathbf{x},\mathbf{z})\phi_j(\mathbf{z})\, d\rho(\mathbf{z}) \\ &= \frac{1}{\lambda_j}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}} \sigma(\mathbf{v}^T\mathbf{x})\sigma(\mathbf{v}^T\mathbf{z})\phi_j(\mathbf{z})\, d\rho(\mathbf{v})\, d\rho(\mathbf{z}) \\ &= \int_{\mathbb{S}^{d-1}} a_j(\mathbf{v})\sigma(\mathbf{v}^T\mathbf{x})\, d\rho(\mathbf{v}).\end{aligned}$$

By the orthonormality of ϕ_j in $L_2(\mathbb{S}^{d-1})$, we have

$$\begin{aligned}\langle \phi_j, \phi_j\rangle_{\mathcal{H}_\rho} &= \int_{\mathbb{S}^{d-1}} |a_j(\mathbf{v})|^2\, d\rho(\mathbf{v}) \\ &= \frac{1}{\lambda_j^2}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}} \sigma(\mathbf{v}^T\mathbf{x})\phi_j(\mathbf{x})\, d\rho(\mathbf{x})\int_{\mathbb{S}^{d-1}} \sigma(\mathbf{v}^T\mathbf{z})\phi_j(\mathbf{z})\, d\rho(\mathbf{z})\, d\rho(\mathbf{v}) \\ &= \frac{1}{\lambda_j^2}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}} H_\rho(\mathbf{x},\mathbf{z})\phi_j(\mathbf{x})\phi_j(\mathbf{z})\, d\rho(\mathbf{x})\, d\rho(\mathbf{z}) \\ &= \frac{1}{\lambda_j}\int_{\mathbb{S}^{d-1}} |\phi_j(\mathbf{x})|^2\, d\rho(\mathbf{x}) = \frac{1}{\lambda_j}.\end{aligned}$$

Thus, by the Cauchy–Schwarz inequality,

$$\|\phi_j\|_{\mathcal{S}}^2 = \left(\int_{\mathbb{S}^{d-1}} |a_j(\mathbf{v})|\, d\rho(\mathbf{v})\right)^2 \le \int_{\mathbb{S}^{d-1}} |a_j(\mathbf{v})|^2\, d\rho(\mathbf{v}) = \frac{1}{\lambda_j}.$$

Now consider the class of functions

$$\mathcal{I}_{J_n} = \left\{\sum_{j=1}^{J_n} \beta_j\phi_j : \boldsymbol{\beta} = (\beta_1,\ldots,\beta_{J_n})^T \in \mathbb{B}_1^{J_n}\right\},$$

where $\mathbb{B}_1^{J_n}$ denotes the unit ℓ_1-ball in $\mathbb{R}^{J_n}$. By the subadditivity of the $\mathcal{S}$-norm, we have

$$\left\|\sum_{j=1}^{J_n}\beta_j\phi_j\right\|_{\mathcal{S}} \le \sum_{j=1}^{J_n}|\beta_j|\|\phi_j\|_{\mathcal{S}} \le \sum_{j=1}^{J_n}|\beta_j|\lambda_j^{-1/2} \le \lambda_{J_n}^{-1/2},$$

and hence $\mathcal{I}_{J_n} \subset \mathcal{G}(\mathbb{S}^{d-1})$. Thus, it suffices to prove the minimax lower bound for the class $\mathcal{I}_{J_n}$. For any $f^{(k)} = \sum_{j=1}^{J_n}\beta_j^{(k)}\phi_j \in \mathcal{I}_{J_n}$, $k = 1, 2$, let $P^{(k)}$ denote the data distribution under the true regression function $f^{(k)}$. By the orthonormality of ϕ_j in $L_2(\mathbb{S}^{d-1})$, the Kullback–Leibler (KL) divergence between $P^{(1)}$ and $P^{(2)}$ can be expressed as

$$\begin{aligned}D_{\mathrm{KL}}(P^{(1)} \parallel P^{(2)}) = \frac{n}{2\sigma_\varepsilon^2}\|f^{(1)} - f^{(2)}\|_2^2 &= \frac{n}{2\sigma_\varepsilon^2}\left\|\sum_{j=1}^{J_n}\beta_j^{(1)}\phi_j - \sum_{j=1}^{J_n}\beta_j^{(2)}\phi_j\right\|_2^2 \\ &= \frac{n}{2\sigma_\varepsilon^2}\sum_{j=1}^{J_n}(\beta_j^{(1)} - \beta_j^{(2)})^2 = \frac{n}{2\sigma_\varepsilon^2}\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\|_2^2.\end{aligned}$$

Let $\mathcal{N}(\xi, \mathcal{I}_{J_n}, D_{\mathrm{KL}}^{1/2})$ denote the ξ-covering number of $\mathcal{I}_{J_n}$ in the square-root KL divergence. By Example 5.32 of Wainwright (2019), we have

$$\log \mathcal{N}(\xi, \mathcal{I}_{J_n}, D_{\mathrm{KL}}^{1/2}) = \log \mathcal{N}\left(\sqrt{\frac{2}{n}}\sigma_\varepsilon \xi, \mathbb{B}_1^{J_n}, \|\cdot\|_2\right) \leq \frac{C_1 n \log J_n}{2\sigma_\varepsilon^2 \xi^2} \tag{61}$$

for some constant $C_1 > 0$. Also, it follows from Lemma 2 of Raskutti et al. (2011) that

$$\log \mathcal{N}(\delta, \mathbb{B}_1^{J_n}, \|\cdot\|_2) \geq C_2 \left(\frac{1}{\delta}\right)^2 \log J_n,$$

provided that

$$C_3 \sqrt{\frac{\log J_n}{J_n^{c_1}}} \leq \delta < 1, \tag{62}$$

for some constants $C_2, C_3 > 0$ and $c_1 \in (0,1)$. Combined with Lemma 5.5 of Wainwright (2019), the δ-packing number of $\mathbb{B}_1^{J_n}$ in the $\|\cdot\|_2$-norm satisfies

$$\log \mathcal{P}(\delta, \mathbb{B}_1^{J_n}, \|\cdot\|_2) \geq \log \mathcal{N}(\delta, \mathbb{B}_1^{J_n}, \|\cdot\|_2) \geq C_2 \left(\frac{1}{\delta}\right)^2 \log J_n. \tag{63}$$

To apply the Yang–Barron method (Wainwright, 2019, Proposition 15.12 and Lemma 15.21), we need to find $\delta, \xi > 0$ such that

$$\log \mathcal{P}(2\delta, \mathcal{I}_{J_n}, \|\cdot\|_2) \geq 2(\xi^2 + \log \mathcal{N}(\xi, \mathcal{I}_{J_n}, D_{\mathrm{KL}}^{1/2}) + \log 2).$$

In view of (61) and (63), it suffices to choose $\xi^2 \asymp \sqrt{n \log J_n}/\sigma_\varepsilon$ and $\delta^2 \asymp \sigma_\varepsilon \sqrt{\log J_n / n}$. Furthermore, the choice $J_n = n^{c_2}$ with $c_2 > 1/2$ satisfies the requirement (62). We then conclude that

$$\inf_{\widehat{f}} \sup_{f^* \in \mathcal{G}(\mathbb{S}^{d-1})} \mathbb{E}\|\widehat{f} - f^*\|_2^2 \geq \inf_{\widehat{f}} \sup_{f^* \in \mathcal{I}_{J_n}} \mathbb{E}\|\widehat{f} - f^*\|_2^2 \geq \frac{\delta^2}{2} \geq C\sigma_\varepsilon \sqrt{\frac{\log n}{n}}$$

for some constant $C > 0$. ■

We now turn to the proof of Proposition 12.

Proof of Proposition 12 We first observe that, for every choice of $\lambda > 0$, the random feature estimator $h(\cdot; \widehat{\mathbf{a}}(\lambda))$ is a linear combination of at most $r \equiv \min(m, n)$ fixed basis functions that do not depend on f^*; see, for example, Hastie et al. (2009, Chapter 5). For any function f and any set of functions $\mathcal{H}$ in $L_2(\mu)$, let $d(f, \mathcal{H}) = \inf_{h \in \mathcal{H}} \|f - h\|_2$ denote the distance between f and $\mathcal{H}$. Also, by Fatou's lemma,

$$\inf_{\lambda > 0} \mathbb{E}\|h_{\rho_0}(\cdot; \widehat{\mathbf{a}}(\lambda)) - f^*\|_2 \geq \mathbb{E} \inf_{\lambda > 0} \|h_{\rho_0}(\cdot; \widehat{\mathbf{a}}(\lambda)) - f^*\|_2.$$

Thus, it suffices to find a lower bound for the approximation error between $\mathcal{G}_M$ and the span of $h_1, \ldots, h_r$ that holds uniformly over all choices of fixed basis functions $h_1, \ldots, h_r$:

$$\inf_{h_1, \ldots, h_r} \sup_{f^* \in \mathcal{G}_M} d(f^*, \mathrm{span}(h_1, \ldots, h_r)). \tag{64}$$

Our next step is to construct an orthogonal set of functions in $\mathcal{G}_M$, which will allow us to apply a key lemma of Barron (1993, Lemma 6) for approximating a $2r$-dimensional space by an r-dimensional linear subspace. To this end, define

$$h^*_{\mathbf{k}}(\mathbf{x}) = \sqrt{2}\cos(\pi\sqrt{d}\mathbf{k}^T\mathbf{x})$$

for $\mathbf{k} \in \{0, 1, \dots\}^d \setminus \{\mathbf{0}\}$. It is easy to verify that $\langle h^*_{\mathbf{k}_1}, h^*_{\mathbf{k}_2}\rangle_{L_2(\mu)} = I(\mathbf{k}_1 \neq \mathbf{k}_2)$. Also, note that $h^*_{\mathbf{k}}$ are ridge functions of the form $h^*_{\mathbf{k}}(\mathbf{x}) = \phi_{\mathbf{k}}(\mathbf{k}_0^T\mathbf{x})$, where $\phi_{\mathbf{k}}(t) = \sqrt{2}\cos(\pi\sqrt{d}\|\mathbf{k}\|_2 t)$ and $\mathbf{k}_0 = \mathbf{k}/\|\mathbf{k}\|_2$. Using Lemma 22 and after some computation, we find that

$$\|h^*_{\mathbf{k}}\|_{\mathcal{S}} = 2|\phi'_{\mathbf{k}}(0)| + \int_{-1}^{1} |\phi''_{\mathbf{k}}(t)|\, dt \leq C_1 d\|\mathbf{k}\|_2^2$$

for some universal constant $C_1 > 0$. Hence, $Mh^*_{\mathbf{k}}/(C_1 d\|\mathbf{k}\|_2^2) \in \mathcal{G}_M$.

Now let the indices $\mathbf{k}_1, \mathbf{k}_2, \dots$ be ordered in terms of increasing ℓ_2-norm, and let $\mathcal{H}^*_{2r} = \operatorname{span}(h^*_{\mathbf{k}_1}, \dots, h^*_{\mathbf{k}_{2r}})$. By projecting $h_1, \dots, h_r$ onto $\mathcal{H}^*_{2r}$, we need only bound the infimum in (64) over all r-dimensional linear subspaces $\mathcal{H}_r$ of $\mathcal{H}^*_{2r}$:

$$\inf_{h_1,\dots,h_r} \sup_{f^* \in \mathcal{G}_M} d(f^*, \operatorname{span}(h_1, \dots, h_r)) \geq \inf_{\mathcal{H}_r} \sup_{f^* \in \mathcal{H}^*_{2r} \cap \mathcal{G}_M} d(f^*, \mathcal{H}_r).$$

By Lemma 6 of Barron (1993), there exists some $1 \leq j \leq 2r$ such that $d^2(h^*_{\mathbf{k}_j}, \mathcal{H}_r) \geq 1/2$. Combined with the fact that $\|\mathbf{k}_{2r}\|_2 \asymp r^{1/d}$, we further obtain

$$\inf_{\mathcal{H}_r} \sup_{f^* \in \mathcal{H}^*_{2r} \cap \mathcal{G}_M} d(f^*, \mathcal{H}_r) \geq \frac{M}{C_1 d\|\mathbf{k}_{2r}\|_2^2} \inf_{\mathcal{H}_r} \sup_{1 \leq j \leq 2r} d(h^*_{\mathbf{k}_j}, \mathcal{H}_r) \geq \frac{CM}{dr^{2/d}}$$

for some universal constant $C > 0$. This completes the proof. ■

Appendix F. Technical Lemmas

The following lemma is needed for bounding the supremum of the empirical process in Lemma 14. Let $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_n)$ be a vector of independent Rademacher variables, that is, $\mathbb{P}(\zeta_i = 1) = \mathbb{P}(\zeta_i = -1) = 1/2$.

Lemma 23 *Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be vectors in $\mathbb{B}^d$. Then, for any $m \geq 1$ and $F > 0$,*

$$\mathbb{E}_{\boldsymbol{\zeta}} \sup_{f \in \mathcal{F}(m,F)} \left| \sum_{i=1}^{n} \zeta_i f(\mathbf{x}_i) \right| \leq 2F\sqrt{2n}.$$

Proof Note that $\|\widetilde{\mathbf{x}}_i\|_2 = \|(\mathbf{x}_i^T, 1)^T\|_2 \leq \sqrt{2}$. By the definition of $\mathcal{F}(m, F)$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\zeta}} \sup_{f \in \mathcal{F}(m,F)} \left| \sum_{i=1}^n \zeta_i f(\mathbf{x}_i) \right| &= \mathbb{E}_{\boldsymbol{\zeta}} \sup_{\nu(\boldsymbol{\theta}) \leq F} \left| \sum_{i=1}^n \zeta_i \sum_{k=1}^m a_k \sigma(\mathbf{w}_k^T \widetilde{\mathbf{x}}_i) \right| \\
&\leq \mathbb{E}_{\boldsymbol{\zeta}} \sup_{\nu(\boldsymbol{\theta}) \leq F} \sup_{\|\mathbf{u}_k\|_2 = 1} \left| \sum_{i=1}^n \zeta_i \sum_{k=1}^m a_k \|\mathbf{w}_k\|_2 \sigma(\mathbf{u}_k^T \widetilde{\mathbf{x}}_i) \right| \\
&= \mathbb{E}_{\boldsymbol{\zeta}} \sup_{\nu(\boldsymbol{\theta}) \leq F} \sup_{\|\mathbf{u}_k\|_2 = 1} \left| \sum_{k=1}^m a_k \|\mathbf{w}_k\|_2 \sum_{i=1}^n \zeta_i \sigma(\mathbf{u}_k^T \widetilde{\mathbf{x}}_i) \right| \\
&\leq \sup_{\nu(\boldsymbol{\theta}) \leq F} \sum_{k=1}^m |a_k| \|\mathbf{w}_k\|_2 \mathbb{E}_{\boldsymbol{\zeta}} \sup_{\|\mathbf{u}\|_2 = 1} \left| \sum_{i=1}^n \zeta_i \sigma(\mathbf{u}^T \widetilde{\mathbf{x}}_i) \right| \\
&\leq F \mathbb{E}_{\boldsymbol{\zeta}} \sup_{\|\mathbf{u}\|_2 = 1} \left| \sum_{i=1}^n \zeta_i \sigma(\mathbf{u}^T \widetilde{\mathbf{x}}_i) \right|.
\end{aligned}$$

Since $\sigma(\cdot)$ is 1-Lipschitz, by the contraction principle (Boucheron et al., 2013, Theorem 11.6) we have

$$\mathbb{E}_{\boldsymbol{\zeta}} \sup_{\|\mathbf{u}\|_2 = 1} \left| \sum_{i=1}^n \zeta_i \sigma(\mathbf{u}^T \widetilde{\mathbf{x}}_i) \right| \leq 2 \mathbb{E}_{\boldsymbol{\zeta}} \sup_{\|\mathbf{u}\|_2 = 1} \left| \sum_{i=1}^n \zeta_i \mathbf{u}^T \widetilde{\mathbf{x}}_i \right|.$$

By the Cauchy–Schwarz inequality and Jensen's inequality, we further obtain

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\zeta}} \sup_{\|\mathbf{u}\|_2 = 1} \left| \sum_{i=1}^n \zeta_i \mathbf{u}^T \widetilde{\mathbf{x}}_i \right| \leq \mathbb{E}_{\boldsymbol{\zeta}} \left\| \sum_{i=1}^n \zeta_i \widetilde{\mathbf{x}}_i \right\|_2 &\leq \left(\mathbb{E}_{\boldsymbol{\zeta}} \left\| \sum_{i=1}^n \zeta_i \widetilde{\mathbf{x}}_i \right\|_2^2 \right)^{1/2} \\
&= \left(\sum_{i=1}^n \|\widetilde{\mathbf{x}}_i\|_2^2 \right)^{1/2} \leq \sqrt{2n}.
\end{aligned}$$

Combining these pieces leads to the desired result. ■

We now prove Lemma 14.

Proof of Lemma 14 By the symmetrization inequality (Boucheron et al., 2013, Lemma 11.4),

$$\mathbb{E} Z_n \leq 2 \mathbb{E}_{\boldsymbol{\zeta}, \mathbf{x}} \sup_{f \in \mathcal{F}^*(m,1)} \left| \frac{1}{n} \sum_{i=1}^n \zeta_i f^2(\mathbf{x}_i) \right|. \tag{65}$$

Note that $\sup_{f \in \mathcal{F}^*(m,1)} \sup_{\mathbf{x} \in \mathbb{B}^d} |f(\mathbf{x})| \leq 2$ and $\phi(x) = x^2$ is 4-Lipschitz on $[-2, 2]$. By the contraction principle (Boucheron et al., 2013, Theorem 11.6), it follows that

$$\mathbb{E} Z_n \leq 2 \mathbb{E}_{\boldsymbol{\zeta}, \mathbf{x}} \sup_{f \in \mathcal{F}^*(m,1)} \left| \frac{1}{n} \sum_{i=1}^n \zeta_i \phi(f(\mathbf{x}_i)) \right| \leq 16 \mathbb{E}_{\boldsymbol{\zeta}, \mathbf{x}} \sup_{f \in \mathcal{F}^*(m,1)} \left| \frac{1}{n} \sum_{i=1}^n \zeta_i f(\mathbf{x}_i) \right|.$$

For some fixed $\widetilde{m} \geq n^{(d+3)/d}$, let $\widetilde{f} \in \mathcal{F}(\widetilde{m}, 1)$ be the two-layer ReLU network of width $\widetilde{m}$ that best approximates f^* under the $L_\infty(\mathbb{B}^d)$-norm in Theorem 2. Then $\|\widetilde{f} - f^*\|_\infty \leq$

$C\|f^*\|_{\mathcal{S}}\widetilde{m}^{-(d+3)/(2d)} \le Cn^{-1/2}$ for some constant $C > 0$. By decomposing $f - f^* = f - \widetilde{f} + \widetilde{f} - f^*$, noting that $f - \widetilde{f} \in \mathcal{F}(m + \widetilde{m}, 2)$, and using Lemma 23, we have

$$\begin{aligned}
\mathbb{E}Z_n &\le 16\mathbb{E}_{\boldsymbol{\zeta},\mathbf{x}} \sup_{f\in\mathcal{F}(m,1)}\left|\frac{1}{n}\sum_{i=1}^{n}\zeta_i\big(f(\mathbf{x}_i) - \widetilde{f}(\mathbf{x}_i)\big)\right| + 16\mathbb{E}_{\boldsymbol{\zeta},\mathbf{x}}\left|\frac{1}{n}\sum_{i=1}^{n}\zeta_i\big(\widetilde{f}(\mathbf{x}_i) - f^*(\mathbf{x}_i)\big)\right| \\
&\le 16\mathbb{E}_{\boldsymbol{\zeta},\mathbf{x}} \sup_{f\in\mathcal{F}(m+\widetilde{m},2)}\left|\frac{1}{n}\sum_{i=1}^{n}\zeta_i f(\mathbf{x}_i)\right| + 16\|\widetilde{f} - f^*\|_\infty \qquad (66)\\
&\le \frac{64\sqrt{2} + 16C}{\sqrt{n}} \equiv \frac{C_{\mathcal{F}}}{\sqrt{n}}.
\end{aligned}$$

To prove the bound (34), we apply Bousquet's inequality for suprema of empirical processes (Boucheron et al., 2013, Theorem 12.5) to obtain

$$\mathbb{P}(Z_n - \mathbb{E}Z_n \ge t) \le \exp\left(-\frac{nt^2}{2(K_n + Ut/3)}\right), \qquad (67)$$

where

$$U = \sup_{f\in\mathcal{F}^*(m,1)} \sup_{\mathbf{x}\in\mathbb{B}^d} |f(\mathbf{x})|^2, \quad K_n = 2U\mathbb{E}Z_n + \eta^2,$$

and

$$\eta^2 = \sup_{f\in\mathcal{F}^*(m,1)} \mathrm{Var}(|f(\mathbf{x})|^2).$$

Note that

$$U \le 2\sup_{f\in\mathcal{F}(m,1)}\sup_{\mathbf{x}\in\mathbb{B}^d}|f(\mathbf{x})|^2 + 2\sup_{\mathbf{x}\in\mathbb{B}^d}|f^*(\mathbf{x})|^2 \le 4$$

and

$$\eta^2 \le \sup_{f\in\mathcal{F}(m,1)} \mathbb{E}|f(\mathbf{x})|^4 \le U^2 \le 16.$$

Combining these bounds with (66) and (67) leads to

$$\mathbb{P}\left(Z_n \ge \frac{C_{\mathcal{F}}}{\sqrt{n}} + t\right) \le \mathbb{P}(Z_n - \mathbb{E}Z_n \ge t) \le \exp\left(-\frac{nt^2}{C_1 + C_2 t}\right),$$

where $C_1 = 16C_{\mathcal{F}} + 32$ and $C_2 = 8/3$. ■

The following metric entropy bound is useful in the proofs of Theorem 16 and Lemma 17.

Lemma 24 *The $L_\infty(\mathbb{B}^d)$-metric entropy of $\mathcal{F}(m,1)$ satisfies*

$$\log\mathcal{N}(\delta, \mathcal{F}(m,1), \|\cdot\|_\infty) \le m(d+1)\log(1 + 2\sqrt{2}/\delta).$$

Proof For any two-layer ReLU networks $g(\cdot;\boldsymbol{\theta}^{(1)}), g(\cdot;\boldsymbol{\theta}^{(2)}) \in \mathcal{F}(m,1)$ with parameters $\boldsymbol{\theta}^{(1)} = \big(a_1^{(1)}, \dots, a_m^{(1)}, \mathbf{w}_1^{(1)T}, \dots, \mathbf{w}_m^{(1)T}\big)^T$ and $\boldsymbol{\theta}^{(2)} = \big(a_1^{(2)}, \dots, a_m^{(2)}, \mathbf{w}_1^{(2)T}, \dots, \mathbf{w}_m^{(2)T}\big)^T$, we can assume without loss of generality that $\|\mathbf{w}_k^{(j)}\|_2 = 1$ for all k, j, so that $g(\mathbf{x}; \boldsymbol{\theta}_j) \in \mathcal{F}(m, 1)$ is equivalent to $\sum_{k=1}^m |a_k^{(j)}| \leq 1$ for $j = 1, 2$. Note that $\|\widetilde{\mathbf{x}}\|_2 = \|(\mathbf{x}^T, 1)^T\|_2 \leq \sqrt{2}$. Then

$$\begin{aligned}
&|g(\mathbf{x}; \boldsymbol{\theta}^{(1)}) - g(\mathbf{x}; \boldsymbol{\theta}^{(2)})| \\
&\quad = \Big| \sum_{k=1}^m a_k^{(1)} \sigma(\widetilde{\mathbf{x}}^T \mathbf{w}_k^{(1)}) - \sum_{k=1}^m a_k^{(2)} \sigma(\widetilde{\mathbf{x}}^T \mathbf{w}_k^{(2)}) \Big| \\
&\quad \leq \Big| \sum_{k=1}^m (a_k^{(1)} - a_k^{(2)}) \sigma(\widetilde{\mathbf{x}}^T \mathbf{w}_k^{(1)}) \Big| + \Big| \sum_{k=1}^m a_k^{(2)} \big(\sigma(\widetilde{\mathbf{x}}^T \mathbf{w}_k^{(1)}) - \sigma(\widetilde{\mathbf{x}}^T \mathbf{w}_k^{(2)}) \big) \Big| \\
&\quad \leq \max_{1 \leq k \leq m} \sigma(\widetilde{\mathbf{x}}^T \mathbf{w}_k^{(1)}) \sum_{k=1}^m |a_k^{(1)} - a_k^{(2)}| + \sum_{k=1}^m |a_k^{(2)}| \max_{1 \leq k \leq m} \big\| \sigma(\widetilde{\mathbf{x}}^T \mathbf{w}_k^{(1)}) - \sigma(\widetilde{\mathbf{x}}^T \mathbf{w}_k^{(2)}) \big\|_2 \\
&\quad \leq \sqrt{2} \sum_{k=1}^m |a_k^{(1)} - a_k^{(2)}| + \sqrt{2} \max_{1 \leq k \leq m} \|\mathbf{w}_k^{(1)} - \mathbf{w}_k^{(2)}\|_2.
\end{aligned}$$

Let $\mathbb{B}_1^n$ denote the unit ℓ_1-ball in $\mathbb{R}^n$. The above inequality implies that, in order to cover $\mathcal{F}(m, 1)$ with respect to $\|\cdot\|_\infty$, we need only cover $\mathbb{B}_1^m$ with respect to $\|\cdot\|_1$ and m many $\mathbb{B}^d$ with respect to $\|\cdot\|_2$ simultaneously. By Example 5.8 of Wainwright (2019), we have

$$\log \mathcal{N}(\delta, \mathbb{B}_1^m, \|\cdot\|_1) \leq m \log(1 + 2/\delta), \quad \log \mathcal{N}(\delta, \mathbb{B}^d, \|\cdot\|_2) \leq d \log(1 + 2/\delta).$$

Thus,

$$\begin{aligned}
\log \mathcal{N}(\delta, \mathcal{F}(m, 1), \|\cdot\|_\infty) &\leq \log \mathcal{N}(\delta/\sqrt{2}, \mathbb{B}_1^m, \|\cdot\|_1) + \log \mathcal{N}(\delta/\sqrt{2}, \mathbb{B}^d, \|\cdot\|_2) \\
&\leq m(d+1) \log(1 + 2\sqrt{2}/\delta),
\end{aligned}$$

completing the proof. ■

Finally, we prove Lemma 17.

Proof of Lemma 17 As in the proof of Lemma 14, we have

$$\mathbb{E} Z_n(\gamma) \leq 16 \mathbb{E}_{\boldsymbol{\zeta}, \mathbf{x}} \sup_{f \in \mathcal{B}_{\mathcal{F}}(\gamma)} \frac{1}{n} \Big| \sum_{i=1}^n \zeta_i f(\mathbf{x}_i) \Big|.$$

Let $\{g_j\}_{j=1}^N$ be a minimal $(1/n)$-covering of $\mathcal{B}_{\mathcal{F}}(\gamma)$ with respect to the $L_\infty(\mathbb{B}^d)$-norm. Then, for any fixed $f \in \mathcal{B}_{\mathcal{F}}(\gamma)$, there exists some j^* such that $\|f - g_{j^*}\|_\infty \leq 1/n$. By the triangle inequality,

$$\begin{aligned}
\Big| \sum_{i=1}^n \zeta_i f(\mathbf{x}_i) \Big| &\leq \Big| \sum_{i=1}^n \zeta_i \big(f(\mathbf{x}_i) - g_{j^*}(\mathbf{x}_i) \big) \Big| + \max_{1 \leq j \leq N} \Big| \sum_{i=1}^n \zeta_i g_j(\mathbf{x}_i) \Big| \\
&\leq 1 + \max_{1 \leq j \leq N} \Big| \sum_{i=1}^n \zeta_i \frac{g_j(\mathbf{x}_i)}{\|g_j\|_n} \Big| \sqrt{\max_{1 \leq j \leq N} \|g_j\|_n^2} \\
&\equiv 1 + M_1 \sqrt{M_2}.
\end{aligned}$$

Applying Massart's lemma (Mohri et al., 2018, Theorem 3.7) yields

$$\mathbb{E}_{\boldsymbol{\zeta}} M_1 \leq \sqrt{2n \log(2N)}.$$

Moreover, since $g_j \in \mathcal{B}_{\mathcal{F}}(\gamma)$, we have $\max_j \|g_j\|_2 \leq \gamma$, and hence

$$M_2 \leq \gamma^2 + \max_{1 \leq j \leq N} \big| \|g_j\|_n^2 - \|g_j\|_2^2 \big|.$$

Note that $\max_j \|g_j\|_\infty \leq 2$ and $\max_j \|g_j^2\|_2^2 \leq 4 \max_j \|g_j\|_2^2 \leq 4\gamma^2$. Applying the maximal inequality for finite maxima (van der Vaart, 1998, Lemma 19.33), we have

$$\mathbb{E} \max_{1 \leq j \leq N} \big| \|g_j\|_n^2 - \|g_j\|_2^2 \big| \leq C_1 \bigg(\frac{\log N}{n} + \gamma \sqrt{\frac{\log N}{n}} \bigg)$$

for some constant $C_1 > 0$. Combining these pieces and using Jensen's inequality, we obtain

$$\begin{aligned} \frac{1}{n} \mathbb{E}_{\boldsymbol{\zeta}, \mathbf{x}} (M_1 \sqrt{M_2}) &= \frac{1}{n} \mathbb{E}_{\mathbf{x}} \{ \mathbb{E}_{\boldsymbol{\zeta}} (M_1 \sqrt{M_2} \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) \} \\ &\leq \sqrt{\frac{2 \log(2N)}{n}} \sqrt{\mathbb{E}_{\mathbf{x}} M_2} \leq C_2 \bigg(\gamma \sqrt{\frac{\log N}{n}} + \frac{\log N}{n} \bigg) \end{aligned}$$

for some constant $C_2 > 0$, and hence

$$\mathbb{E} Z_n(\gamma) \leq \frac{16}{n} + 16 C_2 \bigg(\gamma \sqrt{\frac{\log N}{n}} + \frac{\log N}{n} \bigg) \leq C_3 \bigg(\gamma \sqrt{\frac{\log N}{n}} + \frac{\log N}{n} \bigg) \tag{68}$$

for some constant $C_3 > 0$.

It remains to find an upper bound for the covering number $N \equiv \mathcal{N}(1/n, \mathcal{B}_{\mathcal{F}}(\gamma), \|\cdot\|_\infty)$. Since $\mathcal{B}_{\mathcal{F}}(\gamma) \subset \mathcal{F}^*(m, 1)$, it follows from Lemma 24 that

$$\begin{aligned} \log N &\leq \log \mathcal{N}(1/n, \mathcal{F}^*(m, 1), \|\cdot\|_\infty) = \log \mathcal{N}(1/n, \mathcal{F}(m, 1), \|\cdot\|_\infty) \\ &\leq m(d+1) \log(1 + 2\sqrt{2}n) \leq C_4 m d \log n \end{aligned}$$

for some constant $C_4 > 0$. Substituting into (68) concludes the proof. ■

References

B. Adlam and J. Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in Neural Information Processing Systems*, 33:11022–11032, 2020.

S. Arora, N. Cohen, and E. Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning*, pages 244–253, 2018.

S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 322–332, 2019.

F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reply to loog et al.: Looking beyond the peaking phenomenon. *Proceedings of the National Academy of Sciences*, 117(20):10627, 2020a.

M. Belkin, D. Hsu, and J. Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020b.

C. M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

T. Bos and J. Schmidt-Hieber. Convergence rates of deep ReLU networks for multiclass classification. *Electronic Journal of Statistics*, 16(1):2724–2773, 2022.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

J. Bourgain, J. Lindenstrauss, and V. Milman. Approximation of zonoids by zonotopes. *Acta Mathematica*, 162(1):73–141, 1989.

G. Brown and R. Ali. Bias/variance is not the same as approximation/estimation. *Transactions on Machine Learning Research*, 2024.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Berlin, 2011.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

L. Chen, M. Lukasik, W. Jitkrittum, C. You, and S. Kumar. On bias-variance alignment in deep models. In *International Conference on Learning Representations*, 2024.

G. Chinot, M. Löffler, and S. van de Geer. On the robustness of minimum norm interpolators and regularized empirical risk minimizers. *The Annals of Statistics*, 50(4):2306–2333, 2022.

L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32:2937–2947, 2019.

T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14 (3):326–334, 1965.

Z. Deng, A. Kammoun, and C. Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference*, 11(2):435–495, 2022.

A. Derumigny and J. Schmidt-Hieber. On lower bounds for the bias-variance trade-off. *The Annals of Statistics*, 51(4):1510–1533, 2023.

X. Dou and T. Liang. Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. *Journal of the American Statistical Association*, 116(535):1507–1520, 2021.

W. E and S. Wojtowytsch. On the Banach spaces associated with multi-layer ReLU networks: Function representation, approximation theory and gradient descent dynamics. *CSIAM Transactions on Applied Mathematics*, 1(3):387–440, 2020.

W. E, C. Ma, and L. Wu. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019.

W. E, C. Ma, and L. Wu. A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Science China Mathematics*, 63(7):1235–1258, 2020.

B. Efron. Prediction, estimation, and attribution. *Journal of the American Statistical Association*, 115(530):636–655, 2020.

T. Ergen and M. Pilanci. Revealing the structure of deep neural networks via convex duality. In *Proceedings of the 38th International Conference on Machine Learning*, pages 3004–3014, 2021.

A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542 (7639):115–118, 2017.

M. H. Farrell, T. Liang, and S. Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models.* Cambridge University Press, Cambridge, 2016.

N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. *Information and Inference*, 9(2):473–504, 2020.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning.* MIT Press, Cambridge, MA, 2016.

E. Greenshtein. Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint. *The Annals of Statistics*, 34(5):2367–2386, 2006.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, 2nd edition, 2009.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

S. Hayakawa and T. Suzuki. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks*, 123:343–361, 2020.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

R. W. Hoerl. Ridge regression: A historical context. *Technometrics*, 62(4):420–425, 2020.

D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):1–6, 2012.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pages 8580–8589, 2018.

K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *IEEE 12th International Conference on Computer Vision*, pages 2146–2153, 2009.

Z. Ji and M. Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019.

Z. Ji and M. Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. In *International Conference on Learning Representations*, 2020.

L. K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics*, 20(1):608–613, 1992.

J. M. Klusowski and A. R. Barron. Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Transactions on Information Theory*, 64(12):7649–7656, 2018.

M. Kohler and S. Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012.

A. Krogh and J. A. Hertz. A simple weight decay can improve generalization. *Advances in Neural Information Processing Systems*, 4:950–957, 1991.

X. Li and X.-L. Meng. A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association*, 116(533):353–367, 2021.

T. Liang and P. Sur. A precise high-dimensional asymptotic theory for boosting and minimum-ℓ_1-norm interpolated classifiers. *The Annals of Statistics*, 50(3):1669–1695, 2022.

T. Liang, A. Rakhlin, and X. Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Proceedings of the 33rd Conference on Learning Theory*, pages 2683–2711, 2020.

L. Lin and E. Dobriban. What causes the test error? Going beyond bias-variance via ANOVA. *Journal of Machine Learning Research*, 22(155):1–82, 2021.

F. Liu, Z. Liao, and J. A. K. Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 649–657, 2021.

T. Luo, Z.-Q. J. Xu, Z. Ma, and Y. Zhang. Phase diagram for two-layer ReLU neural networks at infinite-width limit. *Journal of Machine Learning Research*, 22(71):1–47, 2021.

H. Maennel, O. Bousquet, and S. Gelly. Gradient descent quantizes ReLU network features. *arXiv preprint arXiv:1803.08367*, 2018.

Y. Makovoz. Random approximants and neural networks. *Journal of Approximation Theory*, 85(1):98–109, 1996.

J. Matoušek. Improved upper bounds for approximation by zonotopes. *Acta Mathematica*, 177(1):55–73, 1996.

S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, Cambridge, MA, 2018.

A. Montanari and Y. Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5): 2816–2847, 2022.

V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai. Harmless interpolation of moisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 67–83, 2020.

P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021a.

P. Nakkiran, P. Venkat, S. Kakade, and T. Ma. Optimal regularization can mitigate double descent. In *International Conference on Learning Representations*, 2021b.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

B. Neyshabur, R. Tomioka, and N. Srebro. Norm-based capacity control in neural networks. In *Proceedings of the 28th Conference on Learning Theory*, pages 1376–1401, 2015a.

B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations*, 2015b.

B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019.

G. Ongie, R. Willett, D. Soudry, and N. Srebro. A function space view of bounded norm infinite width ReLU nets: The multivariate case. In *International Conference on Learning Representations*, 2020.

R. Parhi and R. D. Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021.

R. Parhi and R. D. Nowak. Near-minimax optimal estimation with shallow ReLU neural networks. *IEEE Transactions on Information Theory*, 69(2):1125–1140, 2023.

M. Pilanci and T. Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7695–7705, 2020.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pages 1177–1184, 2007.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

G. M. Rotskoff and E. Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and D. Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

J. W. Siegel. Optimal approximation of zonoids and uniform approximation by shallow neural networks. *Constructive Approximation*, 62(2):441–469, 2025.

J. W. Siegel and J. Xu. Approximation rates for neural networks with general activation functions. *Neural Networks*, 128:313–321, 2020.

J. W. Siegel and J. Xu. Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *Foundations of Computational Mathematics*, 24(2): 481–537, 2024.

J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

J. Sjöberg and L. Ljung. Overtraining, regularization and searching for a minimum, with application to neural networks. *International Journal of Control*, 62(6):1391–1407, 1995.

M. Soltanolkotabi, A. Javanmard, and J. D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2019.

N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, volume 17, pages 1329–1336, 2004.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27, pages 3104–3112, 2014.

T. Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Cambridge, 2019.

Y. Wang, J. Lacotte, and M. Pilanci. The hidden convex optimization landscape of regularized two-layer ReLU networks: An exact characterization of optimal solutions. In *International Conference on Learning Representations*, 2022.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

Z. Yang, Y. Yu, C. You, J. Steinhardt, and Y. Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10767–10777, 2020.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67, 2006.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.