NONASYMPTOTIC THEORY FOR TWO-LAYER NEURAL NETWORKS: BEYOND THE BIAS–VARIANCE TRADE-OFF

By Huiyuan Wang[a] and Wei Lin[b]

School of Mathematical Sciences and Center for Statistical Science, Peking University, [a]*huiyuan.wang@pku.edn.cn,*
[b]*weilin@math.pku.edu.cn*

Large neural networks have proved remarkably effective in modern deep learning practice, even in the overparametrized regime where the number of active parameters is much larger than 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 all network widths, which reproduce the double descent phenomenon. Moreover, the overparametrized minimum risk is lower than the underparametrized minimum risk when the signal is strong, and nearly attains the minimax optimal rate 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.

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, Vinyals and Le, 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 model parameters is much larger than the number of training samples. For example, the AlexNet (Krizhevsky, Sutskever and Hinton, 2012) involves 60 million parameters trained on 1.2 million images; the model achieving state-of-the-art performance on the ImageNet dataset as of 2022 has reached 2.1 billion parameters (Yu et al., 2022).

Theoretical insights into overparametrized neural networks have been obtained from the optimization viewpoint (Arora, Cohen and Hazan, 2018; Soltanolkotabi, Javanmard and Lee, 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 (or approximation error) and variance (or estimation error). The so-called "double descent" curve (Belkin et al., 2019) 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.

MSC2020 subject classifications: Primary 62G08; secondary 62J07, 68T07.

Keywords and phrases: Double descent, generalization, neural network, nonparametric regression, overparametrization, regularization.

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, Löffler and van de Geer, 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, Bengio and Courville, 2016, p. 277).

Neural networks are pure prediction algorithms in the sense of Efron (2020), which operate in a nonparametric and nonparsimonious way. The nonparametric view of 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 fast and near-minimax rates of convergence and circumvent the curse of dimensionality (Schmidt-Hieber, 2020; Hayakawa and Suzuki, 2020; Farrell, Liang and Misra, 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, Tomioka and Srebro (2015a) and Golowich, Rakhlin and Shamir (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 the underparametrized minimum risk 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

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

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 $\boldsymbol{\theta} = (a_1, \dots, a_m, \mathbf{v}_1^T, \dots, \mathbf{v}_m^T, b_1, \dots, b_m)^T$. By appropriately restricting the function class to which f^* belongs, we do not include an intercept in the output. Assumptions on f^*, $\mathbf{x}_i$, and ε_i are detailed in Section 2.3.

By incorporating a *scaled variation* regularizer to be defined in Section 2.2, our main result (Theorem 5) shows that the prediction (or generalization) error of the regularized network estimator $g(\cdot;\widehat{\boldsymbol{\theta}})$ is of order

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

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 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 (16). We further prove that the overparametrized minimum risk is nearly minimax rate-optimal over a suitable class of functions (Theorem 6). By contrast, overparametrized random feature models suffer from the curse of dimensionality and thus are suboptimal (Proposition 5). 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, which is known as (standard) *weight decay* in deep learning (Krogh and Hertz, 1991) and, in general, does not induce sparsity. On the other hand, a linearization of the ReLU function by parameter space partitioning transforms the regularized problem into a group lasso. This gives a key insight

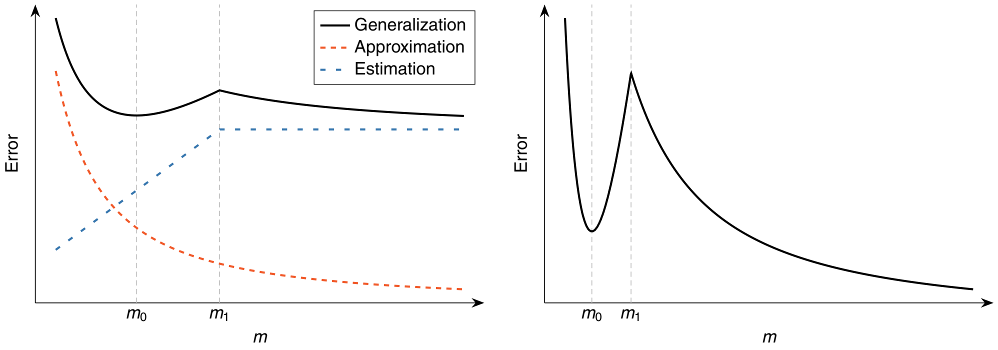

FIG 1. *Risk curves for varying network width* m *from the prediction bound* (3) *with* $\|f^*\|_{\mathcal{S}}^2/\sigma_\varepsilon^2 = 1$, $d = 6$, *and* $n = 1000$. *The left panel shows the decomposition of prediction error into approximation and estimation errors. The right panel shows the same plot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

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.

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 (2022) 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, 2022). 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 (2022) to allow the identification of affine functions.

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 empirical risk minimization problem. For example, Bach (2017) and Parhi and Nowak (2022) considered a variational problem by constraining the network estimator to within a ball in the function space; Parhi and Nowak (2022) showed that such network estimators are nearly minimax optimal. A representer theorem of Parhi and Nowak (2021) ensures the existence of a solution to the variational problem in the form of a finite-width network with a skip connection. However, the network width of the solution is required to be smaller than the sample size, thus providing no clue about the effect of overparametrization. One exception is the work of E, Ma and Wu (2019), which obtained generalization bounds for finite-width 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.

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, Montanari and Nguyen, 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, Gabriel and Hongler, 2018) and the associated kernel methods. This "lazy training" regime (Chizat, Oyallon and Bach, 2019) entails a large performance gap between realistic and linearized networks and hence does not explain the power of fully trained neural networks (E, Ma and Wu, 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 the effort has focused on linear regression and, in particular, minimum norm interpolators and ridge estimators, and has recovered the phenomenon under specific generative models for the random predictors (e.g., Belkin, Hsu and Xu, 2020; Hastie et al., 2022; Muthukumar et al., 2020). 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 been derived for random feature models (Mei and Montanari, 2022) and classification problems (Deng, Kammoun and Thrampoulidis, 2022; Liang and Sur, 2022). Li and Meng (2021) and Liang, Rakhlin and Zhai (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 the emergence of double descent from the oversimplified model assumptions and asymptotic regimes. Also, it remains elusive how these theories extend to realistic neural networks and fit in with our current understanding of the bias–variance trade-off (Geman, Bienenstock and Doursat, 1992; Derumigny and Schmidt-Hieber, 2023).

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 Appendix and Supplemental Material.

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 $\mathbb{S}^{d-1}$ be the ℓ_2 unit ball and unit 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 $\|f\|_{L_\infty(D)}$ denote the L_∞-norm on D, and $\|f\|_2$ and $\|f\|_n$ the L_2-norm under the distribution μ 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. Define the *scaled variation norm* of the finite-width network $g(\cdot;\boldsymbol{\theta})$ by

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

where $\mathbf{w}_k = (\mathbf{v}_k^T, b_k)^T$. This regularizer was also considered by Parhi and Nowak (2021) and Parhi and Nowak (2022) for two-layer ReLU networks. It coincides with the ℓ_1 path norm proposed by Neyshabur, Tomioka and Srebro (2015a) when d degenerates to zero. For any $d \geq 1$, however, it is not separable in the first-layer weights and hence not a path norm. We will show in Section 3 that the scaled variation regularizer (4) has some desirable properties that are key to our theoretical analysis.

The network (2) 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

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

Motivated by this observation, we can naturally represent an infinite-width two-layer ReLU network associated with a signed measure $\alpha \in \mathcal{M}(\mathbb{R}^{d+1})$ as

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

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 $\mathcal{M}_2(\mathbb{R}^{d+1}) \equiv \big\{\alpha \in \mathcal{M}(\mathbb{R}^{d+1}) : \int_{\mathbb{R}^{d+1}} \|\mathbf{v}\|_2 \, d|\alpha|(\mathbf{w}) < \infty\big\}$. Moreover, functions in $\mathcal{G}$ are exactly those that can be approximated by two-layer ReLU networks with finite scaled variation norm. A formal statement is given by Proposition S.1 in the Supplementary Material. 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

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

Clearly, the $\mathcal{S}$-norm is a functional version of the scaled variation norm except for the omission of the bias term owing to the centering by $\sigma(b)$. In fact, the $\mathcal{S}$-norm of a finite-width two-layer ReLU network $g(\cdot;\boldsymbol{\theta})$ is $\|g(\cdot;\boldsymbol{\theta})\|_{\mathcal{S}} = \sum_{k=1}^m |a_k|\|\mathbf{v}_k\|_2$, which is bounded above by the scaled variation norm (4).

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 bounded variation spaces in the Radon domain considered by Ongie et al. (2020) and Parhi and Nowak (2021) and the variation spaces considered by Bach (2017) and Siegel and Xu (2022), which in turn contain the spectral Barron spaces and Sobolev spaces (Klusowski and Barron, 2018; Parhi and Nowak, 2022). Our definition of the $\mathcal{S}$-norm is more transparent in that 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 $f_{\boldsymbol{\beta}}(\mathbf{x}) = \boldsymbol{\beta}^T\mathbf{x} + c$, $\|f_{\boldsymbol{\beta}}\|_{\mathcal{S}} = 2\|\boldsymbol{\beta}\|_2$ 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 Supplementary Material E.

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 $\mathcal{G}$ to $\mathbb{B}^d$; denote the space of such restrictions by $\mathcal{G}(\mathbb{B}^d)$. An important consequence from Corollary S.1 in the Supplementary Material is that, for any $f \in \mathcal{G}(\mathbb{B}^d)$, there exists a signed measure $\widetilde{\alpha}_f \in \mathcal{M}(\mathbb{S}^{d-1} \times [-1,1])$ such that

$$f(\mathbf{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} \sigma(\mathbf{v}^T\mathbf{x} + b)\, d\widetilde{\alpha}_f(\mathbf{w}) + c, \quad \mathbf{x} \in \mathbb{B}^d. \tag{6}$$

Compared with a similar integral representation in Parhi and Nowak (2022, Remark 3), note that no skip connection appears in (6). 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); Siegel and Xu (2022) showed that this rate is sharp and not improvable. We have the following approximation result for functions in $\mathcal{G}_M$, which is a direct consequence of Bach (2017, Proposition 1) and the integral representation (6).

THEOREM 1. *For any $f \in \mathcal{G}_M$, 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*

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

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

The construction in Theorem 1 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.

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

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

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

$$\widehat{\boldsymbol{\theta}} = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_m} J_n(\boldsymbol{\theta};\lambda). \tag{7}$$

In a related work, Parhi and Nowak (2022) studied a variational problem in the variation space associated with two-layer ReLU networks, where regularization is imposed as a constraint on the variation norm of the network function. A representer theorem guarantees the existence of a finitely supported solution of width $m \le n-(d+1)$ to the variational problem. However, the finite-dimensional network learning problem is equivalent to the variational problem only when $m \ge n-(d+1)$. 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 (7) 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 (7), 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 (7) as the ℓ_2-regularized ERM problem

$$\widehat{\boldsymbol{\theta}}_{\ell_2} = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_m}\left\{\frac{1}{2n}\sum_{i=1}^{n}\big(y_i - g(\mathbf{x}_i;\boldsymbol{\theta})\big)^2 + \frac{\lambda}{2}\sum_{k=1}^{m}(a_k^2 + \|\mathbf{w}_k\|_2^2)\right\}. \tag{8}$$

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 (8) 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 1. *Any solution $\widehat{\boldsymbol{\theta}}_{\ell_2}$ to the optimization problem* (8) *is a solution to the problem* (7). *Conversely, if $\widehat{\boldsymbol{\theta}}$ is a solution to the optimization problem* (7), *then $\mathcal{T}_1(\widehat{\boldsymbol{\theta}})$ is a solution to the problem* (8).

Proposition 1 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 2. *Consider the gradient flow for the optimization problem* (7) *defined by*

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

and for the problem (8) *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.*

Observations similar to Proposition 1 have been noted in slightly different forms by, for example, Neyshabur, Tomioka and Srebro (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." The messages of the above results are twofold. First, since ℓ_2 regularization does not induce entrywise sparsity in the parameters (but see Srebro, Rennie and Jaakkola (2004) for an unusual example where it does induce sparsity in spectral structures), we are assured that a sufficiently wide two-layer network can be intrinsically overparametrized. Second, 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 (7) 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, is at most

$$2\sum_{j=0}^{d}\binom{n-1}{j} = O\left(d\left(\frac{en}{d}\right)^d\right), \tag{9}$$

which 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 3. *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* (7)*, 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

$$\widetilde{a}_k = s_j, \quad \widetilde{\mathbf{w}}_k = \frac{1}{|S_j|} \sum_{\ell \in S_j} |a_\ell| \mathbf{w}_\ell, \quad k \in S_j \tag{10}$$

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. Proposition 3 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 3 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

$$\sum_{k=1}^{m} a_k \sigma(\mathbf{X}\mathbf{w}_k) = \sum_{j=1}^{2p} \mathbf{D}_j \mathbf{X} \boldsymbol{\beta}_j(\boldsymbol{\theta}), \tag{11}$$

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 (7) 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 4. *For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}_m$, the reparametrization $\widetilde{\boldsymbol{\theta}} = \mathcal{T}_2(\boldsymbol{\theta})$ defined in* (10) *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* (7) *satisfies*

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

The group lasso formulation allows for borrowing ideas from high-dimensional statistics to derive generalization bounds. 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, Lacotte and Pilanci (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 4, which places no restriction on the minimum network width.

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 nearly 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 (7), 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, 2022) can be effective and give sharp bounds. Moving into the overparametrized 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 2. *Under Conditions* (C1), (C3), *and the assumptions that* $\|\mathbf{x}_i\|_2 \le 1$ *and* $\sqrt{en} > d$, *the regularized network estimator* $g(\cdot;\widehat{\boldsymbol{\theta}})$ *with* $\lambda = C_1\sigma_\varepsilon\sqrt{d\log(en/d)/n}$ *satisfies*

$$\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n^2 \le C\left\{\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2)\sqrt{\frac{d\log(en/d)}{n}}\right\} \tag{12}$$

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

$$\mathbb{E}\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_n^2 \le C\left\{\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2)\sqrt{\frac{d\log(en/d)}{n}}\right\}, \tag{13}$$

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 2 differs from the standard group lasso theory for sparse linear regression in two aspects. First, it requires an extension of the theory to the case where the linear model is only approximate, as discussed in, for example, Bühlmann and van de Geer (2011, Section 6.2.3). Here the linear model represents the reparametrized two-layer neural network, whose approximation error has been given by Theorem 1. Second, and more importantly, there is no guarantee that this linear model will be sparse or its design will satisfy a compatibility or restricted eigenvalue condition that is often imposed in high-dimensional linear regression. As a result, we can only prove a prediction bound at a slower rate, which is analogous to those for the lasso obtained by Bühlmann and van de Geer (2011, Corollary 6.1) and Bartlett, Mendelson and Neeman (2012).

The error bounds in Theorem 2 decompose into a bias term or approximation error that arises from using a finite-width neural network to approximate the nonparametric model (1), and a variance term or 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 bias always decreases while the variance remains constant. To appreciate why this is possible, note first that the variance scales as $O(\sqrt{\log p/n})$ as a consequence of the lack of parameter identifiability. Moreover, the *effective* dimension p is bounded by $O(d(en/d)^d)$ from (9), 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 variance. This result extends beyond the classical bias–variance trade-off and demonstrates the virtues of overparametrization in two-layer neural networks.

Combining Theorem 2 with a maximal inequality, we obtain similar bounds on the prediction error of the regularized network estimator.

THEOREM 3. *Under Conditions* (C1)–(C3)*, if* $m \geq C_1(n\log n/d)^{d/(2(d+3))}$*, then the regularized network estimator* $g(\cdot;\widehat{\boldsymbol{\theta}})$ *with* $\lambda = \lambda_1 \equiv C_2\sigma_\varepsilon\sqrt{d\log(en/d)/n}$ *satisfies*

$$\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_2^2 \leq C\left\{\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2)\sqrt{\frac{d\log(en/d)}{n}}\right\} \tag{14}$$

with probability at least $1 - O(n^{-C_3})$*, and*

$$\mathbb{E}\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*\|_2^2 \leq C\left\{\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + (\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2)\sqrt{\frac{d\log(en/d)}{n}}\right\}, \tag{15}$$

for some constants $C_1, C_2, C_3, C > 0$ *and large enough* 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, Ma and Wu (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 is algorithm-independent and is valid for any global optimum.

E, Ma and Wu (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

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

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 (9) 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 bias–variance trade-off.

THEOREM 4. *Under Conditions* (C1)–(C3)*, if* $m < n/(d \log(en/d))$*, then the regularized network estimator* $g(\cdot; \widehat{\boldsymbol{\theta}})$ *with* $\lambda = \lambda_2 \equiv C_1 \sigma_\varepsilon \max\{m^{-(d+3)/d}, md \log(en/d)/n\}$ *satisfies*

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

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

The proof technique used for Theorem 4 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, Liang and Misra (2021), and Parhi and Nowak (2022).

Finally, noting that the ranges of allowable m in Theorems 3 and 4 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 5. *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 3 and 4, respectively, satisfies*

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

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

The implications of Theorem 5 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 bias and variance. 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 $m_1 \asymp \sqrt{n/(d\log n)}$, the intrinsic model complexity and hence the variance of the network estimator become saturated and remain constant, while the bias 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

$$\kappa \equiv \frac{\|f^*\|_{\mathcal{S}}^2}{\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2} > \left(\frac{1}{2}\right)^{(2d+3)/d} \left(\frac{n}{d\log n}\right)^{3/(2d)}. \tag{16}$$

When $d \gg \log n$, the above condition approximately becomes $\kappa > 1/4$, or the signal-to-noise ratio $\|f^*\|_{\mathcal{S}}^2/\sigma_\varepsilon^2 = \kappa/(1-\kappa) > 1/3$. An example with $\kappa = 1$, $d = 6$, and $n = 1000$ was given in Figure 1. This makes intuitive sense since the reduction in approximation error plays a more important role when the signal is stronger. 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.

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, 2022). In fact, the underparametrized result (Theorem 4) 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 nearly minimax optimal over the larger function class $\mathcal{G}$.

To gain intuition for the optimal rate, for any probability measure ρ on $\mathbb{S}^{d-1} \times [-1,1]$, we consider the reproducing kernel Hilbert space (RKHS)

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

associated with the kernel

$$H_\rho(\mathbf{x},\mathbf{z}) = \mathbb{E}_{\mathbf{w}\sim\rho}\big(\sigma(\mathbf{v}^T\mathbf{x}+b)\sigma(\mathbf{v}^T\mathbf{z}+b)\big). \tag{17}$$

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 $\gamma > 1$. Noting that $\mathcal{H}_\rho \subset \mathcal{G}$ for all ρ and letting $\gamma \to 1$, we see that the desired minimax optimal rate should be $n^{-1/2}$. This heuristic argument is formalized in the following minimax result.

THEOREM 6. *Assume that* $\mathbf{x}_i \sim \mathrm{Uniform}(\mathbb{B}^d)$ *and* $\varepsilon_i \sim N(0,1)$. *Then there exists a constant* $C > 0$ *such that*

$$\inf_{\widehat{f}} \sup_{f^*\in\mathcal{G}} \mathbb{E}\|\widehat{f} - f^*\|_2^2 \geq \frac{C}{\sqrt{n\log n}},$$

where the infimum is taken over all estimators.

This result says that the upper bounds on the overparametrized minimum risk in Theorems 3 and 5 are sharp up to logarithmic factors. 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. Random feature models (Rahimi and Recht, 2007) provide a stochastic approximation to kernel methods by first mapping the input into a randomized feature space and then applying standard linear methods. Alternatively, they can be interpreted 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 on $\mathbb{S}^{d-1} \times [-1, 1]$ 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 ridge estimator $\widehat{\mathbf{a}}(\lambda) = (\mathbf{K} + n\lambda \mathbf{I}_n)^{-1} \mathbf{y}$, 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 (17). The following result establishes a lower bound on the worst-case performance of optimally tuned ridge estimators in random feature models.

PROPOSITION 5. *Under Conditions* (C2) *and* (C3)*, there exists a universal constant* $C > 0$ *such that*

$$\sup_{f^* \in \mathcal{G}_M} \inf_{\lambda > 0} \mathbb{E} \|h_{\rho_0}(\cdot; \widehat{\mathbf{a}}(\lambda)) - f^*\|_2^2 \geq \frac{CM}{d(\min(m, n))^{1/d}}.$$

The proof of Proposition 5 builds on an approximation result of Barron (1993, Theorem 6) for linear subspaces with fixed basis functions. Similar lower bounds have been obtained by E, Ma and Wu (2020) for random feature models trained by gradient descent with noiseless data. The exponential dependence of the rate on d manifests the curse of dimensionality in random feature models for learning functions beyond an RKHS.

6. Discussion. The debate over double descent and the virtues of overparametrization casts a cloud over 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, Rakhlin and Zhai (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 (16), 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., 2021; 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. (2020), 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 of bias and variance in the overparametrized regime (Hastie et al., 2022; Mei and Montanari, 2022): while the variance always decreases, the bias may increase (for well-specified linear models), decrease (for random feature models), or first decrease and then increase (for misspecified linear models). These somewhat peculiar behaviors are partly due to the fact that the ground truth in these settings is parametric and varying with the dimensionality. Neural networks, however, are intrinsically nonparametric, and the bias–variance trade-off should be discussed within this framework (Geman, Bienenstock and Doursat, 1992). Embracing this viewpoint, our results show that the bias always decreases, while the variance remains constant after the saturation threshold. Although there is no more trade-off in the overparametrized regime, the general principle of bias–variance trade-off in the sense of Derumigny and Schmidt-Hieber (2023) still seems to hold: the variance is lower bounded if the bias is small.

Our framework may be extended in several directions. The most important would be the development for deep neural networks, by following the idea of 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. 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. Finally, it would be worthwhile to extend our theory to classification problems and more network architectures such as convolutional and recurrent neural networks.

APPENDIX A: PROOFS FOR SECTION 3

In this appendix we provide the proofs of Propositions 1–4. To simplify the notation, we write $\widetilde{\mathbf{x}}_i = (\mathbf{x}_i^T, 1)^T$.

PROOF OF PROPOSITION 1. Let $\widehat{\boldsymbol{\theta}}_{\ell_2}$ be an arbitrary solution to problem (8) and $\widehat{\boldsymbol{\theta}}$ an arbitrary solution to problem (7). By the optimality of $\widehat{\boldsymbol{\theta}}_{\ell_2}$ and $\widehat{\boldsymbol{\theta}}$, we have

$$J_n^{\ell_2}(\widehat{\boldsymbol{\theta}}_{\ell_2};\lambda) \le J_n^{\ell_2}(\mathcal{T}_1(\widehat{\boldsymbol{\theta}});\lambda), \qquad J_n(\widehat{\boldsymbol{\theta}};\lambda) \le J_n(\widehat{\boldsymbol{\theta}}_{\ell_2};\lambda), \tag{18}$$

where $J_n^{\ell_2}(\boldsymbol{\theta};\lambda)$ is the objective function of (8). 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 (18) 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 2. By direct differentiation, the gradient flow $d\boldsymbol{\theta}(t)/dt = -\nabla_{\boldsymbol{\theta}} J_n(\boldsymbol{\theta}(t);\lambda)$ for problem (7) can be written as

$$\frac{d}{dt}a_j(t) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - g(\mathbf{x}_i;\boldsymbol{\theta}(t))\big)^2\sigma(\widetilde{\mathbf{x}}_i^T\mathbf{w}_j(t)) - \lambda\|\mathbf{w}_j(t)\|_2\partial|a_j(t)|, \tag{19}$$

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

for $j = 1,\dots,m$, 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 (19) by $a_j(t)$ and (20) by $\mathbf{w}_j(t)^T$ gives

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

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 $t \ge 0$,

$$|a_j(t)|^2 = \|\mathbf{w}_j(t)\|_2^2, \quad j = 1,\dots,m. \tag{21}$$

Similarly, the gradient flow for problem (8) can be written as

$$\frac{d}{dt} a_j^{\ell_2}(t) = \frac{1}{n}\sum_{i=1}^{n} \big(y_i - g(\mathbf{x}_i; \boldsymbol{\theta}_{\ell_2}(t))\big)^2 \sigma(\widetilde{\mathbf{x}}_i^T \mathbf{w}_j^{\ell_2}(t)) - \lambda a_j^{\ell_2}(t), \tag{22}$$

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

for $j = 1, \dots, m$. Using (21) 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 (19)–(20) and (22)–(23) are identical. Initialized at the same point, their trajectories must coincide. □

To prove Propositions 3 and 4, we first introduce the following lemma, which says that the ReLU function is linear over each cone Q_j.

LEMMA 1. *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

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

PROOF OF PROPOSITION 3. 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 1 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 4. By Lemma 1 and the definition of $\widetilde{\boldsymbol{\theta}} = \mathcal{T}_2(\boldsymbol{\theta})$ in (10), we have

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

and

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

$$= \sum_{j=1}^{2p} \Big\| \sum_{k\in S_j} |a_k| \mathbf{w}_k \Big\|_2 \le \sum_{k=1}^{m} |a_k| \|\mathbf{w}_k\|_2 = \nu(\boldsymbol{\theta}).$$

To prove the second assertion, from (11) we have, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}_m$,

$$\frac{1}{2n}\Big\| \mathbf{y} - \sum_{k=1}^{m} a_k \sigma(\mathbf{X}\mathbf{w}_k) \Big\|_2^2 = \frac{1}{2n}\Big\| \mathbf{y} - \sum_{j=1}^{2p} \mathbf{D}_j \mathbf{X} \boldsymbol{\beta}_j(\boldsymbol{\theta}) \Big\|_2^2. \tag{24}$$

Also, by the collinearity property of $\widehat{\boldsymbol{\theta}}$ from Proposition 3,

$$\|\mathbf{B}(\widehat{\boldsymbol{\theta}})\|_{2,1} = \sum_{j=1}^{2p} \Big\| \sum_{k\in S_j} |\widehat{a}_k| \widehat{\mathbf{w}}_k \Big\|_2 = \sum_{k=1}^{m} |\widehat{a}_k| \|\widehat{\mathbf{w}}_k\|_2 = \nu(\widehat{\boldsymbol{\theta}}). \tag{25}$$

Combining (24) and (25) yields the expression for $J_n(\widehat{\boldsymbol{\theta}}; \lambda)$. □

APPENDIX B: PROOFS OF RESULTS IN THE OVERPARAMETRIZED REGIME

In this appendix we provide the proofs of the high-probability bounds (12) and (14) for the regularized estimator under the overparametrized regime. Their expectation counterparts (13) and (15) are deferred to Supplementary Material C.

We first introduce some notation. We write $\overline{p} \equiv d(en/d)^d$ and define the class of two-layer ReLU networks with a bounded scaled variation norm as $\mathcal{F}(m, F) = \big\{g(\mathbf{x}; \boldsymbol{\theta}) : \nu(\boldsymbol{\theta}) \le F\big\}$. For any $f^* \in \mathcal{G}_M$, denote by $g(\cdot; \boldsymbol{\theta}^*)$ the best approximation of f^* under the $L_\infty(\mathbb{B}^d)$ norm in Theorem 1, where $\boldsymbol{\theta}_m^* = (a_1^*, \dots, a_m^*, \mathbf{w}_1^{*T}, \dots, \mathbf{w}_m^{*T})^T$.

PROOF OF (12) IN THEOREM 2. By the optimality of $\widehat{\boldsymbol{\theta}}$, we have

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

Rearranging terms gives

$$\begin{aligned}
&\frac{1}{2}\|g(\cdot; \widehat{\boldsymbol{\theta}}) - f^*(\cdot)\|_n^2 \\
&\le \lambda\big(\nu(\boldsymbol{\theta}^*) - \nu(\widehat{\boldsymbol{\theta}})\big) + \frac{1}{2}\|g(\cdot; \boldsymbol{\theta}^*) - f^*(\cdot)\|_n^2 + \frac{1}{n}\Big|\sum_{i=1}^{n} \varepsilon_i\big\{g(\mathbf{x}_i; \widehat{\boldsymbol{\theta}}) - g(\mathbf{x}_i; \boldsymbol{\theta}^*)\big\}\Big| \\
&\equiv T_1 + T_2 + T_3.
\end{aligned} \tag{26}$$

Write $\mathbf{B}(\boldsymbol{\theta}^*)$ as $\mathbf{B}^* = (\boldsymbol{\beta}_1^*, \dots, \boldsymbol{\beta}_{2p}^*)$ and $\mathbf{B}(\widehat{\boldsymbol{\theta}})$ as $\widehat{\mathbf{B}} = (\widehat{\boldsymbol{\beta}}_1, \dots, \widehat{\boldsymbol{\beta}}_{2p})$. Since we need only evaluate $g(\cdot; \boldsymbol{\theta}^*)$ on the training sample, by Proposition 4 we assume without loss of generality $\nu(\boldsymbol{\theta}^*) = \|\mathbf{B}^*\|_{2,1}$. It also follows from Proposition 4 that $\nu(\widehat{\boldsymbol{\theta}}) = \|\widehat{\mathbf{B}}\|_{2,1}$. These facts, together with the triangle inequality, give

$$T_1 = \lambda\big(\|\mathbf{B}^*\|_{2,1} - \|\widehat{\mathbf{B}}\|_{2,1}\big) \le 2\lambda\|\mathbf{B}^*\|_{2,1} - \lambda\|\mathbf{B}^* - \widehat{\mathbf{B}}\|_{2,1}. \tag{27}$$

To bound T_2, applying Theorem 1 yields

$$T_2 = \frac{1}{2n}\sum_{i=1}^{n}\big(g(\mathbf{x}_i;\boldsymbol{\theta}^*) - f^*(\mathbf{x}_i)\big)^2 \le C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} \tag{28}$$

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

$$T_3 = \frac{1}{n}\left|\boldsymbol{\varepsilon}^T\sum_{i=1}^{2p}\mathbf{D}_i\mathbf{X}(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^*)\right| \le n^{-1/2}\max_{1\le j\le 2p}\|\mathbf{v}_j\|_2\sum_{i=1}^{2p}\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^*\|_2, \tag{29}$$

where $\mathbf{v}_j^T = \boldsymbol{\varepsilon}^T\mathbf{D}_j\mathbf{X}/\sqrt{n}$ and $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)^T$. Thus, combining (27)–(29), choosing $\lambda \ge 2n^{-1/2}\max_j\|\mathbf{v}_j\|_2$, and noting that $\nu(\boldsymbol{\theta}^*) \le 6\|f^*\|_{\mathcal{S}}$ in Theorem 1, we obtain

$$\begin{aligned}
&\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*(\cdot)\|_n^2 \\
&\le 2C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + 4\lambda\nu(\boldsymbol{\theta}^*) + 2\left(\frac{1}{\sqrt{n}}\max_{j=1,\ldots,p}\|\mathbf{v}_j\|_2 - \lambda\right)\|\mathbf{B}^* - \widehat{\mathbf{B}}\|_{2,1} \qquad (30)\\
&\le 2C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + 24\lambda\|f^*\|_{\mathcal{S}}.
\end{aligned}$$

It remains to bound $n^{-1/2}\max_j\|\mathbf{v}_j\|_2$. For simplicity, let $\mathbf{H}_j = \mathbf{D}_j\mathbf{X}\mathbf{X}^T\mathbf{D}_j/n$, so that $\mathbf{v}_j^T\mathbf{v}_j = \boldsymbol{\varepsilon}^T\mathbf{H}_j\boldsymbol{\varepsilon}$. By the definition of $\mathbf{D}_j$ and the fact that $\|\mathbf{x}_i\|_2 \le 1$ for all i, we have

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

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

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

By the union bound, $P\big(\max_j\|\mathbf{v}_j\|_2^2 \ge 2\sigma_\varepsilon^2 + 4\sigma_\varepsilon^2\sqrt{t} + 4\sigma_\varepsilon^2 t\big) \le 2pe^{-t}$. Recall that $p \le 2d(en/d)^d = 2\overline{p}$. Choosing $t = 5\log(4\overline{p})$ and noting that $\log(4\overline{p}) > 1$ yields

$$\max_{1\le j\le 2p}\|\mathbf{v}_j\|_2^2 < 2\sigma_\varepsilon^2 + 4\sigma_\varepsilon^2\sqrt{5\log(4\overline{p})} + 20\sigma_\varepsilon^2\log(4\overline{p}) < 49\sigma_\varepsilon^2\log(4\overline{p}) \tag{31}$$

with probability at least $1 - O(\overline{p}^{-4})$. Thus, for $\lambda \ge 2n^{-1/2}\max_j\|\mathbf{v}_j\|_2$ to hold with at least the same probability, it suffices to set $\lambda = 14\sigma_\varepsilon\sqrt{n^{-1}\log(4\overline{p})}$. Substituting this λ into (30) gives

$$\|g(\cdot;\widehat{\boldsymbol{\theta}}) - f^*(\cdot)\|_n^2 \le 2C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d} + 168\big(\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2\big)\sqrt{\frac{\log(4\overline{p})}{n}},$$

where we have used the inequality $2\sigma_\varepsilon\|f^*\|_{\mathcal{S}} \le \sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2$. To complete the proof, note that when $\sqrt{en} > d$,

$$-2\log\overline{p} \le 2d\log(d) - 2d\log(en) \le -d\log(en). \tag{32}$$

The desired result follows if we choose $C_2 = 2d$ so that $\overline{p}^{-4} \le (en)^{-C_2}$. □

To prove (14) in Theorem 3, we need the following maximal inequality whose proof can be found in the Supplementary Material G.1.

LEMMA 2. *Suppose that $\mathbf{x}_1,\ldots,\mathbf{x}_n$ are independently sampled from the distribution μ. Let $\mathcal{F}^*(m,1) = \{f - f^*\colon f \in \mathcal{F}(m,1), f^* \text{ is fixed with } \|f^*\|_{\mathcal{S}} \le 1\}$, 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 $C_{\mathcal{F}} > 0$ only depending on d. Furthermore, if $n \ge C_{\mathcal{F}}^2$, then*

$$P\left(Z_n \ge \frac{C_{\mathcal{F}}}{\sqrt{n}} + t\right) \le \exp\left\{-\frac{n}{32}\min\left(\frac{t^2}{12e}, t\right)\right\}. \tag{33}$$

PROOF OF (14) IN THEOREM 3. Let $\widehat{f}(\cdot) = g(\cdot;\widehat{\boldsymbol{\theta}})$ and $\rho_n = 6\sigma_\varepsilon\sqrt{n^{-1}\log(4\overline{p})}$. Define the event $E_0 = \{\rho_n \geq n^{-1/2}\max_j \|\mathbf{v}_j\|_2\}$. It follows from (31) that $P(E_0) \geq 1 - O(\overline{p}^{-4})$.

On the event E_0, by (30) we have

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

Since $\nu(\boldsymbol{\theta}^*) = \|\mathbf{B}^*\|_{2,1} \leq 6\|f^*\|_{\mathcal{S}}$ and $\|\widehat{\mathbf{B}}\|_{2,1} = \nu(\widehat{\boldsymbol{\theta}})$, by choosing $\lambda = 2\rho_n$ we further obtain

$$\lambda\nu(\widehat{\boldsymbol{\theta}}) \leq 3\lambda\nu(\boldsymbol{\theta}^*) + 2C_1\|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d}. \tag{34}$$

If $m \geq C_2\big(n/\log\overline{p}\big)^{d/(2(d+3))}$ for some constant $C_2 > 0$ such that

$$C_1\big(\|f^*\|_{\mathcal{S}}/\sigma_\varepsilon\big)m^{-(d+3)/d}\sqrt{n/\{49\log(4\overline{p})\}} < 1, \tag{35}$$

then $\nu(\widehat{\boldsymbol{\theta}}) \leq 20\|f^*\|_{\mathcal{S}}$ also holds with probability $1 - O(\overline{p}^{-4})$. Let $\widehat{\Delta} = \widehat{f} - f^*$ and let $\|\widehat{\Delta}\|_{\mathcal{S}}$ be the $\mathcal{S}$-norm of $\widehat{\Delta}$. By the definition of $\mathcal{S}$-norm and the triangle inequality, $\|\widehat{\Delta}\|_{\mathcal{S}} \leq \|\widehat{f}\|_{\mathcal{S}} + \|f^*\|_{\mathcal{S}} \leq \nu(\widehat{\boldsymbol{\theta}}) + \|f^*\|_{\mathcal{S}}$. Therefore, with the same probability

$$\frac{\widehat{\Delta}}{21\|f^*\|_{\mathcal{S}}} = \frac{\widehat{f}}{21\|f^*\|_{\mathcal{S}}} - \frac{f^*}{21\|f^*\|_{\mathcal{S}}} \in \mathcal{F}^*(m,1). \tag{36}$$

By Lemma 2, we have, with probability at least $1 - \exp\big\{-nt^2/384\big\}$,

$$\|\widehat{\Delta}\|_2^2 \leq \|\widehat{\Delta}\|_n^2 + \frac{441}{\sqrt{n}}C_{\mathcal{F}}\|f^*\|_{\mathcal{S}}^2 + 441\|f^*\|_{\mathcal{S}}^2 t \tag{37}$$

for any $t \leq 12e$. To bound $\|\widehat{\Delta}\|_n^2$, it follows from Theorem 2 that with probability at least $1 - O(\overline{p}^{-4})$,

$$\|\widehat{\Delta}\|_n^2 = \|\widehat{f} - f^*\|_n^2 \leq C_2\bigg\{\big(\sigma_\varepsilon^2 + \|f^*\|_{\mathcal{S}}^2\big)\sqrt{\frac{\log\overline{p}}{n}} + \|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d}\bigg\} \tag{38}$$

for some constant $C_2 > 0$.

Combining (37)–(38) and taking $t = \|f^*\|_{\mathcal{S}}^{-1}\sigma_\varepsilon\sqrt{n^{-1}\log\overline{p}} \leq 12e$ yields

$$\|\widehat{f} - f^*\|_2^2 \leq C_3\bigg\{\big(\sigma^2 + \|f^*\|_{\mathcal{S}}^2\big)\sqrt{\frac{\log\overline{p}}{n}} + \|f^*\|_{\mathcal{S}}^2 m^{-(d+3)/d}\bigg\}$$

with probability at least $1 - O(\overline{p}^{-4} + \overline{p}^{-\tau_0})$ for some constant $C_3 > 0$ and $\tau_0 = \sigma_\varepsilon^2/(384\|f^*\|_{\mathcal{S}}^2)$. The result follows from (32) that $O(\overline{p}^{-4} + \overline{p}^{-\tau_0}) \leq O(n^{-C_4})$ for some constant $C_4 > 0$. □

SUPPLEMENTARY MATERIAL

Supplement to "Nonasymptotic theory for two-layer neural networks: Beyond the bias–variance trade-off" (; .pdf). The supplement contains the remaining proofs and technical details.

Funding. This research was supported by Beijing Natural Science Foundation grant Z190001 and National Natural Science Foundation of China grants 12171012 and 12292981.

REFERENCES

ARORA, S., COHEN, N. and HAZAN, E. (2018). On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning* 244–253.

ARORA, S., DU, S. S., HU, W., LI, Z. and WANG, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning* 322–332.

BACH, F. (2017). Breaking the curse of dimensionality with convex neural networks. *J. Mach. Learn. Res.* **18** 1–53.

BARRON, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory* **39** 930–945.

BARRON, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Mach. Learn.* **14** 115–133.

BARTLETT, P. L., MENDELSON, S. and NEEMAN, J. (2012). ℓ_1-regularized linear regression: Persistence and oracle inequalities. *Probab. Theory Related Fields* **154** 193–224.

BARTLETT, P. L., LONG, P. M., LUGOSI, G. and TSIGLER, A. (2020). Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. USA* **117** 30063–30070.

BELKIN, M., HSU, D. and XU, J. (2020). Two models of double descent for weak features. *SIAM J. Math. Data Sci.* **2** 1167–1180.

BELKIN, M., HSU, D., MA, S. and MANDAL, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci. USA* **116** 15849–15854.

BELKIN, M., HSU, D., MA, S. and MANDAL, S. (2020). Reply to Loog et al.: Looking beyond the peaking phenomenon. *Proc. Natl. Acad. Sci. USA* **117** 10627.

BISHOP, C. M. (1995). Training with noise is equivalent to Tikhonov regularization. *Neural Comput.* **7** 108–116.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Berlin.

CAPONNETTO, A. and DE VITO, E. (2007). Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **7** 331–368.

CHINOT, G., LÖFFLER, M. and VAN DE GEER, S. (2022). On the robustness of minimum norm interpolators and regularized empirical risk minimizers. *Ann. Statist.* **50** 2306–2333.

CHIZAT, L., OYALLON, E. and BACH, F. (2019). On lazy training in differentiable programming. *Advances in Neural Information Processing Systems* **32** 2937–2947.

COVER, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comput.* **EC-14** 326–334.

DENG, Z., KAMMOUN, A. and THRAMPOULIDIS, C. (2022). A model of double descent for high-dimensional binary linear classification. *Inf. Inference* **11** 435–495.

DERUMIGNY, A. and SCHMIDT-HIEBER, J. (2023). On lower bounds for the bias-variance trade-off. *Ann. Statist.*

DOU, X. and LIANG, T. (2021). Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. *J. Amer. Statist. Assoc.* **116** 1507–1520.

E, W., MA, C. and WU, L. (2019). A priori estimates of the population risk for two-layer neural networks. *Commun. Math. Sci.* **17** 1407–1425.

E, W., MA, C. and WU, L. (2020). A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Sci. China Math.* **63** 1235–1258.

EFRON, B. (2020). Prediction, estimation, and attribution. *J. Amer. Statist. Assoc.* **115** 636–655.

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

ESTEVA, A., KUPREL, B., NOVOA, R. A., KO, J., SWETTER, S. M., BLAU, H. M. and THRUN, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542** 115–118.

FARRELL, M. H., LIANG, T. and MISRA, S. (2021). Deep neural networks for estimation and inference. *Econometrica* **89** 181–213.

GEMAN, S., BIENENSTOCK, E. and DOURSAT, R. (1992). Neural networks and the bias/variance dilemma. *Neural Comput.* **4** 1–58.

GHORBANI, B., MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2021). Linearized two-layers neural networks in high dimension. *Ann. Statist.* **49** 1029–1054.

GOLOWICH, N., RAKHLIN, A. and SHAMIR, O. (2020). Size-independent sample complexity of neural networks. *Inf. Inference* **9** 473–504.

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA.

HASTIE, T., MONTANARI, A., ROSSET, S. and TIBSHIRANI, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Statist.* **50** 949–986.

HAYAKAWA, S. and SUZUKI, T. (2020). On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Netw.* **123** 343–361.

HE, K., ZHANG, X., REN, S. and SUN, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778.

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

HSU, D., KAKADE, S., ZHANG, T. et al. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability* **17**.

JACOT, A., GABRIEL, F. and HONGLER, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems* **31** 8580–8589.

JARRETT, K., KAVUKCUOGLU, K., RANZATO, M. and LECUN, Y. (2009). What is the best multi-stage architecture for object recognition? In *IEEE 12th International Conference on Computer Vision* 2146–2153.

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

JONES, L. K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Statist.* **20** 608–613.

KLUSOWSKI, J. M. and BARRON, A. R. (2018). Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Trans. Inf. Theory* **64** 7649–7656.

KOHLER, M. and LANGER, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *Ann. Statist.* **49** 2231–2249.

KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* **25** 1097–1105.

KROGH, A. and HERTZ, J. A. (1991). A simple weight decay can improve generalization. *Advances in Neural Information Processing Systems* **4** 950–957.

LI, X. and MENG, X.-L. (2021). A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *J. Amer. Statist. Assoc.* **116** 353–367.

LIANG, T., RAKHLIN, A. and ZHAI, X. (2020). On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Proceedings of the 33rd Conference on Learning Theory* 2683–2711.

LIANG, T. and SUR, P. (2022). A precise high-dimensional asymptotic theory for boosting and minimum-ℓ_1-norm interpolated classifiers. *Ann. Statist.* **50** 1669–1695.

MAKOVOZ, Y. (1996). Random approximants and neural networks. *J. Approx. Theory* **85** 98–109.

MATOUŠEK, J. (1996). Improved upper bounds for approximation by zonotopes. *Acta Math.* **177** 55–73.

MEI, S., MONTANARI, A. and NGUYEN, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA* **115** E7665–E7671.

MEI, S. and MONTANARI, A. (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Comm. Pure Appl. Math.* **75** 667–766.

MONTANARI, A. and ZHONG, Y. (2022). The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *Ann. Statist.* **50** 2816–2847.

MUTHUKUMAR, V., VODRAHALLI, K., SUBRAMANIAN, V. and SAHAI, A. (2020). Harmless interpolation of moisy data in regression. *IEEE J. Sel. Areas Inform. Theory* **1** 67-83.

NAKKIRAN, P., VENKAT, P., KAKADE, S. and MA, T. (2021). Optimal regularization can mitigate double descent. In *International Conference on Learning Representations*.

NEYSHABUR, B., TOMIOKA, R. and SREBRO, N. (2015a). Norm-based capacity control in neural networks. In *Proceedings of the 28th Conference on Learning Theory* 1376–1401.

NEYSHABUR, B., TOMIOKA, R. and SREBRO, N. (2015b). In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations*.

NEYSHABUR, B., LI, Z., BHOJANAPALLI, S., LECUN, Y. and SREBRO, N. (2019). The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*.

ONGIE, G., WILLETT, R., SOUDRY, D. and SREBRO, N. (2020). A function space view of bounded norm infinite width ReLU nets: The multivariate case. In *International Conference on Learning Representations*.

PARHI, R. and NOWAK, R. D. (2021). Banach space representer theorems for neural networks and ridge splines. *J. Mach. Learn. Res.* **22** 1–40.

PARHI, R. and NOWAK, R. D. (2022). Near-minimax optimal estimation with shallow ReLU neural networks. *IEEE Trans. Inf. Theory*.

PILANCI, M. and ERGEN, T. (2020). 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* 7695–7705.

RAHIMI, A. and RECHT, B. (2007). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems* **20** 1177–1184.

ROTSKOFF, G. M. and VANDEN-EIJNDEN, E. (2022). Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Comm. Pure Appl. Math.* **75** 1889–1935.

SCHMIDT-HIEBER, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.* **48** 1875–1897.

SCHRITTWIESER, J., ANTONOGLOU, I., HUBERT, T., SIMONYAN, K., SIFRE, L., SCHMITT, S., GUEZ, A., LOCKHART, E., HASSABIS, D., GRAEPEL, T., LILLICRAP, T. and SILVER, D. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* **588** 604–609.

SIEGEL, J. W. and XU, J. (2020). Approximation rates for neural networks with general activation functions. *Neural Netw.* **128** 313–321.

SIEGEL, J. W. and XU, J. (2022). Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *Found. Comput. Math.*

SIRIGNANO, J. and SPILIOPOULOS, K. (2020). Mean field analysis of neural networks: A law of large numbers. *SIAM J. Appl. Math.* **80** 725–752.

SJÖBERG, J. and LJUNG, L. (1995). Overtraining, regularization and searching for a minimum, with application to neural networks. *Internat. J. Control* **62** 1391–1407.

SOLTANOLKOTABI, M., JAVANMARD, A. and LEE, J. D. (2019). Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Trans. Inf. Theory* **65** 742–769.

SREBRO, N., RENNIE, J. D. M. and JAAKKOLA, T. S. (2004). Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems* **17** 1329–1336.

SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I. and SALAKHUTDINOV, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15** 1929–1958.

SUTSKEVER, I., VINYALS, O. and LE, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems* **27** 3104–3112.

WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Univ. Press, Cambridge.

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

YU, J., WANG, Z., VASUDEVAN, V., YEUNG, L., SEYEDHOSSEINI, M. and WU, Y. (2022). CoCa: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.*

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67.

ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* **64** 107–115.