The Annals of Statistics
2021, Vol. 49, No. 2, 1029–1054
https://doi.org/10.1214/20-AOS1990

LINEARIZED TWO-LAYERS NEURAL NETWORKS IN HIGH DIMENSION

BY BEHROOZ GHORBANI[1], SONG MEI[2], THEODOR MISIAKIEWICZ[3] AND ANDREA MONTANARI[4]

[1]*Department of Electrical Engineering, Stanford University, ghorbani@stanford.edu*
[2]*Institute for Computational and Mathematical Engineering, Stanford University, songmei@stanford.edu*
[3]*Department of Statistics, Stanford University, misiakie@stanford.edu*
[4]*Department of Electrical Engineering and Department of Statistics, Stanford University, montanari@stanford.edu*

We consider the problem of learning an unknown function $f_\star$ on the d-dimensional sphere with respect to the square loss, given i.i.d. samples $\{(y_i, \boldsymbol{x}_i)\}_{i\le n}$ where $\boldsymbol{x}_i$ is a feature vector uniformly distributed on the sphere and $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$. We study two popular classes of models that can be regarded as linearizations of two-layers neural networks around a random initialization: the random features model of Rahimi–Recht (RF); the neural tangent model of Jacot–Gabriel–Hongler (NT). Both these models can also be regarded as randomized approximations of kernel ridge regression (with respect to different kernels), and enjoy universal approximation properties when the number of neurons N diverges, for a fixed dimension d.

We consider two specific regimes: the infinite-sample finite-width regime, in which $n = \infty$ while d and N are large but finite, and the infinite-width finite-sample regime in which $N = \infty$ while d and n are large but finite. In the first regime, we prove that if $d^{\ell+\delta} \le N \le d^{\ell+1-\delta}$ for small $\delta > 0$, then RF effectively fits a degree-ℓ polynomial in the raw features, and NT fits a degree-$(\ell+1)$ polynomial. In the second regime, both RF and NT reduce to kernel methods with rotationally invariant kernels. We prove that, if the sample size satisfies $d^{\ell+\delta} \le n \le d^{\ell+1-\delta}$, then kernel methods can fit at most a degree-ℓ polynomial in the raw features. This lower bound is achieved by kernel ridge regression, and near-optimal prediction error is achieved for vanishing ridge regularization.

1. Introduction and main results. In the canonical statistical learning problem, we are given independent and identically distributed (i.i.d.) pairs $(y_i, \boldsymbol{x}_i)$, $1 \le i \le n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is a feature vector and $y_i \in \mathbb{R}$ is a label or response variable. We would like to construct a function f which allows us to predict future responses. Throughout this paper, we will measure the quality of a predictor f via its square prediction error (risk): $R(f) \equiv \mathbb{E}\{(y - f(\boldsymbol{x}))^2\}$.

1.1. *Background.* For a number of important applications, state-of-the-art performances are obtained by representing the function f as a multi-layers neural network. The simplest model in this class is given by two-layers networks (NN):

$$\mathcal{F}_{\mathsf{NN}} \equiv \left\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \sigma\big(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle\big) : a_i \in \mathbb{R}, \boldsymbol{w}_i \in \mathbb{R}^d \ \forall i \le N \right\}. \tag{NN}$$

Here, N is the number of neurons and $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function.

Two-layers neural networks have been extensively studied in the nineties, with a focus on two goals: (i) Establishing approximation guarantees over classical function spaces; (ii)

Received July 2019; revised June 2020.
MSC2020 subject classifications. Primary 62G08; secondary 62J07.
Key words and phrases. Two-layers neural networks, random features, neural tangent kernel, approximation bounds, Kernel ridge regression.

Controlling the generalization error via Rademacher complexity arguments. We refer to [3, 49] for surveys of these results.

Computational aspects were notably underrepresented within these early theoretical contributions. On the contrary, it is nowadays increasingly clear that computational and statistical aspects cannot be separated in the analysis of neural networks (see, e.g., [16, 44, 54]). Indeed, the optimization algorithm does not simply compute the unique minimizer of a regularized empirical risk: it instead selects one among many possible near-minimizers, whose generalization properties can vary significantly. Therefore, the specific optimization algorithm is an integral part of the definition of the regularization method.

A concrete scenario in which this interplay can be understood precisely is the so-called "neural tangent kernel" regime. First, explicitly described in [35], this regime has attracted considerable amount of work. The basic idea is that, for highly overparametrized networks, the network weights barely change from their random initialization. We can therefore replace the nonlinear function class $\mathcal{F}_{\mathsf{NN}}$ by its first-order Taylor expansion around this initialization.

Denoting by $(a_{0,i}, \boldsymbol{w}_{0,i})_{i\le N}$ the weights at initialization, a first-order Taylor expansion yields

$$f_{\mathsf{NN}}(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \sigma\big(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle\big) \approx f_{\mathsf{NN},0}(\boldsymbol{x}) + \sum_{i=1}^{N} (a_i - a_{0,i}) \sigma\big(\langle \boldsymbol{w}_{0,i}, \boldsymbol{x}\rangle\big) + \sum_{i=1}^{N} a_{0,i} \langle \boldsymbol{w}_i - \boldsymbol{w}_{0,i}, \boldsymbol{x}\rangle \sigma'\big(\langle \boldsymbol{w}_{0,i}, \boldsymbol{x}\rangle\big),$$

where $f_{\mathsf{NN},0}$ is the neural network at initialization. In other words, $f_{\mathsf{NN}} - f_{\mathsf{NN},0}$ is a function in the direct sum $\mathcal{F}_{\mathsf{NT}}(\boldsymbol{W}) \oplus \mathcal{F}_{\mathsf{RF}}(\boldsymbol{W})$, where we defined

$$\mathcal{F}_{\mathsf{RF}}(\boldsymbol{W}) \equiv \left\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \sigma\big(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle\big) : a_i \in \mathbb{R}\ \forall i \le N \right\}, \tag{RF}$$

$$\mathcal{F}_{\mathsf{NT}}(\boldsymbol{W}) \equiv \left\{ f(\boldsymbol{x}) = \sum_{i=1}^{N} \langle \boldsymbol{a}_i, \boldsymbol{x}\rangle \sigma'\big(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle\big) : \boldsymbol{a}_i \in \mathbb{R}^d\ \forall i \le N \right\}. \tag{NT}$$

Here, $\boldsymbol{W} \in \mathbb{R}^{N\times d}$ is a matrix whose ith row is the vector $\boldsymbol{w}_i$, and σ' is the derivative of the activation function with respect to its argument (if $\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle$ has a density, σ only needs to be weakly differentiable).

We will refer to $\mathcal{F}_{\mathsf{RF}}(\boldsymbol{W})$ as the "random features" (RF) model: it amounts to fixing the bottom-layer weights, and only optimizing the top-layer weights. Equivalently, $\mathcal{F}_{\mathsf{RF}}(\boldsymbol{W})$ corresponds to the first-order Taylor expansion of f_{NN} with respect to the top-layer weights $(a_i)_{i\le N}$. This model can be traced back to the work of Neal [47], and was successfully developed by Rahimi and Recht [50] as a randomized approximation to kernel methods.

The second function class $\mathcal{F}_{\mathsf{NT}}(\boldsymbol{W})$ corresponds to the first-order Taylor expansion of f_{NN} with respect to the bottom-layer weights $(\boldsymbol{w}_i)_{i\le N}$ [35]. We will refer to $\mathcal{F}_{\mathsf{NT}}(\boldsymbol{W})$ as the neural tangent class.[1]

A sequence of recent papers proves that, in a certain overparametrized regime, gradient descent (GD) applied to the nonlinear neural network class $\mathcal{F}_{\mathsf{NN}}$ effectively converges to a model in $\mathcal{F}_{\mathsf{NT}}(\boldsymbol{W}) \oplus \mathcal{F}_{\mathsf{RF}}(\boldsymbol{W})$. Namely, if the number of neurons N is larger than a threshold $N_0(n, d)$, and training is initialized with $f_0(\boldsymbol{x}) = N^{-1/2} \sum_{i=1}^{N} a_{0,i} \sigma(\langle \boldsymbol{w}_{0,i}, \boldsymbol{x}\rangle)$

[1] Often the term "neural tangent" is reserved for the direct sum $\mathcal{F}_{\mathsf{NT}}(\boldsymbol{W}) \oplus \mathcal{F}_{\mathsf{RF}}(\boldsymbol{W})$. We find it more convenient to give distinct names to each of the two terms, especially since $\mathcal{F}_{\mathsf{RF}}(\boldsymbol{W})$ has much smaller dimension than $\mathcal{F}_{\mathsf{NT}}(\boldsymbol{W})$ for large d.

where $\{(a_{0,i}, \boldsymbol{w}_{0,i})\}_{i\le N} \sim_{\text{i.i.d.}} \mathsf{N}(0,1) \otimes \mathsf{N}(0, \mathbf{I}_d/d)$, then gradient descent converges exponentially fast to weights $\{(a_i, \boldsymbol{w}_i)\}_{i\le N}$ such that $f - f_0$ is well approximated by a function in $\mathcal{F}_{\mathsf{NT}}(\boldsymbol{W}) \oplus \mathcal{F}_{\mathsf{RF}}(\boldsymbol{W})$. The specific value of the threshold $N_0(n,d)$ for the onset of this NT regime has been steadily pushed down over the last year [2, 4, 22, 23, 59].

Does the NT regime explain the power of multilayers neural networks, when trained by gradient descent methods? From an empirical point of view, the evidence is not univocal [17, 27, 36]. From a theoretical point of view, while the expressivity of neural networks is superior to the one of NT models, this hypothesis is not easy to dismiss for at least two reasons. First, neural networks learned by gradient descent algorithms form a significantly smaller class than general networks. Second, the answer depends on the data distribution, the target function $f_\star$ and the sample size.

In order to clarify this question, we explore the behavior of RF and NT models in the high-dimensional setting. More precisely, we consider two specific asymptotic regimes:

(i) The infinite-sample finite-width regime in which $n = \infty$, and N, d diverge while being polynomially related. In this case, the prediction error reduces to the approximation error $\inf_{f\in\mathcal{F}_{\mathsf{M}}} \mathbb{E}\{[f_\star(\boldsymbol{x}) - f(\boldsymbol{x})]^2\}$, for either model $\mathsf{M} \in \{\mathsf{NT}, \mathsf{RF}\}$.

(ii) The infinite-width finite-sample regime in which $N = \infty$ and n, d diverge while being polynomially related. In this case (and under a suitable bound on the ℓ_2 norm of the coefficients) both classes $\mathcal{F}_{\mathsf{RF}}$, $\mathcal{F}_{\mathsf{NT}}$ reduce to certain reproducing kernel Hilbert spaces (RKHS).

In both cases, we obtain sharp results, up to errors vanishing as $d \to \infty$. Crucially, our results hold *pointwise*, that is, they provide a characterization of approximation and generalization error which hold *for a given function* $f_\star$.

In summary, a large number of recent papers argue that large neural networks, *when trained using practical algorithms*, such as stochastic gradient descent, behave as linearized neural networks. Our findings suggest that the performance gap between neural network and its linearization is large, and hence the linear theory does not fully capture the behavior of neural networks, even if we limit ourselves to those trained via gradient-based algorithms.

1.2. *A parenthesis.* The approximation properties of neural networks have been studied for over three decades [8, 19, 20, 31, 34, 39, 45, 46, 48, 49]. It is useful to discuss the relation between the questions outlined above and existing literature.

A number of results are available on the approximation of functions in certain smoothness classes by two-layers neural networks. In particular, [8] controls smoothness by the average frequency content in the Fourier transform (the "Barron norm"), while [39, 40, 45, 48] use classical Sobolev norms. For instance, [40] proves that N-neurons NN approximate functions in the Sobolev ball $W_2^{r,d}$ with worst case error

$$C_1(r,d,\delta)N^{-r/d-\delta} \le \sup_{\|f\|_{W_2^{r,d}}\le 1} \inf_{\hat f\in\mathcal{F}_{\mathsf{NN}}} \|f - \hat f\|_{L^2} \le C_2(r,d,\delta)N^{-r/d+\delta}. \tag{1}$$

for any $\delta > 0$ and for some functions C_1, C_2 that are independent of N. (Similar results are found in [48].) These results cannot be used for our purposes.

First of all, we are interested in the NT class which is potentially much less powerful than NN.

Second, even if bounds of the type (1) were available for NT, it would be hard to use them to prove separation results between NN and NT. Since the lower bound in (1) is for the worst case function, in order to prove a separation result, we would have to prove that neural networks trained by gradient descent have good approximation properties, uniformly over Sobolev balls. This objective is currently out of reach. Our pointwise approximation results make it much easier to prove separation statements.

Third, bounds of the type (1) have weak implications when both d and N are large, say $d = 100$, $N = 10^6$. We will instead prove sharp asymptotic results that are valid in this regime. As illustrated in the next section, our analysis captures the actual behavior in a quantitative manner, already when $d \geq 100$.

Quantitative results in the high-dimensional regime have been proved only recently. In particular, Bach [6] established quantitative upper and lower bounds for the approximation error in the RF model. However, these results do not have direct implications on the NT model which is our main interest here. Further, lower bounds in [6] are, as before, worst case over a certain RKHS. (See also [1, 5, 52] for related work.)

Similar considerations apply to the generalization error of kernel methods. While this is a classical topic [14, 18, 37, 52], earlier work proves minimax upper and lower bounds. Establishing pointwise lower bounds is instead important in order to understand precisely the separation between neural networks and their linearized counterparts. We refer to Section 4 for further discussion of related work.

1.3. *A numerical experiment.* In order to illustrate the approximation behavior of RF and NT models, we present a simple simulation study. We consider feature vectors normalized so that $\|\boldsymbol{x}_i\|_2^2 = d$, and otherwise uniformly random, and responses $y_i = f_\star(\boldsymbol{x}_i)$, for a certain function $f_\star$. Indeed, this will be the setting throughout the paper: $\boldsymbol{x}_i \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ (where $\mathbb{S}^{d-1}(r)$ denotes the sphere with radius r in d dimensions) and $f_\star : \mathbb{S}^{d-1}(\sqrt{d}) \to \mathbb{R}$. We draw random weights $(\boldsymbol{w}_i)_{i\leq N} \sim_{\text{i.i.d.}} \mathsf{Unif}(\mathbb{S}^{d-1}(1))$. We use n samples to fit a model in $\mathcal{F}_{\mathsf{RF}}(\boldsymbol{W})$ or $\mathcal{F}_{\mathsf{NT}}(\boldsymbol{W})$. We learn the model parameters using least squares. If the model is overparametrized, we select the minimum ℓ_2-norm solution. (We refer to Appendix F in the Supplementary Material [30] for simulations using ridge regression instead.) We estimate the risk (test error) using $n_{\text{test}} = 1500$ fresh samples, and normalize it by the risk of the trivial model $R_0 = \mathbb{E}\{f_\star(\boldsymbol{x})^2\}$.

Figures 1, 2, 3 report the results of such a simulation using RF—for Figure 1—and NT—for Figures 2 and 3. We use shifted ReLU activations $\sigma(u) = \max(u - u_0, 0)$, $u_0 = 0.5$. The choice of $u_0 = 0.5$ is not essential: (Lebesgue-)almost every $u_0 \neq 0$ has similar behavior. In contrast, the case $u_0 = 0$ is degenerate because $\max(u, 0)$ is equal to a linear function plus an even function.[2]

The target functions $f_\star$ in these examples are quite simple. 1 and 2 use a quadratic function $f_{\star,2}(\boldsymbol{x}) = \sum_{i\leq\lfloor d/2\rfloor} x_i^2 - \sum_{i>\lfloor d/2\rfloor} x_i^2$. In Figure 3, the target function is a third-order polynomial $f_{\star,3}(\boldsymbol{x}) = \sum_{i=1}^d (x_i^3 - 3x_i)$.

The results are somewhat disappointing: in two cases (first and third figures) RF and NT models do not beat the trivial predictor. In one case (the second one), the NT model surpasses the trivial baseline, and it appears to decrease to 0 as the number of samples n increase. We also note that the risk shows a cusp when $n \approx p$, with p being the number of parameters ($p = N$ for RF, and $p = Nd$ for NT). This phenomenon is related to overparametrization, and will not be discussed further in this paper (see [9, 10, 33, 43] for relevant work). We will instead focus on the population behavior $n \to \infty$.

In other words, the RF model does not appear to be able to learn a simple quadratic function, and the NT model does not appear to be able to learn a third order polynomial. Our main theorems (presented in the next sections) capture this behavior in a precise manner. In particular,

[2]Note that ReLU and shifted ReLU are not equivalent in the present setting because we are not allowing for biases, namely a neuron maps $\boldsymbol{x} \mapsto \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle)$ rather than $\boldsymbol{x} \mapsto \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle + u_i)$. A slightly more general setting would introduce random biases u_i in the RF and NT models. However, this makes the proof more cumbersome, without being substantially different from the constant bias u_0.

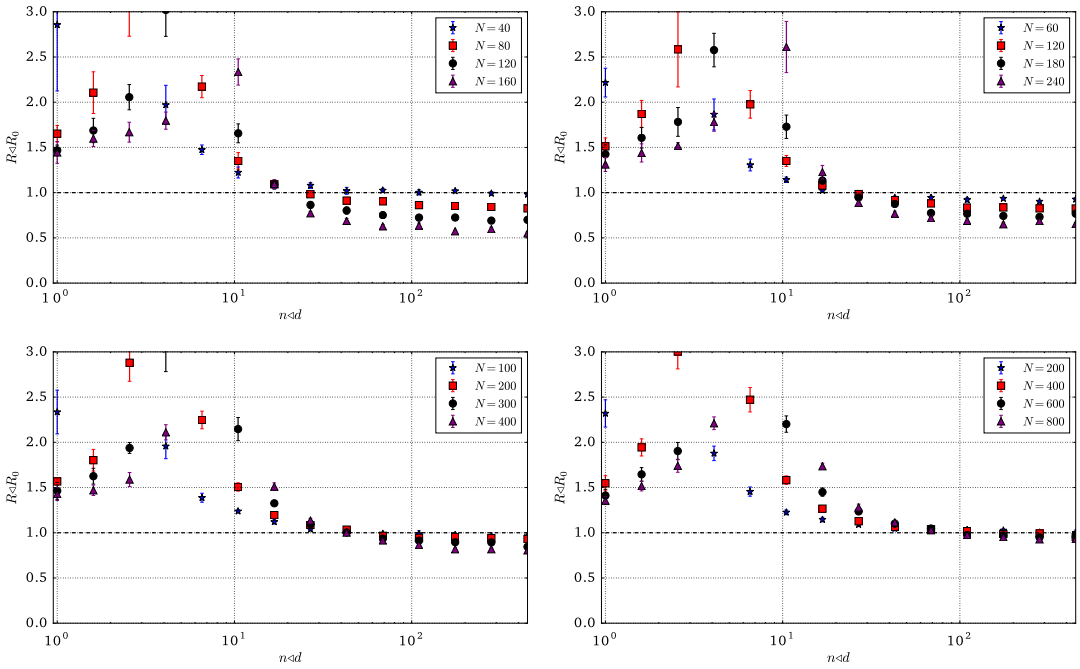

FIG. 1. *Risk of the random features model for learning a quadratic function* $f_{\star,2}$, *for* $d = 20$ (*top left*), $d = 30$ (*top right*), $d = 50$ (*bottom left*) *and* $d = 100$ (*bottom right*). *We use least square to estimate the model coefficients from* n *samples and report the test error over* $n_{\text{test}} = 1500$ *fresh samples. Data points correspond to averages over* 10 *independent repetitions, and the risk is normalized by the risk* R_0 *of the trivial* (*constant*) *predictor.*

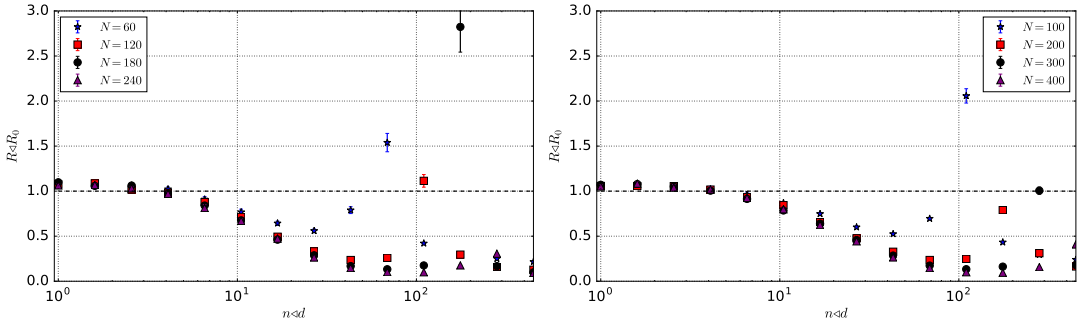

FIG. 2. *Risk* (*test error*) *of the neural tangent model in learning a quadratic function* $f_{\star,2}$, *for* $d = 30$ (*left frame*) *and* $d = 50$ (*right frame*). *The other settings are the same as in Figure* 1.

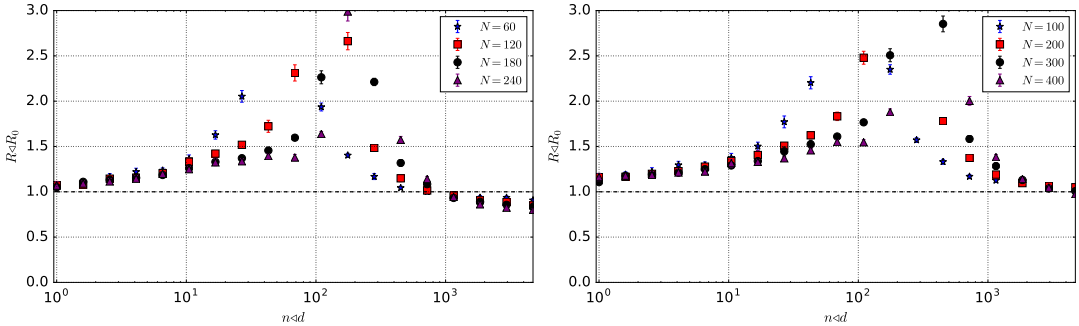

FIG. 3. *Risk* (*test error*) *of the neural tangent model in learning a third order polynomial* $f_{\star,3}$, *for* $d = 30$ (*left frame*) *and* $d = 50$ (*right frame*). *The other settings are the same as in Figures* 1 *and* 2.

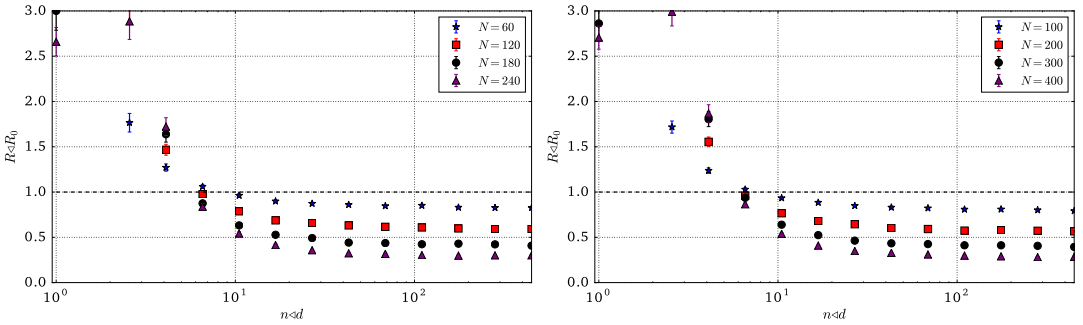

FIG. 4. *Upper bounds on the optimal risk of the neural network model* $\mathcal{F}_{\mathsf{NN}}$ *when used to learn the third-order polynomial* $f_{\star,3}$ (*same target function as in Figure* 3), *for* $d = 30$ (*left frame*) *and* $d = 50$ (*right frame*). *We use* n *train samples and report the test error over* $n_{\text{test}} = 1500$ *fresh samples. Data points correspond to averages over* 50 *independent repetitions, and the risk is normalized by the risk* R_0 *of the trivial* (*constant*) *predictor. Training uses oracle knowledge of the function* $f_{\star,3}$.

- We will prove that for $N = O_d(d^{2-\delta})$, RF does not outperform the trivial predictor on *any function* that has vanishing projection on linear functions. Similarly, NT does not outperform the trivial predictor on *any function* that has vanishing projection on linear and quadratic functions.
- In contrast, there exists neural networks in $\mathcal{F}_{\mathsf{NN}}$ with $N = O_d(d)$ neurons, and a small approximation error both for $f_{\star,2}$ and $f_{\star,3}$ (see, e.g., [6] or [44], Proposition 1).

These two points illustrate the gap in approximation power between NT (or RF) and NN.

We demonstrate the second point empirically in Figure 4 by choosing weight vectors $\boldsymbol{w}_i = s_i \boldsymbol{e}_{r(i)}$, where $r(i) \sim \mathsf{Unif}([n])$ are i.i.d. uniformly random indices, and the scaling factor is $s_i \sim \mathsf{N}(0, 1)$. Fixing these random bottom-layer weights, we fit the top-layer weights a_i by least squares. The risk achieved is an upper bound on the minimum risk in the NN model, namely $R_{\mathsf{NN}}(f_\star) \equiv \inf_{f \in \mathcal{F}_{\mathsf{NN}}} \mathbb{E}\{(f_\star(\boldsymbol{x}) - f(\boldsymbol{x}))^2\}$, and is significantly smaller than the baseline R_0. (The risk reported in Figure 4 can also be interpreted as a "random features" risk. However, the specific distribution of the vectors $\boldsymbol{w}_i$ is tailored to the function $f_\star$, and hence not achievable within the RF model.)

1.4. *Summary of main results. Approximation error of* RF *models.* If $d^{1+\delta} < N \le d^{2-\delta}$ for some $\delta > 0$, then the approximation error of RF is asymptotically equivalent to the approximation error of fitting a linear function in the raw covariates $\boldsymbol{x}$ (i.e., least squares with the model $f(\boldsymbol{x}) = b_0 + \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle$, $b_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^d$). More generally, if $d^{\ell+\delta} \le N \le d^{\ell+1-\delta}$, then RF is equivalent to fitting a linear function over all monomials of degree at most ℓ in $\boldsymbol{x}$.

The equivalence between RF regression and polynomial regression holds *pointwise* for target function $f_\star$.

Approximation error of NT *models.* If $d^{1+\delta} \le N \le d^{2-\delta}$, then the approximation error of NT is asymptotically equivalent to the approximation error of fitting a linear function over monomials of degree at most two in $\boldsymbol{x}$ (i.e., least squares with the model $f(\boldsymbol{x}) = b_0 + \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle + \langle \boldsymbol{x}, \boldsymbol{B}\boldsymbol{x} \rangle$, $b_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^d$, $\boldsymbol{B} \in \mathbb{R}^{d \times d}$). More generally, if $d^{\ell+\delta} \le N \le d^{\ell+1-\delta}$, then NT is equivalent to fitting a linear function over all monomials of degree at most $\ell + 1$ in $\boldsymbol{x}$.

Again, this result holds pointwise over the choice of $f_\star$.

Generalization error of kernel methods. We study the generalization error of kernel methods under the same data distribution described above, for any rotationally invariant kernel on the sphere $\mathbb{S}^{d-1}(\sqrt{d})$. We prove two results:

1. If the sample size is $n \le d^{\ell+1-\delta}$, then the generalization error of *any kernel method* is lower bounded by the approximation error of linear regression over monomials of degree at most ℓ in $\boldsymbol{x}$.

2. If the sample size satisfies $d^{\ell+\delta} \le n \le d^{\ell+1-\delta}$, then the generalization error of Kernel Ridge Regression (KRR) is given by the approximation error of linear regression over monomials of degree at most ℓ in $\boldsymbol{x}$.

It is worth emphasizing two aspects of this last result. The first one is its generality. The NT kernel associated to an infinitely wide *multilayers* fully connected neural network is always rotational invariant (assuming an i.i.d. Gaussian initialization of weights, which is common in practice). Therefore, in the NT regime, multilayers neural networks cannot outperform the trivial predictor on a target function $f_\star(\boldsymbol{x})$ that has vanishing projection onto degree-ℓ polynomials, unless the sample size satisfies $n \ge d^{\ell+1-\delta}$. (For instance, they cannot outperform the trivial predictor for $f_\star(\boldsymbol{x}) = x_1^3 - 3x_1$ unless $n \ge d^{3-\delta}$.)

The second aspect can be summarized as follows.

Optimality of near interpolators. For $d^{\ell+\delta} \le n \le d^{\ell+1-\delta}$, the ideal behavior of KRR is achieved for all regularization values $\lambda \le \lambda_*$, with λ_* depending on N, d and the activation function. In particular, it is achieved by "near interpolators" (corresponding to $\lambda \approx 0$), that is, functions $\hat{f}$ that have negligible training error.

The statistical properties of interpolators have been the object of a large number of recent papers [11, 12, 33, 37]. Our analysis provides a sharper optimality guarantee in certain cases.

2. Approximation error of linearized neural networks. In this section, we state formally our results about the approximation error of RF and NT models. We define the minimum population error for any of the models $\mathsf{M} \in \{\mathsf{RF}, \mathsf{NT}\}$ by

$$R_{\mathsf{M}}(f_\star, \boldsymbol{W}) = \inf_{f \in \mathcal{F}_{\mathsf{M}}(\boldsymbol{W})} \mathbb{E}\big[\big(f_\star(\boldsymbol{x}) - f(\boldsymbol{x})\big)^2\big], \quad \mathsf{M} \in \{\mathsf{RF}, \mathsf{NT}\}. \tag{2}$$

Notice that this is a random variable because of the random features encoded in the matrix $\boldsymbol{W} \in \mathbb{R}^{N\times d}$. Also, it depends implicitly on d, N, but we will make this dependence explicit only when necessary.

For $\ell \in \mathbb{N}$, we denote by $\mathsf{P}_{\le\ell} : L^2(\mathbb{S}^{d-1}(\sqrt{d})) \to L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ the orthogonal projector onto the subspace of polynomials of degree at most ℓ. (We also let $\mathsf{P}_{>\ell} = \mathbf{I} - \mathsf{P}_{\le\ell}$.) In other words, $\mathsf{P}_{\le\ell} f$ is the function obtained by linear regression of f onto monomials of degree at most ℓ. Throughout this paper, "with high probability" means 'with probability converging to one as $d, N \to \infty$'. The notation $s_d = \omega_d(t_d)$, $s_d = o_d(t_d)$, $s_d = O_d(t_d)$, $s_d = \Omega_d(t_d)$ mean, respectively, $\lim_{d\to\infty} |s_d/t_d| = \infty$, $\lim_{d\to\infty} |s_d/t_d| = 0$, $\limsup_{d\to\infty} |s_d/t_d| < \infty$, $\liminf_{d\to\infty} |s_d/t_d| > 0$. Given random variables X_d, and deterministic quantities t_d, we write $X_d = o_{d,\mathbb{P}}(t_d)$ (and so on) if the above holds in probability.

2.1. *Approximation error of random features models*.

ASSUMPTION 1 (Assumptions for the RF model at level $\ell \in \mathbb{N}$). Let $\{\sigma_d\}_{d\ge1}$ be a sequence of functions $\sigma_d : \mathbb{R} \to \mathbb{R}$.

(a) $\sigma_d \in L^2([-\sqrt{d}, \sqrt{d}], \tau_{d-1}^1)$, where τ_{d-1}^1 is the distribution of $\langle \boldsymbol{x}, \boldsymbol{e}\rangle$ for $\boldsymbol{x} \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, and $\boldsymbol{e} = (1, 0, \dots, 0)^{\mathsf{T}} \in \mathbb{R}^d$.

(b) We have

$$\Big[d^\ell \cdot \min_{k\le\ell} \lambda_{d,k}(\sigma_d)^2\Big] / \big\|\sigma_d\big(\langle \boldsymbol{e}, \cdot\rangle\big)\big\|^2_{L^2(\mathbb{S}^{d-1}(\sqrt{d}))} = \Omega_d(1),$$

where $\lambda_{d,k}(\sigma_d) = \langle \sigma_d(\langle \boldsymbol{e}, \cdot\rangle), Q_k(\sqrt{d}\langle \boldsymbol{e}, \cdot\rangle)\rangle_{L^2(\mathbb{S}^{d-1}(\sqrt{d}))}$, and Q_k is the kth Gegenbauer polynomial (see Section 5).

THEOREM 1 (Risk of the RF model). *Let $\{f_d \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))\}_{d\ge 1}$ be a sequence of functions. Let $\boldsymbol{W} = (\boldsymbol{w}_i)_{i\in[N]}$ with $(\boldsymbol{w}_i)_{i\in[N]} \sim \mathsf{Unif}(\mathbb{S}^{d-1})$ independently. Then the following hold*:

(a) *Assume $N \le d^{\ell+1-\delta_d}$ for a fixed integer ℓ and any sequence δ_d such that $\delta_d^2 \log d \to \infty$ (in particular, $N \le d^{\ell+1-\delta}$ is sufficient for any fixed $\delta > 0$). Let $\{\sigma_d\}_{d\ge 1}$ satisfy Assumption 1(a). Then, for any $\varepsilon > 0$, the following holds with high probability*:

$$\begin{aligned} &\big|R_{\mathsf{RF}}(f_d, \boldsymbol{W}) - R_{\mathsf{RF}}(\mathsf{P}_{\le \ell} f_d, \boldsymbol{W}) - \|\mathsf{P}_{>\ell} f_d\|_{L^2}^2\big| \\ &\qquad \le \varepsilon \|f_d\|_{L^2} \|\mathsf{P}_{>\ell} f_d\|_{L^2}. \end{aligned} \tag{3}$$

(b) *Assume $N = \omega_d(d^\ell)$ for some integer ℓ, and $\{\sigma_d\}_{d\ge 1}$ satisfy Assumption 1(b) at level ℓ. Then for any $\varepsilon > 0$, the following holds with high probability*:

$$0 \le R_{\mathsf{RF}}(\mathsf{P}_{\le \ell} f_d, \boldsymbol{W}) \le \varepsilon \|\mathsf{P}_{\le \ell} f_d\|_{L^2}^2. \tag{4}$$

See Appendix A for the proof of lower bound, and Appendix B for the proof of upper bound in the Ssupplementary Material.

In words, equation (3) amounts to say that when $N = O_d(d^{\ell+1-\delta_d})$, the risk of the random feature model can be approximately decomposed in two parts, each nonnegative, and each with a simple interpretation:

$$R_{\mathsf{RF}}(f_d, \boldsymbol{W}) \approx R_{\mathsf{RF}}(\mathsf{P}_{\le \ell} f_d, \boldsymbol{W}) + \|\mathsf{P}_{>\ell} f_d\|_{L^2}^2. \tag{5}$$

The second contribution, $\|\mathsf{P}_{>\ell} f_d\|_{L^2}^2$ is simply the risk achieved by linear regression with respect to polynomials of degree at most ℓ. The first contribution $R_{\mathsf{RF}}(\mathsf{P}_{\le \ell} f_d, \boldsymbol{W})$ is the risk of the RFmodel when applied to the low-degree component of f_d. Equation (4) implies that when $N = \omega_d(d^\ell)$, the first contribution $R_{\mathsf{RF}}(\mathsf{P}_{\le \ell} f_d, \boldsymbol{W})$ vanishes asymptotically.

If both Assumptions 1(a) and 1(b) hold and $\omega_d(d^\ell) \le N \le O_d(d^{\ell+1-\delta})$ for some integer ℓ, we thus obtain

$$R_{\mathsf{RF}}(f_d, \boldsymbol{W}) = \|\mathsf{P}_{>\ell} f_d\|_{L^2}^2 + \|f_d\|_{L^2}^2 \cdot o_{d,\mathbb{P}}(1).$$

In particular, this shows that RFfits a linear function over polynomials of maximum degree ℓ.

REMARK 1. Note that Theorem 1(a) holds under very weak conditions on the activation function, which may depend on the dimension d. The condition $\sigma_d(\langle \boldsymbol{e}_1, \cdot\rangle) \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ can also be rewritten as $\sigma_d \in L^2(\mathbb{R}, \tau_{d-1}^1)$, where τ_{d-1}^1 is the one-dimensional projection of the uniform measure over $\mathbb{S}^{d-1}(\sqrt{d})$. In particular:

(i) τ_{d-1}^1 is supported on $[-\sqrt{d}, \sqrt{d}]$. It is therefore sufficient that $\sup_{|u|\le\sqrt{d}} |\sigma_d(u)| = C_1(d) < \infty$.

(ii) By an explicit calculation, the density is given by $\tau_{d-1}^1(\mathrm{d}u) = C_2(d) \times (1 - u^2/d)^{(d-3)/2}\,\mathrm{d}u$. Since this density is bounded, it is sufficient that σ_d is square integrable with respect to the Lebesgue measure on $[-\sqrt{d}, \sqrt{d}]$.

REMARK 2. If the activation σ is independent of d, Assumption 1(b) is satisfied as long as $\mu_k(\sigma) \ne 0$ for $k = 0, \dots, \ell$, where $\mu_k(\sigma)$ is the kth Hermite coefficient of σ (see Section 5 for definitions).

REMARK 3. The conclusion of Theorem 1(a) can be established[3] by a somewhat simpler proof if the activation function σ is independent of d and satisfies the following regularity conditions: (i) $\sigma(u)^2 \le c_0 \exp(c_1 u^2/2)$ for some $c_1 < 1$; (ii) σ is not a polynomial of degree smaller than $2\ell + 3$. Under these conditions, the conclusion holds for $N = o_d(d^{\ell+1})$.

[3]The first version of this manuscript, posted on arXiv, assumed such conditions.

Note that Assumption 1(b) requires in particular that σ is not a polynomial of degree strictly smaller than ℓ. This is easily seen to be a necessary condition, since any linear combination of polynomials of degree $k < \ell$ is a polynomial of degree k. For the same reason, this condition also arises in the approximation theory of neural networks [49].

2.2. *Approximation error of neural tangent models.* For the NT model, the proof, while following the same scheme as for RF, is more challenging. We restrict our setting to a fixed activation function σ (independent of dimensions) which is weakly differentiable, with weak derivative σ' that does not grow too fast (in particular, exponential growth is fine). We further require the Hermite decomposition of σ' to satisfy a mild "genericity" condition. Recall that the kth Hermite coefficient of a function h can be defined as $\mu_k(h) \equiv \mathbb{E}_{G\sim \mathsf{N}(0,1)}\{h(G)\mathrm{He}_k(G)\}$, where $\mathrm{He}_k(x)$ is the kth Hermite polynomial (see Section 5 for further background).

ASSUMPTION 2 (Assumptions for the NT model at level $\ell \in \mathbb{N}$). Let σ be an activation function $\sigma : \mathbb{R} \to \mathbb{R}$.

(a) The function σ is weakly differentiable, with weak derivative σ' such that $\sigma'(u)^2 \le c_0 \exp(c_1 u^2/2)$ for some constants c_0, c_1, with $c_1 < 1$.

(b) The Hermite coefficients $\{\mu_k(\sigma')\}_{k\ge 0}$ are such that there exist $k_1, k_2 \ge 2\ell + 7$ such that $\mu_{k_1}(\sigma'), \mu_{k_2}(\sigma') \ne 0$ and

$$\frac{\mu_{k_1}(x^2\sigma')}{\mu_{k_1}(\sigma')} \ne \frac{\mu_{k_2}(x^2\sigma')}{\mu_{k_2}(\sigma')}. \tag{6}$$

(c) The Hermite coefficients of σ satisfy $\mu_k(\sigma) \ne 0$ for any $k \le \ell + 1$.

THEOREM 2 (Risk of the NT model). *Let $\{f_d \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))\}_{d\ge 1}$ be a sequence of functions. Let $\boldsymbol{W} = (\boldsymbol{w}_i)_{i\in[N]}$ with $(\boldsymbol{w}_i)_{i\in[N]} \sim \mathsf{Unif}(\mathbb{S}^{d-1})$ independently. We have the following results*:

(a) *Assume $N = o_d(d^{\ell+1})$ for a fixed integer ℓ, and let σ satisfy Assumptions 2(a) and 2(b) at level ℓ. Then, for any $\varepsilon > 0$, the following holds with high probability*:

$$\begin{aligned} &\big|R_{\mathsf{NT}}(f_d, \boldsymbol{W}) - R_{\mathsf{NT}}(\mathsf{P}_{\le \ell+1} f_d, \boldsymbol{W}) - \|\mathsf{P}_{>\ell+1} f_d\|_{L^2}^2\big| \\ &\quad \le \varepsilon \|f_d\|_{L^2} \|\mathsf{P}_{>\ell+1} f_d\|_{L^2}. \end{aligned} \tag{7}$$

(b) *Assume $N = \omega_d(d^\ell)$ for some integer ℓ, and let σ satisfy Assumptions 2(a) and 2(c) at level ℓ. Then for any $\varepsilon > 0$, the following holds with high probability*:

$$0 \le R_{\mathsf{NT}}(\mathsf{P}_{\le \ell+1} f_d, \boldsymbol{W}) \le \varepsilon \|\mathsf{P}_{\le \ell+1} f_d\|_{L^2}^2. \tag{8}$$

See Section 6 for the proof of lower bound, and Appendix D in the Supplementary Material for the proof of upper bound.

REMARK 4. It is easy to check that Assumptions 2(a) and 2(b) hold for all ℓ, for all commonly used activations.

For instance, the ReLU activation $\sigma(u) = \max(u, 0)$ and its weak derivative $\sigma'(x) = \mathbf{1}_{x\ge 0}$ have subexponential growth. Its Hermite coefficients are easily computed. Indeed, $\mu_0(\sigma') = \mathbb{E}_{G\sim \mathsf{N}(0,1)}\{\sigma'(G)\} = 1/2$. Further recall that $\mathrm{He}_k(x) \equiv (-1)^k \phi^{(k)}(x)/\phi(x)$ with

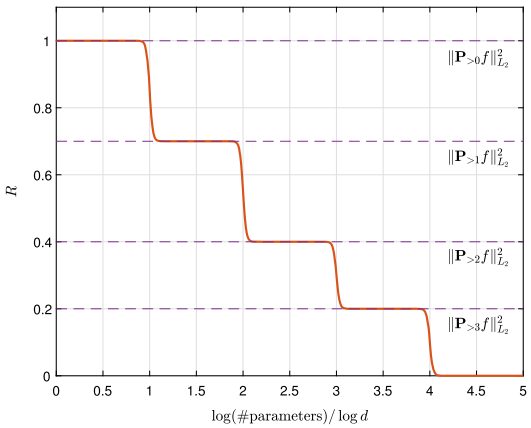

FIG. 5. *A cartoon of the approximation error versus number of parameters in the* RF*and* NT *models. (This sketch is an illustration of the predictions of Theorems* 1 *and* 2*. It is not a numerical result by of solving a finite-N approximation problem.)*

$\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$ the standard Gaussian density and $\phi^{(k)}(x)$ its kth derivative. We thus get, for $k \geq 1$,

$$\begin{aligned}\mu_k(\sigma') &= \mathbb{E}_{G\sim\mathsf{N}(0,1)}\{\sigma'(G)\mathrm{He}_k(G)\} = \int_0^\infty (-1)^k \phi^{(k)}(x)\,\mathrm{d}x \\ &= (-1)^{k-1}\phi^{(k-1)}(0) = \frac{(-1)^{(k-1)/2}}{\sqrt{2\pi}}(k-2)!!\mathbf{1}_{k\text{ odd}},\end{aligned} \tag{9}$$

which satisfy the required condition of Theorem 2(a) for each ℓ. (In checking the condition, it might be useful to notice the relation $\mu_k(x^2\sigma') = \mu_{k+2}(\sigma') + (2k+1)\mu_k(\sigma') + k(k-1)\mu_{k-2}(\sigma')$.)

Assumption 2(c) does not hold for ReLU activation $\sigma(u) = \max(u, 0)$, since $\mu_k(\sigma) = 0$ for k even. However, it holds for shifted ReLU $\sigma(u) = \max(u - u_0, 0)$, for a generic value of the shift u_0.

Theorems 1 and 2 can be illustrated by a cartoon, which is shown as Figure 5. In words, the approximation error plotted as a function of $\log(\#\text{parameters})/\log d$ follows a staircase: it drops at the integer values of this ratio, with the size of the ℓth drop corresponding to $\|\mathsf{P}_\ell f_\star\|_{L^2}^2$. We can extract three useful statistical insights from these findings:

1. There is no difference between RFand NT in terms of approximation power, once we compare them at fixed number of parameters p. Note that $p = N$ for RF, and $p = Nd$ for NT. The recent work [28] actually shows some advantage for the RFmodel, although in a special case. It is worth mentioning that the same equivalence holds in the infinite-width finite-sample regime; see Section 3.

We notice however an important computational advantage for NT over RFat the same number of parameters. Indeed, the complexity at prediction time is $O(Nd) = O(p)$ for NT, while it is $O(Nd) = O(pd)$ for RF.

2. RFor NT models have similar performance to polynomial regressions. Note that the space of degree-ℓ polynomials in d dimension has $\Theta_d(d^\ell)$ degrees of freedom. For any function $f_\star$, degree-ℓ polynomial regression gives approximation error $\|\mathsf{P}_{>\ell} f_\star\|_{L^2}^2$.

3. Our results also suggest interesting directions to improve random features methods. First, if $f_\star$ is known to primarily depend on $\boldsymbol{x}$ projected onto a low-dimensional subspace $\mathcal{V} \subseteq \mathbb{R}^d$ with $\dim(\mathcal{V}) = d_1 \ll d$, there will be a significant advantage in choosing the bottom-layer

weights $\boldsymbol{w}_i$'s along that d_1-dimensional subspace. Second, if the distribution of feature $\boldsymbol{x}$ lies close to such a subspace $\mathcal{V}$, one might hope that—even if the $\boldsymbol{w}_i$'s are sampled isotropically on $\mathbb{S}^{d-1}$—random features methods will be sensitive to d_1 rather than d. We reported on these topics in a follow-up publication [29].

2.3. *Separation between* NN *and* RF, NT. Theorems 1 and 2 imply a separation of approximation power between two-layers neural networks and their linearization. As a simple example, consider the target function $f_\star(\boldsymbol{x}) = \sigma(\langle \boldsymbol{w}_\star, \boldsymbol{x}\rangle)$, for $\|\boldsymbol{w}_\star\|_2 = 1$. This can be represented exactly by a neural network with $N = 1$, that is, by a single neuron. On the other hand, the above results imply that either RFor NT model is lower bounded by a non-vanishing population error. If $d^{\ell+\delta} \le N \le d^{\ell+1-\delta}$, provided that σ satisfies Assumptions 1 and 2, we get

$$
\begin{aligned}
R_{\mathsf{RF}}(\sigma, \boldsymbol{W}) &= \|\sigma_{>\ell}\|^2_{L^2(\mathbb{R},\gamma)} + o_{d,\mathbb{P}}(1)\cdot \|\sigma\|^2_{L^2(\mathbb{R},\gamma)},\\
R_{\mathsf{NT}}(\sigma, \boldsymbol{W}) &= \|\sigma_{>\ell+1}\|^2_{L^2(\mathbb{R},\gamma)} + o_{d,\mathbb{P}}(1)\cdot \|\sigma\|^2_{L^2(\mathbb{R},\gamma)}.
\end{aligned} \tag{10}
$$

Here, $\sigma_{>k}(x)$ is the projection of σ orthogonal to the subspace of polynomials of maximum degree k, in $L^2(\mathbb{R},\gamma)$, where $\gamma(\mathrm{d}x) = e^{-x^2/2}\,\mathrm{d}x/\sqrt{2\pi}$ is the standard Gaussian measure.

Crucially, as proven in [41], running gradient descent over the space of neural networks consisting of a single neuron allows to learn the target function $f_\star(\boldsymbol{x}) = \sigma(\langle \boldsymbol{w}_\star, \boldsymbol{x}\rangle)$ efficiently. In other words, we not only have a separation between linearized neural networks ($\mathcal{F}_{\mathsf{RF}}$ and $\mathcal{F}_{\mathsf{NT}}$) and the function class $\mathcal{F}_{\mathsf{NN}}$, but also a separation between linearized neural networks, and neural networks trained by gradient descent.

The same blueprint can be followed to prove further separation results. For instance, consider $f_\star(\boldsymbol{x}) = \varphi(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{x})$, for $\boldsymbol{Q} \in \mathbb{R}^{d\times r}$ an orthogonal matrix and $\varphi : \mathbb{R}^r \to \mathbb{R}$ a bounded smooth function, which is not a polynomial. If r is kept constant as $d^{\ell+\delta} \le N \le d^{\ell+1-\delta}$, Theorems 1 and 2 can be used to show that $R_{\mathsf{RF}}(f_\star, \boldsymbol{W})$, $R_{\mathsf{NT}}(f_\star, \boldsymbol{W})$ are bounded away from zero and to compute their limits. On the other hand, the classical results [39] can be used to show that such $f_\star$ can be approximated arbitrarily well by neural networks with $O_d(1)$ neurons (with bottom-layer weights $\boldsymbol{w}_i$ in the span of columns of $\boldsymbol{Q}$). Unfortunately, we are not aware of general results implying that such neural networks can be learnt by gradient descent, although we expect this to be the case for certain choices of φ. Whenever such a result is available, it implies a separation between linearized neural networks and neural networks trained by gradient descent.

Let us remark that a similar setting was independently considered by Yehudai and Shamir[4] in concurrent work [58]. (The setting of their work is slightly different as they consider $\boldsymbol{x}_i \sim \mathsf{N}(0, \mathbf{I}_d)$ instead of our setting $\boldsymbol{x}_i \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$.) Translating to our notation, they prove that there exist finite constants $c_0, c_1 > 0$ such that, if $N \le \exp\{c_1 d\}$ and the coefficients a_i and $\boldsymbol{a}_i$ in (RF) and (NT) have magnitude at most $\exp\{c_1 d\}$, then there exists $b_\star$, with $|b_\star| \le 6d + 1$, such that the following happens for any $\boldsymbol{w}_\star$ with $\|\boldsymbol{w}_\star\|_2 = 1$: if $f_\star(\boldsymbol{x}) = (\langle \boldsymbol{w}_\star, \boldsymbol{x}\rangle + b_\star)_+$, then $R_{\mathsf{RF}}(f_\star, \boldsymbol{W})$, $R_{\mathsf{NT}}(f_\star, \boldsymbol{W}) \ge c_0/d^6$ with high probability. An important difference of their result with our analysis is that Theorems 1 and 2 apply to any function $f_\star$ (and equation (10) can be generalized to other functions of low-dimensional projections), while the result of [58] applies to the specific function $f_\star(\boldsymbol{x}) = (\langle \boldsymbol{w}_\star, \boldsymbol{x}\rangle + b_\star)_+$ for a certain $b_\star$. Let us emphasize that there are other important differences between our setting and the one of [58], and neither of the two analysis implies the other.

[4]We refer here to the latest available version of their result, posted as arXiv:1904.00687v3, Theorem 4.1. We adapt the normalization adopted in that paper, dividing $f_\star$ by d^3, so that $\|f_\star\|_{L^2} = \Theta(1)$, when $b_\star = O(1)$.

3. Generalization error of kernel methods. We consider next the infinite-width finite-sample regime. Namely, we let $N = \infty$, and n, d diverge while being polynomially related. It is known since the work of Rahimi and Recht [50] that ridge regression over the function class $\mathcal{F}_{\mathsf{RF}}(\boldsymbol{W})$ converges in this limit to kernel ridge regression (KRR) with respect to the kernel (here expectation is with respect to $\boldsymbol{w} \sim \mathsf{Unif}(\mathbb{S}^{d-1}(1))$)

$$H_d^{\mathsf{RF}}(\boldsymbol{x}_1, \boldsymbol{x}_2) := h_d^{\mathsf{RF}}\big(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d\big) = \mathbb{E}\big\{\sigma\big(\langle \boldsymbol{w}, \boldsymbol{x}_1\rangle\big)\sigma\big(\langle \boldsymbol{w}, \boldsymbol{x}_2\rangle\big)\big\}. \tag{11}$$

Analogously, ridge regression in $\mathcal{F}_{\mathsf{NT}}(\boldsymbol{W})$ can be shown to converge to KRR with respect to the kernel

$$\begin{aligned} H_d^{\mathsf{NT}}(\boldsymbol{x}_1, \boldsymbol{x}_2) &:= h_d^{\mathsf{NT}}\big(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d\big) \\ &= \big(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d\big)\mathbb{E}\big\{\sigma'\big(\langle \boldsymbol{w}, \boldsymbol{x}_1\rangle\big)\sigma'\big(\langle \boldsymbol{w}, \boldsymbol{x}_2\rangle\big)\big\}. \end{aligned} \tag{12}$$

We will denote the corresponding RKHS by $\mathcal{H}_{\mathsf{RF}}$ and $\mathcal{H}_{\mathsf{NT}}$. Quantitative estimates on the relation between $\mathcal{F}_{\mathsf{RF}}(\boldsymbol{W})$ and $\mathcal{H}_{\mathsf{RF}}$ are obtained in [6], which shows that the unit ball of $\mathcal{H}_{\mathsf{RF}}$ is well approximated by the unit ball of $\mathcal{F}_{\mathsf{RF}}(\boldsymbol{W})$ (endowed with the ℓ_2 norm of the coefficients $(a_i)_{i\le N}$), for N large enough.

Notice that both kernels H_d^{RF}, H_d^{NT} are rotationally invariant, namely $H_d(\boldsymbol{S}\boldsymbol{x}_1, \boldsymbol{S}\boldsymbol{x}_2) = H_d(\boldsymbol{x}_1, \boldsymbol{x}_2)$ for $H_d \in \{H_d^{\mathsf{RF}}, H_d^{\mathsf{NT}}\}$ and any $d \times d$ orthogonal matrix $\boldsymbol{S}$. Any rotationally invariant kernel on the sphere $\mathbb{S}^{d-1}(\sqrt{d})$ takes the form

$$H_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = h_d\big(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d\big), \tag{13}$$

for some function $h_d : [-1, 1] \to \mathbb{R}$. (The scaling factor d is introduced here to make contact with the normalization used in previous sections, and is not necessary: indeed, h_d can depend itself on d.)

Our results apply to general rotational invariant kernels under very weak conditions on the function h_d. In particular, they apply to *multilayer neural networks* in the neural tangent regime. Namely, consider a L-layers network with matrix weights $\boldsymbol{W}_1 \in \mathbb{R}^{N_1\times d}$, $\boldsymbol{W}_2 \in \mathbb{R}^{N_2\times N_1}, \dots \boldsymbol{W}_{L-1} \in \mathbb{R}^{N_{L-1}\times N_{L-2}}$, $\boldsymbol{a} \in \mathbb{R}^{N_{L-1}}$. As long as all the weights are initialized as independent centered Gaussians, with variance dependent only on the layer, the resulting NT kernel is rotationally invariant. The recent papers [2, 4, 22, 23, 59] provide conditions under which the NT approximation is accurate for SGD-trained multilayer neural networks.

Section 3.1 presents a lower bound on the prediction error of general kernel methods, and Section 3.2 derives an upper bound for kernel ridge regression.

Throughout this section, we consider the same data model as in the previous sections: we observe pairs $\{(y_i, \boldsymbol{x}_i)\}_{i\le n}$, with $(\boldsymbol{x}_i)_{i\le n} \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, and $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$, $f_\star \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ and $\varepsilon_i \sim \mathsf{N}(0, \tau^2)$ independently.

3.1. *Lower bound for general kernel methods.* Consider any regression method of the form

$$\hat{f}_\lambda = \arg\min_f \left\{\sum_{i=1}^n \ell\big(y_i, f(\boldsymbol{x}_i)\big) + \lambda \|f\|_H^2\right\}, \tag{14}$$

where $\|f\|_H$ is the reproducing kernel Hilbert space (RKHS) norm with respect to the kernel H of the form (13). By the representer theorem [13], there exist coefficients $\hat{a}_1, \dots, \hat{a}_n$ such that

$$\hat{f}_\lambda(\boldsymbol{x}) = \sum_{i=1}^n \hat{a}_i h_d\big(\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle/d\big). \tag{15}$$

We are therefore led to define the following data-dependent prediction risk function for kernel methods

$$R_H(f_\star, \boldsymbol{X}) \equiv \min_{\boldsymbol{a}} \mathbb{E}_{\boldsymbol{x}} \left\{ \left(f_\star(\boldsymbol{x}) - \sum_{i=1}^{n} a_i h_d\big(\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle/d\big)\right)^2 \right\}. \tag{16}$$

The next theorem provides a decomposition of this generalization error that is analogous to the one given in Theorem 1(a). Notice however that the controlling factor is not the number of neurons N, but instead the sample size n.

THEOREM 3. *Assume $n \le d^{\ell+1-\delta_d}$ for a fixed integer ℓ and any sequence δ_d such that $\delta_d^2 \log d \to \infty$ (in particular, $n \le d^{\ell+1-\delta}$ is sufficient for any fixed $\delta > 0$). Let $\{f_d \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))\}_{d\ge 1}$ be a sequence of functions, $\{\boldsymbol{x}_i\}_{i\in[n]} \sim_{\text{i.i.d.}} \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ with $y_i = f_d(\boldsymbol{x}_i)$. Assume $h_n(\langle \boldsymbol{e}_1, \cdot\rangle/\sqrt{d}) \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$. Then for any $\varepsilon > 0$, with high probability as $d \to \infty$, we have*

$$\big|R_H(f_d, \boldsymbol{X}) - R_H(\mathsf{P}_{\le \ell} f_d, \boldsymbol{X}) - \|\mathsf{P}_{>\ell} f_d\|_{L^2}^2\big| \le \varepsilon \|f_d\|_{L^2} \|\mathsf{P}_{>\ell} f_d\|_{L^2}. \tag{17}$$

PROOF. This follows immediately from Theorem 1(a). Indeed, setting $\sigma_d(u) = h_d(u/\sqrt{d})$ and $\boldsymbol{w}_i = \boldsymbol{x}_i/\sqrt{d}$, we obtain $R_H(f_d, \boldsymbol{X}) = R_{\mathsf{RF}}(f_d, \boldsymbol{W})$, whence the claim follows by applying equation (3). □

3.2. *Upper bound for kernel ridge regression.* Kernel ridge regression is one specific way of selecting the coefficients $\hat{\boldsymbol{a}}$ in equation (15), namely by using $\ell(\hat{y}, y) = (\hat{y} - y)^2$ in equation (14). Solving for the coefficients yields

$$\hat{\boldsymbol{a}} = (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{y},$$

where the kernel matrix $\boldsymbol{H} = (H_{ij})_{ij\in[n]}$ is given by

$$H_{ij} = h_d\big(\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle/d\big),$$

and $\boldsymbol{y} = (y_1, \dots, y_n)^{\mathsf{T}}$. The prediction function at location $\boldsymbol{x}$ is given by

$$\hat{f}_\lambda(\boldsymbol{x}) = \boldsymbol{y}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{h}(\boldsymbol{x}),$$

where

$$\boldsymbol{h}(\boldsymbol{x}) = \big[h_d\big(\langle \boldsymbol{x}, \boldsymbol{x}_1\rangle/d\big), \dots, h_d\big(\langle \boldsymbol{x}, \boldsymbol{x}_n\rangle/d\big)\big]^{\mathsf{T}}.$$

The test error of empirical kernel ridge regression is defined as

$$R_{\mathsf{KR}}(f_d, \boldsymbol{X}, \lambda) \equiv \mathbb{E}_{\boldsymbol{x}}\big[\big(f_d(\boldsymbol{x}) - \boldsymbol{y}^{\mathsf{T}} (\boldsymbol{H} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{h}(\boldsymbol{x})\big)^2\big].$$

We assume that $\{h_d\}_{d\ge 1}$ are positive-definite kernels, and we consider the associated eigenvalues:

$$\xi_{d,k}(h_d) = \int_{[-\sqrt{d},\sqrt{d}]} h_d(x/\sqrt{d})\, Q_k^{(d)}(\sqrt{d}x)\, \tau_{d-1}^1(\mathrm{d}x), \tag{18}$$

where we recall that $Q_k^{(d)}$ is the kth Gegenbauer polynomial.

ASSUMPTION 3 (Assumption for KRR at level $\ell \in \mathbb{N}$). Let $\{h_d\}_{d\ge 1}$ be a sequence of functions $h_d : \mathbb{R} \to \mathbb{R}$, such that $H_d(\boldsymbol{x}_1, \boldsymbol{x}_2) = h_d(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle/d)$ is a positive semidefinite kernel.

(a) $h_d(\cdot/\sqrt{d}) \in L^2([-\sqrt{d}, \sqrt{d}], \tau_{d-1}^1)$, where τ_{d-1}^1 is the distribution of $\langle \boldsymbol{x}, \boldsymbol{e}\rangle$ for $\boldsymbol{x} \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, where $\boldsymbol{e} = (1, 0, \dots, 0)^{\mathsf{T}} \in \mathbb{R}^d$.

(b) There exists a constant $c_\ell > 0$ such that

$$\frac{d^\ell \min_{k\le\ell} \xi_{d,k}(h_d)}{\sum_{k\ge\ell+1} \xi_{d,k}(h_d) B(d,k)} \ge c_\ell, \tag{19}$$

with $B(d,k) \equiv \frac{2k+d-2}{d-2}\binom{k+d-3}{k}$.

THEOREM 4. *Assume $\omega_d(d^\ell \log d) \le n \le O_d(d^{\ell+1-\delta})$ for some integer ℓ and $\delta > 0$. Let $\{f_d \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))\}_{d\ge1}$ be a sequence of functions. Let $\{h_d\}_{d\ge1}$ be a sequence of kernels satisfying Assumption 3 at level ℓ. Further, define*

$$\lambda_*(d,\ell) := d^\ell \min_{k\le\ell} \xi_{d,k}(h_d). \tag{20}$$

If h_d has zero mean (i.e., $\int h_d(\sqrt{d}\langle \boldsymbol{e}_1, \boldsymbol{x}\rangle)\tau_d(\mathrm{d}\boldsymbol{x}) = 0$) further assume that f_d is centered (i.e., $\int f_d(\boldsymbol{x})\tau_d(\mathrm{d}\boldsymbol{x}) = 0$).

Let $\boldsymbol{X} = (\boldsymbol{x}_i)_{i\in[n]}$ with $(\boldsymbol{x}_i)_{i\in[n]} \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ independently, and $y_i = f_d(\boldsymbol{x}_i) + \varepsilon_i$ and $\varepsilon_i \sim_{\text{i.i.d.}} \mathsf{N}(0,\tau^2)$. Then for any $\varepsilon > 0$, and any regularization parameter $\lambda \in (0, \lambda_*)$ with high probability we have

$$\left|R_{\mathsf{KR}}(f_d, \boldsymbol{X}, \lambda) - \|\mathsf{P}_{>\ell} f_d\|_{L^2}^2\right| \le \varepsilon\left(\|f_d\|_{L^2}^2 + \tau^2\right). \tag{21}$$

See Appendix E in the Supplementary Material for the proof of this theorem.

REMARK 5. Assume $h_d \to h$ as $d \to \infty$, uniformly over $[-\delta, \delta]$, together with its derivatives, and further assume $|h_d(x)| \le c_0 \exp(c_1 x^2/2)$ for some $c_0 > 0$, $c_1 < 1$. We expect this to be the case for many kernels of interest, and in particular it can be shown to be the case for h_d^{RF} and h_d^{NT} under mild conditions on the activation σ. Using Rodrigues' formula described in Section 5.2, by an application of integration by part followed by dominated convergence, we get

$$\xi_{d,k}(h_d) = \frac{1}{d^k} h^{(k)}(0) + o_d\left(d^{-k-1}\right), \tag{22}$$

where $h^{(k)}$ is the kth derivative of h. Notice further that $\xi_{d,k}(h_d) \ge 0$ for all k since h_d is positive semidefinite by definition. Therefore, as long as $h^{(k)}(0) > 0$ for all $k \le \ell$, Assumption 3 is satisfied, and $\lambda_*(d,\ell)$ is bounded away from 0.

REMARK 6. For $h_d = h_d^{\mathsf{RF}}$ and if the activation $\sigma \in L^2(\mathbb{R}, \gamma)$ is independent of d, we have $\xi_{d,k}(h_d) = \mu_k(\sigma)^2 d^{-k} + o_d(d^{-k-1})$ and, therefore, Assumption 3 is satisfied as soon as $\mu_k(\sigma) \ne 0$ for all $k \le \ell$.

Notice that the setting of Theorem 4 is the same as in classical nonparametric regression. However, classical theory typically establishes minimax consistency rates of the form $\mathbb{E}\{[\hat{f}(\boldsymbol{x}) - f_\star(\boldsymbol{x})]^2\} \le C(d) n^{-2\beta/(2\beta+d)}$ [32, 56]. In order to guarantee a fixed (small) error, these bounds require $n \ge \exp\{cd\}$. Modern machine learning typically have $d \ge 100$ and n between 10^4 and 10^8, and it is therefore unrealistic to consider n exponential in d. This regime motivates a new type of question: *assuming $n \asymp d^\alpha$, what is the minimum prediction error that can be achieved?* This question is addressed by Theorem 4.

3.3. *Separation between kernel methods and neural networks.* Repeating the same argument of Section 2.3, we see that Theorems 3 and 4 imply a separation between kernel methods, with rotationally invariant kernels, and gradient-descent trained neural networks.

Namely, consider again the target function $f_\star(\boldsymbol{x}) = \sigma(\langle \boldsymbol{w}_\star, \boldsymbol{x}\rangle)$, for $\|\boldsymbol{w}_\star\|_2 = 1$. As proven in [41], $f_\star$ can be learnt efficiently by minimizing the following empirical risk via gradient descent:

$$\hat{R}_{\mathsf{NN}}(\boldsymbol{w}; \boldsymbol{w}_\star) := \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \sigma\big(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle\big)\big)^2.$$

Namely, if $n \geq Cd\log d$ samples are used (and under some technical conditions on σ), gradient descent reaches prediction error of order $(d\log d)/n$

In contrast, Theorems 3 and 4 imply that, for any integer ℓ, and any $d^{\ell+\delta} \leq n \leq d^{\ell+1-\delta}$, any kernel method has test error bounded away from zero. Namely,

$$R_H(\sigma; \boldsymbol{X}) = \|\sigma_{>\ell}\|^2_{L^2(\mathbb{R},\gamma)} + o_{d,\mathbb{P}}(1)\cdot\|\sigma\|^2_{L^2(\mathbb{R},\gamma)}. \tag{23}$$

This test error is achieved by kernel ridge regression.

3.4. *Near-optimality of interpolators.* Let us emphasize some important statistical aspects of Theorem 4. KRR is proved to achieve near optimal prediction error (matching the lower bound of Theorem 3) *pointwise*, that is, per given function f_d. What is the nature of the predictor $\hat{f}_\lambda$? Theorems 3 and 4 imply that, in ℓ_2 sense, $\hat{f}_\lambda$ must be close to a low-degree approximation of f_d, namely $\mathsf{P}_{\leq\ell} f_d$.

Optimal test error is achieved for any $\lambda < \lambda_*$. In particular, by taking $\lambda \to 0$, we obtain an *interpolator*, that is, a predictor that interpolates the dataset $\{(y_i, \boldsymbol{x}_i)\}_{i\leq n}$. This remark is made quantitative in the following bound on the empirical risk:

$$\hat{R}_{\mathsf{KR}}(f_d, \boldsymbol{X}, \lambda) \equiv \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \hat{f}_\lambda(\boldsymbol{x})\big)^2. \tag{24}$$

THEOREM 5. *Assume $\omega_d(d^\ell\log d) \leq n \leq O_d(d^{\ell+1-\delta})$ for some integer ℓ and $\delta > 0$. Under the same assumptions of Theorem 4, if $\lambda < \lambda_*$, then*

$$\hat{R}_{\mathsf{KR}}(f_d, \boldsymbol{X}, \lambda) \leq \big(1 + o_{d,\mathbb{P}}(1)\big)\big(\|f_d\|^2_{L^2} + \tau^2\big)\bigg(\frac{\lambda}{\lambda+\kappa_h}\bigg)^2, \tag{25}$$

where $\kappa_h = \sum_{k\geq\ell+1}\xi_{d,k}(h_d)B(d,k)$.

PROOF OF THEOREM 5. Recall that the empirical risk of KRR is given by equation (24), where $\hat{\boldsymbol{f}}_\lambda = (\hat{f}_\lambda(\boldsymbol{x}_1), \dots, \hat{f}_\lambda(\boldsymbol{x}_n))$ can be rewritten as

$$\hat{\boldsymbol{f}}_\lambda = \boldsymbol{H}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{y}.$$

Therefore,

$$\begin{aligned}\hat{R}_{\mathsf{KR}}(f_d, \boldsymbol{X}, \lambda) &= \big\|\big[\mathbf{I}_n - \boldsymbol{H}(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\big]\boldsymbol{y}\big\|_2^2/n\\ &= \lambda^2\big\|(\boldsymbol{H} + \lambda\mathbf{I}_n)^{-1}\boldsymbol{y}\big\|_2^2/n.\end{aligned}$$

From the proof of Theorem 4, we have the following lower bound on the eigenvalues $\boldsymbol{H} + \lambda\mathbf{I}_n \succeq (\kappa_h + \lambda + o_{d,\mathbb{P}}(1))\mathbf{I}_n$. We deduce that with high probability

$$\begin{aligned}\hat{R}_{\mathsf{KR}}(f_d, \boldsymbol{X}, \lambda) &\leq \big(1 + o_{d,\mathbb{P}}(1)\big)\big(\lambda/(\kappa_h+\lambda)\big)^2\|\boldsymbol{y}\|_2^2/n\\ &\leq \big(1 + o_{d,\mathbb{P}}(1)\big)\big(\|f_d\|^2_{L^2} + \tau^2\big)\big(\lambda/(\kappa_h+\lambda)\big)^2,\end{aligned}$$

where we simply used the law of large numbers $\|\boldsymbol{y}\|_2^2/n \to \|f_d\|^2_{L^2} + \tau^2$. □

The statistical properties of interpolators have attracted substantial interest over the last 2 years. Classical statistical wisdom suggests that a model that fits too well the training set is too complex and cannot generalize well. In contrast, modern machine learning methods often operate at or near such an interpolation regime. Examples of interpolators with provable optimality properties were presented among others in [11, 12, 37].

3.5. *A conjecture for generalization error of random features model.* Consider random features regression with finite sample size and a finite number of neurons. We fit data $\{(y_i, \boldsymbol{x}_i)\}_{i\le n}$ using ridge regression in the random features (RF) model, with (where $\boldsymbol{w}_i \sim_{\text{i.i.d.}} \mathsf{Unif}(\mathbb{S}^{d-1}(1))$)

$$\hat{\boldsymbol{a}}(\lambda) = \arg\min_{\boldsymbol{a}\in\mathbb{R}^N} \left\{ \frac{1}{n}\sum_{j=1}^{n} \left(y_j - \sum_{i=1}^{N} a_i \sigma\big(\langle \boldsymbol{w}_i, \boldsymbol{x}_j\rangle\big)\right)^2 + \frac{N\lambda}{d}\|\boldsymbol{a}\|_2^2 \right\}. \tag{26}$$

Under the same data model of the previous sections, we are interested in the test error

$$R_{\mathsf{RF}}(f_d, \boldsymbol{X}, \boldsymbol{W}, \lambda) = \mathbb{E}_{\boldsymbol{x}}\left[\left(f_d(\boldsymbol{x}) - \sum_{i=1}^{N} \hat{a}_i(\lambda)\sigma\big(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle\big)\right)^2 \right]. \tag{27}$$

Theorem 1 characterizes the test error $R_{\mathsf{RF}}(f_d, \boldsymbol{X}, \boldsymbol{W}, \lambda)$ in the population limit $n = \infty$, whereas Theorems 3 and 4 characterize the same quantity in the case when $N = \infty$.

What happens when both n and N are finite? In the proportional regime $N \propto d$ and $n \propto d$, the precise asymptotics of $R_{\mathsf{RF}}(f_d, \boldsymbol{X}, \boldsymbol{W}, \lambda)$ was calculated in [43].

What happens beyond the proportional asymptotics? We conjecture that the limiting factor is given by the smallest of n and N. Namely, if $d^{\ell+\delta} \le \min(n, N) \le d^{\ell+1-\delta}$ for some positive δ, then the prediction error is the same as the one of fitting a degree-ℓ polynomial, that is, $R_{\mathsf{RF}}(f_d, \boldsymbol{X}, \boldsymbol{W}, \lambda) = \|\mathsf{P}_{>\ell} f_d\|_{L^2}^2 + \|f_d\|_{L^2}^2 \cdot o_{d,\mathbb{P}}(1)$. We leave this conjecture to future work.

4. Further related work. Donoho and Johnstone [21] study an approximation problem analogous to the one we considered in Section 2, although in $d = 2$ dimensions. Their problem essentially reduces to determining rates of approximation on the unit circle, with the technical difference that the $\boldsymbol{w}_i$'s are equispaced along the circle instead of being random. As for other references mentioned in Section 1.2, the lower bounds of [21] are worst case over differentiable functions.

The limitations of kernel methods in high-dimension are studied by El Karoui in [25] (see also [26]), which analyzes kernel random matrices of the form $\boldsymbol{H} = (h(\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle/d))_{i,j\le n}$. The analysis of [25] is limited to the proportional asymptotics $n \propto d$. and establishes that in this regime $\boldsymbol{H}$ is well approximated by the Gram matrix of raw feature vectors plus a diagonal term: $\boldsymbol{H} \approx (h(1) - h'(0))\mathbf{I}_n + h'(0)\boldsymbol{G}$, where $\boldsymbol{G} = (\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle/d)_{i,j\le n}$. This result is related to our Theorems 3 and 4, which deal with kernel methods. However, our results analyze general polynomial scalings $n = O_d(d^{\ell+1-\delta})$, while [25] assumes $n = \Theta_d(d)$. Also [25] analyzes the spectrum of $\boldsymbol{H}$ but not the prediction error of kernel methods. Finally, a large part of our technical work is devoted to RFand NT models (cf. Theorems 1 and 2), which are not touched upon by [25].

Recent work of Vempala and Wilmes [57] analyzes what amounts to an RFmodel. These authors prove that RFcan learn a degree-ℓ polynomial from $n = d^{O(\ell)}$ samples using $N = d^{O(\ell)}$ neurons, and that at least $d^{\Omega(\ell)}$ queries are needed within the statistical query model. While related, our setting is not directly comparable to theirs. Notice further that we obtain a sharper tradeoff, since we obtain the precise exponents of d.

After the present paper appeared as a preprint, several authors presented important contributions to the same line of work. In particular, Liang, Rakhlin, and Zhai [38] studies kernel

ridge regression in d dimension using $n = O_d(d^\gamma)$ samples. Assuming the target function has bounded RKHS norm, they derive upper and lower bounds on the rate of convergence of the generalization error. This result is related to our Theorem 3. The most important difference is that we do not assume that the target function has bounded RKHS norm. Instead we obtain the precise asymptotics of the generalization error in a regime in which it is nonvanishing. As illustrated in Section 1.3, this asymptotic analysis captures indeed the actual behavior in practically reasonable settings.

From a technical viewpoint, several of our calculations make use of harmonic analysis over the d-dimensional sphere, as it is natural given that $\boldsymbol{x}_i$'s are uniform over the sphere. Spherical harmonics expansion appear in related contexts, for example, in [7, 21, 57].

Let us finally mention that an alternative approach to the analysis of two-layers neural networks in the wide limit, was developed in [16, 42, 44, 51, 53] using mean field theory. Unlike in the neural tangent approach, the evolution of network weights is described beyond the linear regime in this theory.

5. Technical background. In this section, we introduce some notation and technical background which will be useful for the proofs in the next sections. In particular, we will use decompositions in (hyper)spherical harmonics on the $\mathbb{S}^{d-1}(\sqrt{d})$ and in orthogonal polynomials on the real line. All of the properties listed below are classical: we will however prove a few facts that are slightly less standard. We refer the reader to [15, 24, 55] for further information on these topics. As mentioned above, expansions in spherical harmonics were used in the past in the statistics literature, for instance, in [7, 21].

5.1. *Functional spaces over the sphere.* For $d \geq 3$, we let $\mathbb{S}^{d-1}(r) = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = r\}$ denote the sphere with radius r in $\mathbb{R}^d$. We will mostly work with the sphere of radius $\sqrt{d}$, $\mathbb{S}^{d-1}(\sqrt{d})$ and will denote by τ_{d-1} the uniform probability measure on $\mathbb{S}^{d-1}(\sqrt{d})$. All functions in the following are assumed to be elements of $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \tau_{d-1})$, with scalar product and norm denoted as $\langle \cdot, \cdot \rangle_{L^2}$ and $\| \cdot \|_{L^2}$:

$$\langle f, g \rangle_{L^2} \equiv \int_{\mathbb{S}^{d-1}(\sqrt{d})} f(\boldsymbol{x}) g(\boldsymbol{x}) \tau_{d-1}(\mathrm{d}\boldsymbol{x}). \tag{28}$$

For $\ell \in \mathbb{Z}_{\geq 0}$, let $\tilde{V}_{d,\ell}$ be the space of homogeneous harmonic polynomials of degree ℓ on $\mathbb{R}^d$ (i.e., homogeneous polynomials $q(\boldsymbol{x})$ satisfying $\Delta q(\boldsymbol{x}) = 0$), and denote by $V_{d,\ell}$ the linear space of functions obtained by restricting the polynomials in $\tilde{V}_{d,\ell}$ to $\mathbb{S}^{d-1}(\sqrt{d})$. With these definitions, we have the following orthogonal decomposition:

$$L^2\big(\mathbb{S}^{d-1}(\sqrt{d}), \tau_{d-1}\big) = \bigoplus_{\ell=0}^{\infty} V_{d,\ell}. \tag{29}$$

The dimension of each subspace is given by

$$\dim(V_{d,\ell}) = B(d, \ell) = \frac{2\ell + d - 2}{d - 2} \binom{\ell + d - 3}{\ell}. \tag{30}$$

For each $\ell \in \mathbb{Z}_{\geq 0}$, the spherical harmonics $\{Y_{\ell,j}^{(d)}\}_{1 \leq j \in \leq B(d,\ell)}$ form an orthonormal basis of $V_{d,\ell}$:

$$\langle Y_{ki}^{(d)}, Y_{sj}^{(d)} \rangle_{L^2} = \delta_{ij}\delta_{ks}.$$

Note that our convention is different from the more standard one, that defines the spherical harmonics as functions on $\mathbb{S}^{d-1}(1)$. It is immediate to pass from one convention to the other

by a simple scaling. We will drop the superscript d and write $Y_{\ell,j} = Y_{\ell,j}^{(d)}$ whenever clear from the context.

We denote by P_k the orthogonal projections to $V_{d,k}$ in $L^2(\mathbb{S}^{d-1}(\sqrt{d}), \tau_{d-1})$. This can be written in terms of spherical harmonics as

$$\mathsf{P}_k f(\boldsymbol{x}) \equiv \sum_{l=1}^{B(d,k)} \langle f, Y_{kl}\rangle_{L^2} Y_{kl}(\boldsymbol{x}). \tag{31}$$

We also define $\mathsf{P}_{\le \ell} \equiv \sum_{k=0}^{\ell} \mathsf{P}_k$, $\mathsf{P}_{>\ell} \equiv \mathbf{I} - \mathsf{P}_{\le \ell} = \sum_{k=\ell+1}^{\infty} \mathsf{P}_k$ and $\mathsf{P}_{<\ell} \equiv \mathsf{P}_{\le \ell-1}$, $\mathsf{P}_{\ge \ell} \equiv \mathsf{P}_{>\ell-1}$.

5.2. *Gegenbauer polynomials.* The ℓth Gegenbauer polynomial $Q_\ell^{(d)}$ is a polynomial of degree ℓ. Consistently with our convention for spherical harmonics, we view $Q_\ell^{(d)}$ as a function $Q_\ell^{(d)} : [-d, d] \to \mathbb{R}$. The set $\{Q_\ell^{(d)}\}_{\ell \ge 0}$ forms an orthogonal basis on $L^2([-d, d], \tilde{\tau}_{d-1}^1)$, where $\tilde{\tau}_{d-1}^1$ is the distribution of $\sqrt{d}\langle \boldsymbol{x}, \boldsymbol{e}_1\rangle$ when $\boldsymbol{x} \sim \tau_{d-1}$, satisfying the normalization condition:

$$\big\langle Q_k^{(d)}(\sqrt{d}\langle \boldsymbol{e}_1, \cdot\rangle), Q_j^{(d)}(\sqrt{d}\langle \boldsymbol{e}_1, \cdot\rangle)\big\rangle_{L^2(\mathbb{S}^{d-1}(\sqrt{d}))} = \frac{1}{B(d,k)}\delta_{jk}. \tag{32}$$

In particular, these polynomials are normalized so that $Q_\ell^{(d)}(d) = 1$. As above, we will omit the superscript d when clear from the context.

Gegenbauer polynomials are directly related to spherical harmonics as follows. Fix $\boldsymbol{v} \in \mathbb{S}^{d-1}(\sqrt{d})$ and consider the subspace of V_ℓ formed by all functions that are invariant under rotations in $\mathbb{R}^d$ that keep $\boldsymbol{v}$ unchanged. It is not hard to see that this subspace has dimension one, and coincides with the span of the function $Q_\ell^{(d)}(\langle \boldsymbol{v}, \cdot\rangle)$.

We will use the following properties of Gegenbauer polynomials:

1. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}(\sqrt{d})$

$$\big\langle Q_j^{(d)}(\langle \boldsymbol{x}, \cdot\rangle), Q_k^{(d)}(\langle \boldsymbol{y}, \cdot\rangle)\big\rangle_{L^2} = \frac{1}{B(d,k)}\delta_{jk} Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle). \tag{33}$$

2. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}(\sqrt{d})$,

$$Q_k^{(d)}\big(\langle \boldsymbol{x}, \boldsymbol{y}\rangle\big) = \frac{1}{B(d,k)} \sum_{i=1}^{B(d,k)} Y_{ki}^{(d)}(\boldsymbol{x}) Y_{ki}^{(d)}(\boldsymbol{y}). \tag{34}$$

3. Recurrence formula

$$\frac{t}{d} Q_k^{(d)}(t) = \frac{k}{2k+d-2} Q_{k-1}^{(d)}(t) + \frac{k+d-2}{2k+d-2} Q_{k+1}^{(d)}(t). \tag{35}$$

4. Rodrigues' formula

$$Q_k^{(d)}(t) = (-1)^k C_{k,d} \left(1 - \frac{t^2}{d^2}\right)^{(3-d)/2} \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^k \left(1 - \frac{t^2}{d^2}\right)^{k+(d-3)/2}, \tag{36}$$

where $C_{k,d} = (d/2)^k \Gamma((d-1)/2)/\,\Gamma(k + (d-1)/2)$.

Note in particular that property 2 implies that—up to a constant—$Q_k^{(d)}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)$ is a representation of the projector onto the subspace of degree-k spherical harmonics

$$(\mathsf{P}_k f)(\boldsymbol{x}) = B(d,k) \int_{\mathbb{S}^{d-1}(\sqrt{d})} Q_k^{(d)}\big(\langle \boldsymbol{x}, \boldsymbol{y}\rangle\big) f(\boldsymbol{y}) \tau_{d-1}(\mathrm{d}\boldsymbol{y}). \tag{37}$$

For a function $\sigma \in L^2([-\sqrt{d},\sqrt{d}],\tau_{d-1}^1)$ (where τ_{d-1}^1 is the distribution of $\langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle/\sqrt{d}$ when $\boldsymbol{x}_1,\boldsymbol{x}_2 \sim_{\text{i.i.d.}} \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$), denoting its spherical harmonics coefficients $\lambda_{d,k}(\sigma)$ to be

$$\lambda_{d,k}(\sigma) = \int_{[-\sqrt{d},\sqrt{d}]} \sigma(x)\, Q_k^{(d)}(\sqrt{d}x)\tau_{d-1}^1(\mathrm{d}x), \tag{38}$$

then we have the following equation holds in $L^2([-\sqrt{d},\sqrt{d}],\tau_{d-1}^1)$ sense

$$\sigma(x) = \sum_{k=0}^{\infty} \lambda_{d,k}(\sigma) B(d,k)\, Q_k^{(d)}(\sqrt{d}x).$$

To any rotationally invariant kernel $H_d(\boldsymbol{x}_1,\boldsymbol{x}_2) = h_d(\langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle/d)$, with $h_d(\sqrt{d}\cdot) \in L^2([-\sqrt{d},\sqrt{d}],\tau_{d-1}^1)$, we can associate a self-adjoint operator $\mathscr{H}_d : L^2(\mathbb{S}^{d-1}(\sqrt{d})) \to L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ via

$$\mathscr{H}_d f(\boldsymbol{x}) := \int_{\mathbb{S}^{d-1}(\sqrt{d})} h_d\big(\langle \boldsymbol{x},\boldsymbol{x}_1\rangle/d\big) f(\boldsymbol{x}_1)\tau_{d-1}(\mathrm{d}\boldsymbol{x}_1). \tag{39}$$

By rotational invariance, the space V_k of homogeneous polynomials of degree k is an eigenspace of $\mathscr{H}_d$, and we will denote the corresponding eigenvalue by $\xi_{d,k}(h_d)$. In other words, $\mathscr{H}_d f(\boldsymbol{x}) := \sum_{k=0}^{\infty} \lambda_{d,k}(h_d)\mathsf{P}_k f$. The eigenvalues can be computed via

$$\xi_{d,k}(h_d) = \int_{[-\sqrt{d},\sqrt{d}]} h_d(x/\sqrt{d})\, Q_k^{(d)}(\sqrt{d}x)\tau_{d-1}^1(\mathrm{d}x). \tag{40}$$

5.3. *Hermite polynomials.* The Hermite polynomials $\{\mathrm{He}_k\}_{k\ge 0}$ form an orthogonal basis of $L^2(\mathbb{R},\gamma)$, where $\gamma(\mathrm{d}x) = e^{-x^2/2}\mathrm{d}x/\sqrt{2\pi}$ is the standard Gaussian measure, and He_k has degree k. We will follow the classical normalization (here and below, expectation is with respect to $G \sim \mathsf{N}(0,1)$):

$$\mathbb{E}\big\{\mathrm{He}_j(G)\mathrm{He}_k(G)\big\} = k!\delta_{jk}. \tag{41}$$

As a consequence, for any function $g \in L^2(\mathbb{R},\gamma)$, we have the decomposition

$$g(x) = \sum_{k=0}^{\infty} \frac{\mu_k(g)}{k!}\mathrm{He}_k(x),\ \mu_k(g) \equiv \mathbb{E}\big\{g(G)\mathrm{He}_k(G)\big\}. \tag{42}$$

The Hermite polynomials can be obtained as high-dimensional limits of the Gegenbauer polynomials introduced in the previous section. Indeed, the Gegenbauer polynomials (up to a $\sqrt{d}$ scaling in domain) are constructed by Gram–Schmidt orthogonalization of the monomials $\{x^k\}_{k\ge 0}$ with respect to the measure $\tilde{\tau}_{d-1}^1$, while Hermite polynomial are obtained by Gram–Schmidt orthogonalization with respect to γ. Since $\tilde{\tau}_{d-1}^1 \Rightarrow \gamma$ (here $\Rightarrow$ denotes weak convergence), it is immediate to show that, for any fixed integer k,

$$\lim_{d\to\infty} \mathrm{Coeff}\big\{Q_k^{(d)}(\sqrt{d}x)B(d,k)^{1/2}\big\} = \mathrm{Coeff}\bigg\{\frac{1}{(k!)^{1/2}}\mathrm{He}_k(x)\bigg\}. \tag{43}$$

Here and below, for P a polynomial, $\mathrm{Coeff}\{P(x)\}$ is the vector of the coefficients of P. As a consequence, for any fixed integer k, we have

$$\mu_k(\sigma) = \lim_{d\to\infty} \lambda_{d,k}(\sigma)\big(B(d,k)k!\big)^{1/2}, \tag{44}$$

where $\mu_k(\sigma)$ and $\lambda_{d,k}(\sigma)$ are given in equations (42) and (38).

5.4. *Notation.* Throughout the proofs, $O_d(\cdot)$ (resp., $o_d(\cdot)$) denotes the standard big-O (resp., little-o) notation, where the subscript d emphasizes the asymptotic variable. We denote $O_{d,\mathbb{P}}(\cdot)$ (resp., $o_{d,\mathbb{P}}(\cdot)$) the big-O (resp., little-o) in probability notation: $h_1(d) = O_{d,\mathbb{P}}(h_2(d))$ if for any $\varepsilon > 0$, there exists $C_\varepsilon > 0$ and $d_\varepsilon \in \mathbb{Z}_{>0}$, such that

$$\mathbb{P}\big(|h_1(d)/h_2(d)| > C_\varepsilon\big) \le \varepsilon \quad \forall d \ge d_\varepsilon,$$

and respectively: $h_1(d) = o_{d,\mathbb{P}}(h_2(d))$, if $h_1(d)/h_2(d)$ converges to 0 in probability.

We will occasionally hide logarithmic factors using the $\tilde{O}_d(\cdot)$ notation (resp., $\tilde{o}_d(\cdot)$): $h_1(d) = \tilde{O}_d(h_2(d))$ if there exists a constant C such that $h_1(d) \le C(\log d)^C h_2(d)$. Similarly, we will denote $\tilde{O}_{d,\mathbb{P}}(\cdot)$ (resp., $\tilde{o}_{d,\mathbb{P}}(\cdot)$) when considering the big-O in probability notation up to a logarithmic factor.

6. Proof of Theorem 2(a): NT model lower bound.

6.1. *Proof of Theorem 2(a): Outline.* The proof for the NT model follows the same scheme as for the RFcase presented in Appendix A of the Supplementary Material. However, several steps are technically more challenging.

Recall that $(\boldsymbol{w}_i)_{i\in[N]} \sim \mathsf{Unif}(\mathbb{S}^{d-1})$ independently. We define $\boldsymbol{\theta}_i = \sqrt{d}\cdot\boldsymbol{w}_i$ for $i \in [N]$, so that $(\boldsymbol{\theta}_i)_{i\in[N]} \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ independently. Let $\boldsymbol{W} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_N)$, and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)$. We denote $\mathbb{E}_{\boldsymbol{\theta}}$ to be the expectation operator with respect to $\boldsymbol{\theta} \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $\mathbb{E}_{\boldsymbol{x}}$ to be the expectation operator with respect to $\boldsymbol{x} \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and $\mathbb{E}_{\boldsymbol{w}}$ to be the expectation operator with respect to $\boldsymbol{w} \sim \mathsf{Unif}(\mathbb{S}^{d-1}(1))$.

We define the random vector $\boldsymbol{V} = (\boldsymbol{V}_1, \dots, \boldsymbol{V}_N)^{\mathsf{T}} \in \mathbb{R}^{Nd}$, where, for each $j \le N$, $\boldsymbol{V}_j \in \mathbb{R}^d$, and analogously $\boldsymbol{V}_{\le \ell+1} = (\boldsymbol{V}_{1,\le\ell+1}, \dots, \boldsymbol{V}_{N,\le\ell+1})^{\mathsf{T}} \in \mathbb{R}^{Nd}$, $\boldsymbol{V}_{>\ell+1} = (\boldsymbol{V}_{1,>\ell+1}, \dots, \boldsymbol{V}_{N,>\ell+1})^{\mathsf{T}} \in \mathbb{R}^{Nd}$, as follows:

$$\begin{aligned}
\boldsymbol{V}_{i,\le\ell+1} &= \mathbb{E}_{\boldsymbol{x}}\big[[\mathsf{P}_{\le\ell+1} f_d](\boldsymbol{x})\sigma'\big(\langle\boldsymbol{\theta}_i, \boldsymbol{x}\rangle/\sqrt{d}\big)\boldsymbol{x}\big],\\
\boldsymbol{V}_{i,>\ell+1} &= \mathbb{E}_{\boldsymbol{x}}\big[[\mathsf{P}_{>\ell+1} f_d](\boldsymbol{x})\sigma'\big(\langle\boldsymbol{\theta}_i, \boldsymbol{x}\rangle/\sqrt{d}\big)\boldsymbol{x}\big],\\
\boldsymbol{V}_i &= \mathbb{E}_{\boldsymbol{x}}\big[f_d(\boldsymbol{x})\sigma'\big(\langle\boldsymbol{\theta}_i, \boldsymbol{x}\rangle/\sqrt{d}\big)\boldsymbol{x}\big] = \boldsymbol{V}_{i,\le\ell+1} + \boldsymbol{V}_{i,>\ell+1}.
\end{aligned}$$

We define the random matrix $\boldsymbol{U} = (\boldsymbol{U}_{ij})_{i,j\in[N]} \in \mathbb{R}^{Nd\times Nd}$, where, for each $i, j \le N$, $\boldsymbol{U}_{ij} \in \mathbb{R}^{d\times d}$, is given by

$$\boldsymbol{U}_{ij} = \mathbb{E}_{\boldsymbol{x}}\big[\sigma'\big(\langle\boldsymbol{x}, \boldsymbol{\theta}_i\rangle/\sqrt{d}\big)\sigma'\big(\langle\boldsymbol{x}, \boldsymbol{\theta}_j\rangle/\sqrt{d}\big)\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\big]. \tag{45}$$

In what follows, we write $R_{\mathsf{NT}}(f_d) = R_{\mathsf{NT}}(f_d, \boldsymbol{W}) = R_{\mathsf{NT}}(f_d, \boldsymbol{\Theta}/\sqrt{d})$ for the random features risk, omitting the dependence on the weights $\boldsymbol{W} = \boldsymbol{\Theta}/\sqrt{d}$. By the definition and a simple calculation, we have

$$\begin{aligned}
R_{\mathsf{NT}}(f_d) &= \min_{\boldsymbol{a}\in\mathbb{R}^{Nd}} \big\{\mathbb{E}_{\boldsymbol{x}}\big[f_d(\boldsymbol{x})^2\big] - 2\langle\boldsymbol{a}, \boldsymbol{V}\rangle + \langle\boldsymbol{a}, \boldsymbol{U}\boldsymbol{a}\rangle\big\}\\
&= \mathbb{E}_{\boldsymbol{x}}\big[f_d(\boldsymbol{x})^2\big] - \boldsymbol{V}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V},\\
R_{\mathsf{NT}}(\mathsf{P}_{\le\ell+1} f_d) &= \min_{\boldsymbol{a}\in\mathbb{R}^{Nd}} \big\{\mathbb{E}_{\boldsymbol{x}}\big[\mathsf{P}_{\le\ell+1} f_d(\boldsymbol{x})^2\big] - 2\langle\boldsymbol{a}, \boldsymbol{V}_{\le\ell+1}\rangle + \langle\boldsymbol{a}, \boldsymbol{U}\boldsymbol{a}\rangle\big\}\\
&= \mathbb{E}_{\boldsymbol{x}}\big[\mathsf{P}_{\le\ell+1} f_d(\boldsymbol{x})^2\big] - \boldsymbol{V}_{\le\ell+1}^{\mathsf{T}}\boldsymbol{U}^{-1}\boldsymbol{V}_{\le\ell+1}.
\end{aligned}$$

By orthogonality, we have

$$\mathbb{E}_{\boldsymbol{x}}\big[f_d(\boldsymbol{x})^2\big] = \mathbb{E}_{\boldsymbol{x}}\big[[\mathsf{P}_{\le\ell+1} f_d](\boldsymbol{x})^2\big] + \mathbb{E}_{\boldsymbol{x}}\big[[\mathsf{P}_{>\ell+1} f_d](\boldsymbol{x})^2\big],$$

which gives

$$\begin{aligned}
&\big|R_{\mathsf{NT}}(f_d) - R_{\mathsf{NT}}(\mathsf{P}_{\le \ell+1} f_d) - \mathbb{E}_{\boldsymbol{x}}\big[[\mathsf{P}_{>\ell+1} f_d](\boldsymbol{x})^2\big]\big| \\
&\quad = \big|\boldsymbol{V}_{\le \ell+1}^{\mathsf{T}} \boldsymbol{U}^{-1} \boldsymbol{V}_{\le \ell+1} - \boldsymbol{V}^{\mathsf{T}} \boldsymbol{U}^{-1} \boldsymbol{V}\big| \\
&\quad = \big|\boldsymbol{V}_{\le \ell+1}^{\mathsf{T}} \boldsymbol{U}^{-1} \boldsymbol{V}_{\le \ell+1} - (\boldsymbol{V}_{\le \ell+1} + \boldsymbol{V}_{>\ell+1})^{\mathsf{T}} \boldsymbol{U}^{-1} (\boldsymbol{V}_{\le \ell+1} + \boldsymbol{V}_{>\ell+1})\big| \\
&\quad = \big|2\boldsymbol{V}^{\mathsf{T}} \boldsymbol{U}^{-1} \boldsymbol{V}_{>\ell+1} - \boldsymbol{V}_{>\ell+1}^{\mathsf{T}} \boldsymbol{U}^{-1} \boldsymbol{V}_{>\ell+1}\big| \\
&\quad \le 2\big\|\boldsymbol{U}^{-1/2} \boldsymbol{V}_{>\ell+1}\big\|_2 \big\|\boldsymbol{U}^{-1/2} \boldsymbol{V}\big\|_2 + \big\|\boldsymbol{U}^{-1}\big\|_{\mathrm{op}} \|\boldsymbol{V}_{>\ell+1}\|_2^2 \\
&\quad \le 2\big\|\boldsymbol{U}^{-1/2}\big\|_{\mathrm{op}} \|\boldsymbol{V}_{>\ell+1}\|_2 \|f_d\|_{L^2} + \big\|\boldsymbol{U}^{-1}\big\|_{\mathrm{op}} \|\boldsymbol{V}_{>\ell+1}\|_2^2,
\end{aligned}$$

where the last inequality used the fact that

$$0 \le R_{\mathsf{NT}}(f_d) = \|f_d\|_{L^2}^2 - \boldsymbol{V}^{\mathsf{T}} \boldsymbol{U}^{-1} \boldsymbol{V},$$

so that

$$\big\|\boldsymbol{U}^{-1/2} \boldsymbol{V}\big\|_2^2 = \boldsymbol{V}^{\mathsf{T}} \boldsymbol{U}^{-1} \boldsymbol{V} \le \|f_d\|_{L^2}^2.$$

We claim that we have

$$\|\boldsymbol{V}_{>\ell+1}\|_2 / \|\mathsf{P}_{>\ell+1} f_d\|_{L^2} = o_{d,\mathbb{P}}(1), \tag{46}$$

$$\big\|\boldsymbol{U}^{-1}\big\|_{\mathrm{op}} = O_{d,\mathbb{P}}(1), \tag{47}$$

This is achieved in the following two propositions.

PROPOSITION 1 (Expected norm of $\boldsymbol{V}$). *Let σ be an activation function satisfying Assumption 2(a). Define*

$$\begin{aligned}
\mathcal{E}_{\ge \ell} \equiv &\mathbb{E}_{\boldsymbol{\theta}}\big[\big\langle \mathbb{E}_{\boldsymbol{x}}\big[\mathsf{P}_{\ge \ell} f_\star(\boldsymbol{x}) \sigma'\big(\langle \boldsymbol{\theta}, \boldsymbol{x}\rangle/\sqrt{d}\big)\boldsymbol{x}\big], \mathbb{E}_{\boldsymbol{x}}\big[\mathsf{P}_{\ge \ell} f_\star(\boldsymbol{x}) \sigma'\big(\langle \boldsymbol{\theta}, \boldsymbol{x}\rangle/\sqrt{d}\big)\boldsymbol{x}\big]\big\rangle\big] \\
= &\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\big[\mathsf{P}_{\ge \ell} f_\star(\boldsymbol{x}) \mathsf{P}_{\ge \ell} f_\star(\boldsymbol{x}') \mathbb{E}_{\boldsymbol{\theta}}\big[\sigma'\big(\langle \boldsymbol{\theta}, \boldsymbol{x}\rangle/\sqrt{d}\big)\sigma'\big(\langle \boldsymbol{\theta}, \boldsymbol{x}'\rangle/\sqrt{d}\big)\langle \boldsymbol{x}, \boldsymbol{x}'\rangle\big]\big],
\end{aligned}$$

where expectation is with respect to $\boldsymbol{x}, \boldsymbol{x}' \sim_{i.i.d.} \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. Then there exists a constant C (depending only on the constants in Assumption 2(a)) such that, for any $\ell \ge 1$ and $d \ge 6$,

$$\mathcal{E}_{\ge \ell} \le \frac{Cd}{B(d,\ell)} \|\mathsf{P}_{\ge \ell} f_\star\|_{L^2(\mathbb{S}^{d-1}(\sqrt{d}))}^2.$$

PROPOSITION 2 (Lower bound on the kernel matrix). *Let $N = o_d(d^{\ell+1})$ for some $\ell \in \mathbb{Z}_{>0}$, and $(\boldsymbol{\theta}_i)_{i\in[N]} \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ independently. Let σ be an activation that satisfies Assumptions 2(a) and 2(b). Let $\boldsymbol{U} \in \mathbb{R}^{Nd\times Nd}$ be the kernel matrix with i, j block $\boldsymbol{U}_{ij} \in \mathbb{R}^{d\times d}$ defined by equation (45). Then there exists a constant $\varepsilon > 0$ that depends on the activation function σ, such that*

$$\lambda_{\min}(\boldsymbol{U}) \ge \varepsilon$$

with high probability as $d \to \infty$.

Proposition 1 will be proven in the next section, while the longer proof of Proposition 2 is deferred to Appendix C of the Supplementary Material. Proposition 1 shows that

$$\mathbb{E}\big[\|\boldsymbol{V}_{>\ell+1}\|_2^2\big] \le \frac{CNd}{B(d,\ell+2)} \|\mathsf{P}_{>\ell+1} f_d\|_2^2.$$

Note $B(d,\ell+2) = \Theta_d(d^{\ell+2})$, and $N = o_d(d^{\ell+1})$. By Markov inequality, we have equation (46). Equation (47) follows simply by Proposition 2. This proves the theorem.

6.2. *Proof of Proposition* 1. We denote the Gegenbauer decomposition of $\sigma'(\langle \boldsymbol{e}, \cdot \rangle)$ by

$$\sigma'(\langle \boldsymbol{e}, \boldsymbol{x} \rangle) = \sum_{k=0}^{\infty} B(d,k) \lambda_k(\sigma) Q_k(\sqrt{d}\langle \boldsymbol{e}, \boldsymbol{x} \rangle),$$

where

$$\lambda_k(\sigma') = \langle \sigma'(\langle \boldsymbol{e}, \cdot \rangle), Q_k(\sqrt{d}\langle \boldsymbol{e}, \cdot \rangle) \rangle_{L^2}.$$

By Lemma C.1 in the Supplementary Material, applied to function σ' (instead of σ), under Assumption 2(a), we have $\|\sigma'(\langle \boldsymbol{e}, \cdot \rangle)\|_{L^2}^2 \le C$ (for C a constant independent of d). We therefore have (recalling the normalization of the Gegenbauer polynomials in equation (32))

$$\sum_{k=0}^{\infty} \lambda_k(\sigma')^2 B(d,k) = \|\sigma'(\langle \boldsymbol{e}, \cdot \rangle)\|_{L^2}^2 \le C. \tag{48}$$

We define the NT kernel by

$$H(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{\theta}}[\sigma'(\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle / \sqrt{d}) \sigma'(\langle \boldsymbol{\theta}, \boldsymbol{x}' \rangle / \sqrt{d})] \langle \boldsymbol{x}, \boldsymbol{x}' \rangle.$$

Then

$$\begin{aligned} H(\boldsymbol{x}, \boldsymbol{x}') &= \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{k=0}^{\infty} B(d,k) \lambda_k(\sigma') Q_k(\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle) \sum_{l=0}^{\infty} B(d,l) \lambda_l(\sigma') Q_l(\langle \boldsymbol{\theta}, \boldsymbol{x}' \rangle)\right] \langle \boldsymbol{x}, \boldsymbol{x}' \rangle \\ &= \sum_{k=0}^{\infty} B(d,k)^2 \lambda_k(\sigma')^2 \mathbb{E}_{\boldsymbol{\theta}}[Q_k(\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle) Q_k(\langle \boldsymbol{\theta}, \boldsymbol{x}' \rangle)] \langle \boldsymbol{x}, \boldsymbol{x}' \rangle \\ &= \sum_{k=0}^{\infty} d \cdot B(d,k) \lambda_k(\sigma')^2 Q_k(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle) \langle \boldsymbol{x}, \boldsymbol{x}' \rangle / d, \end{aligned} \tag{49}$$

where in the last step we used equation (33). By the recurrence relationship for Gegenbauer polynomials (35), we have

$$\frac{t}{d} Q_k(t) = s_{d,k} Q_{k-1}(t) + t_{d,k} Q_{k+1}(t),$$

where

$$\begin{aligned} s_{d,k} &= \frac{k}{2k+d-2}, \\ t_{d,k} &= \frac{k+d-2}{2k+d-2}. \end{aligned}$$

We use the convention that $t_{d,-1} = 0$. This gives

$$\sup_{d \ge 6, k \ge 0} [s_{d,k+1} + t_{d,k-1}] = \sup_{d \ge 6, k \ge 0} \left[\frac{k+1}{2k+d} + \frac{k+d-3}{2k+d-4}\right] \le 2. \tag{50}$$

Hence we get

$$\begin{aligned} H(\boldsymbol{x}, \boldsymbol{x}') &= \sum_{k=0}^{\infty} d \cdot B(d,k) \lambda_k(\sigma')^2 Q_k(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle) \langle \boldsymbol{x}, \boldsymbol{x}' \rangle / d \\ &= \sum_{k=0}^{\infty} d \cdot B(d,k) \lambda_k(\sigma')^2 [s_{d,k} Q_{k-1}(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle) + t_{d,k} Q_{k+1}(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle)] \\ &= \sum_{k=0}^{\infty} \Gamma_{d,k} Q_k(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle), \end{aligned}$$

where

$$\begin{aligned}\Gamma_{d,k} &= d\cdot\big[t_{d,k-1}\lambda_{k-1}(\sigma')^2 B(d,k-1)+s_{d,k+1}\lambda_{k+1}(\sigma')^2 B(d,k+1)\big]\\ &\le 2dC.\end{aligned}$$

The last inequality follows by equations (48) and (50).

We define

$$\begin{aligned}\mathcal{E}_k &\equiv \mathbb{E}_{\boldsymbol{\theta}}\big[\langle \mathbb{E}_{\boldsymbol{x}}[\mathsf{P}_k f_\star(\boldsymbol{x})\sigma'(\langle\boldsymbol{\theta},\boldsymbol{x}\rangle/\sqrt{d})\boldsymbol{x}], \mathbb{E}_{\boldsymbol{x}}[\mathsf{P}_k f_\star(\boldsymbol{x})\sigma'(\langle\boldsymbol{\theta},\boldsymbol{x}\rangle/\sqrt{d})\boldsymbol{x}]\rangle\big]\\ &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\big[[\mathsf{P}_k f_\star](\boldsymbol{x})H(\boldsymbol{x},\boldsymbol{x}')[\mathsf{P}_k f_\star](\boldsymbol{x}')\big].\end{aligned}$$

Using the fact that the kernel H preserve the decomposition (29), we have

$$\mathcal{E}_{\ge\ell}=\sum_{k\ge\ell}\mathcal{E}_k.$$

Note by equation (49), we have (as always, expectations are with respect to $\boldsymbol{x},\boldsymbol{x}'\sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ independently)

$$\begin{aligned}\mathcal{E}_k &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\big[[\mathsf{P}_k f_\star](\boldsymbol{x})H(\boldsymbol{x},\boldsymbol{x}')[\mathsf{P}_k f_\star](\boldsymbol{x}')\big]\\ &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\Bigg[\sum_{l=1}^{B(d,k)}\lambda_{kl}(f_\star)Y_{kl}(\boldsymbol{x})\Gamma_{d,k}Q_k\big(\langle\boldsymbol{x},\boldsymbol{x}'\rangle\big)\sum_{s=1}^{B(d,k)}\lambda_{ks}(f_\star)Y_{ks}(\boldsymbol{x}')\Bigg]\\ &= \Gamma_{d,k}\sum_{l=1}^{B(d,k)}\sum_{s=1}^{B(d,k)}\lambda_{kl}(f_\star)\lambda_{ks}(f_\star)\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\big[Y_{kl}(\boldsymbol{x})Q_k\big(\langle\boldsymbol{x},\boldsymbol{x}'\rangle\big)Y_{ks}(\boldsymbol{x}')\big]\\ &= \frac{\Gamma_{d,k}}{B(d,k)}\times\sum_{l=1}^{B(d,k)}\sum_{s=1}^{B(d,k)}\lambda_{kl}(f_\star)\lambda_{ks}(f_\star)\delta_{ls}\\ &= \frac{\Gamma_{d,k}}{B(d,k)}\times\|\mathsf{P}_k f_\star\|_{L^2}^2\le\frac{2Cd}{B(d,k)}\cdot\|\mathsf{P}_k f_\star\|_{L^2}^2,\end{aligned}$$

where the fourth equality used the fact that $\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}[Y_{kl}(\boldsymbol{x})Q_k(\langle\boldsymbol{x},\boldsymbol{x}'\rangle)Y_{ks}(\boldsymbol{x}')]=\delta_{ls}/B(d,k)$.

Hence we have

$$\mathcal{E}_{\ge\ell}=\sum_{k=\ell}^{\infty}\mathcal{E}_k\le\frac{2dC}{B(d,\ell)}\cdot\|\mathsf{P}_{\ge\ell}f_\star\|_{L^2}^2,$$

where we used the fact that $B(d,k)$ is nondecreasing in k (see Lemma A.1 in the Supplementary Material). This concludes the proof.

Acknowledgments. The fourth author is supported by NSF Grants DMS-1613091, CCF-1714305, IIS-1741162 and ONR N00014-18-1-2729, NSF DMS-1418362, NSF DMS-1407813.

SUPPLEMENTARY MATERIAL

Supplement to "Linearized two-layers neural networks in high dimension" (DOI: 10.1214/20-AOS1990SUPP; .pdf). The Supplementary Material contains the proofs of Theorem 1(a) in Appendix A, Theorem 1(b) in Appendix B, Proposition 2 in Appendix C, Theorem 2(b) in Appendix D and Theorem 4 in Appendix E. We included additional numerical simulations using Ridge regression in Appendix F.

REFERENCES

[1] ALAOUI, A. E. and MAHONEY, M. W. (2015). Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems* 775–783.

[2] ALLEN-ZHU, Z., LI, Y. and SONG, Z. (2019). A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning* 242–252.

[3] ANTHONY, M. and BARTLETT, P. L. (1999). *Neural Network Learning*: *Theoretical Foundations*. Cambridge Univ. Press, Cambridge. MR1741038 https://doi.org/10.1017/CBO9780511624216

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

[5] BACH, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory* 185–209.

[6] BACH, F. (2017). On the equivalence between kernel quadrature rules and random feature expansions. *J. Mach. Learn. Res.* **18** Paper No. 21, 38. MR3634888

[7] BACH, F. (2017). Breaking the curse of dimensionality with convex neutral networks. *J. Mach. Learn. Res.* **18** Paper No. 19, 53. MR3634886

[8] BARRON, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory* **39** 930–945. MR1237720 https://doi.org/10.1109/18.256500

[9] 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. MR3997901 https://doi.org/10.1073/pnas.1903070116

[10] BELKIN, M., HSU, D. and XU, J. (2019). Two models of double descent for weak features. Available at arXiv:1903.07571.

[11] BELKIN, M., HSU, D. J. and MITRA, P. (2018). Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems* 2300–2311.

[12] BELKIN, M., RAKHLIN, A. and TSYBAKOV, A. B. (2019). Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics* 1611–1619.

[13] BERLINET, A. and THOMAS-AGNAN, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Boston, MA. MR2239907 https://doi.org/10.1007/978-1-4419-9096-9

[14] CAPONNETTO, A. and DE VITO, E. (2007). Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **7** 331–368. MR2335249 https://doi.org/10.1007/s10208-006-0196-8

[15] CHIHARA, T. S. (2011). *An Introduction to Orthogonal Polynomials*. Courier Corporation.

[16] CHIZAT, L. and BACH, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems* 3036–3046.

[17] CHIZAT, L., OYALLON, E. and BACH, F. (2019). On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems* 2933–2943.

[18] CRISTIANINI, N., SHAWE-TAYLOR, J. et al. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge Univ. Press, Cambridge.

[19] CYBENKO, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2** 303–314. MR1015670 https://doi.org/10.1007/BF02551274

[20] DEVORE, R. A., HOWARD, R. and MICCHELLI, C. (1989). Optimal nonlinear approximation. *Manuscripta Math.* **63** 469–478. MR0991266 https://doi.org/10.1007/BF01171759

[21] DONOHO, D. L. and JOHNSTONE, I. M. (1989). Projection-based approximation and a duality with kernel methods. *Ann. Statist.* **17** 58–106. MR0981438 https://doi.org/10.1214/aos/1176347004

[22] DU, S., LEE, J., LI, H., WANG, L. and ZHAI, X. (2019). Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning* 1675–1685.

[23] DU, S. S., ZHAI, X., POCZOS, B. and SINGH, A. (2019). Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*.

[24] EFTHIMIOU, C. and FRYE, C. (2014). *Spherical Harmonics in p Dimensions*. World Scientific Co. Pte. Ltd., Hackensack, NJ. MR3290046 https://doi.org/10.1142/9134

[25] EL KAROUI, N. (2010). The spectrum of kernel random matrices. *Ann. Statist.* **38** 1–50. MR2589315 https://doi.org/10.1214/08-AOS648

[26] EL KAROUI, N. (2010). On information plus noise kernel random matrices. *Ann. Statist.* **38** 3191–3216. MR2722468 https://doi.org/10.1214/10-AOS801

[27] GEIGER, M., SPIGLER, S., JACOT, A. and WYART, M. (2019). Disentangling feature and lazy learning in deep neural networks: An empirical study. Available at arXiv:1906.08034.

[28] GHORBANI, B., MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2019). Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems* 9108–9118.

[29] GHORBANI, B., MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2020). When do neural networks outperform kernel methods? In *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, eds.) 14820–14830. Curran Associates.

[30] GHORBANI, B., MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2021). Supplement to "Linearized two-layers neural networks in high dimension." https://doi.org/10.1214/20-AOS1990SUPP

[31] GIROSI, F., JONES, M. and POGGIO, T. (1995). Regularization theory and neural networks architectures. *Neural Comput.* **7** 219–269.

[32] GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression. Springer Series in Statistics*. Springer, New York. MR1920390 https://doi.org/10.1007/b97848

[33] HASTIE, T., MONTANARI, A., ROSSET, S. and TIBSHIRANI, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. Available at arXiv:1903.08560.

[34] HORNIK, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4** 251–257.

[35] JACOT, A., GABRIEL, F. and HONGLER, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems* 8571–8580.

[36] LEE, J., XIAO, L., SCHOENHOLZ, S., BAHRI, Y., NOVAK, R., SOHL-DICKSTEIN, J. and PENNINGTON, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems* 8570–8581.

[37] LIANG, T. and RAKHLIN, A. (2020). Just interpolate: Kernel "Ridgeless" regression can generalize. *Ann. Statist.* **48** 1329–1347. MR4124325 https://doi.org/10.1214/19-AOS1849

[38] LIANG, T., RAKHLIN, A. and ZHAI, X. (2019). On the risk of minimum-norm interpolants and restricted lower isometry of kernels. Available at arXiv:1908.10292.

[39] MAIOROV, V. E. (1999). On best approximation by ridge functions. *J. Approx. Theory* **99** 68–94. MR1696577 https://doi.org/10.1006/jath.1998.3304

[40] MAIOROV, V. E. and MEIR, R. (2000). On the near optimality of the stochastic approximation of smooth functions by neural networks. *Adv. Comput. Math.* **13** 79–103. MR1759189 https://doi.org/10.1023/A:1018993908478

[41] MEI, S., BAI, Y. and MONTANARI, A. (2018). The landscape of empirical risk for nonconvex losses. *Ann. Statist.* **46** 2747–2774. MR3851754 https://doi.org/10.1214/17-AOS1637

[42] MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2019). Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit. In *Conference on Learning Theory* 2388–2464.

[43] MEI, S. and MONTANARI, A. (2019). The generalization error of random features regression: Precise asymptotics and double descent curve. Available at arXiv:1908.05355.

[44] 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. MR3845070 https://doi.org/10.1073/pnas.1806579115

[45] MHASKAR, H. N. (1996). Neural networks for optimal approximation of smooth and analytic functions. *Neural Comput.* **8** 164–177.

[46] MHASKAR, H. N. and MICCHELLI, C. A. (1994). Dimension-independent bounds on the degree of approximation by neural networks. *IBM J. Res. Develop.* **38** 277–284.

[47] NEAL, R. M. (1996). Priors for infinite networks. In *Bayesian Learning for Neural Networks* 29–53. Springer, Berlin.

[48] PETRUSHEV, P. P. (1999). Approximation by ridge functions and neural networks. *SIAM J. Math. Anal.* **30** 155–189. MR1646689 https://doi.org/10.1137/S0036141097322959

[49] PINKUS, A. (1999). Approximation theory of the MLP model in neural networks. In *Acta Numerica*, 1999. *Acta Numer.* **8** 143–195. Cambridge Univ. Press, Cambridge. MR1819645 https://doi.org/10.1017/S0962492900002919

[50] RAHIMI, A. and RECHT, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems* 1177–1184.

[51] ROTSKOFF, G. M. and VANDEN-EIJNDEN, E. (2018). Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. Available at arXiv:1805.00915.

[52] RUDI, A. and ROSASCO, L. (2017). Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems* 3215–3225.

[53] SIRIGNANO, J. and SPILIOPOULOS, K. (2020). Mean field analysis of neural networks: A central limit theorem. *Stochastic Process. Appl.* **130** 1820–1852. MR4058290 https://doi.org/10.1016/j.spa.2019.06.003

[54] SOUDRY, D., HOFFER, E., NACSON, M. S., GUNASEKAR, S. and SREBRO, N. (2018). The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.* **19** Paper No. 70, 57. MR3899772

[55] SZEGÖ, G. (1939). *Orthogonal Polynomials*. Amer. Math. Soc., New York. MR0000077
[56] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794
[57] VEMPALA, S. and WILMES, J. (2018). Polynomial convergence of gradient descent for training one-hidden-layer neural networks. Available at arXiv:1805.02677.
[58] YEHUDAI, G. and SHAMIR, O. (2019). On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems* 6594–6604.
[59] ZOU, D., CAO, Y., ZHOU, D. and GU, Q. (2020). Gradient descent optimizes over-parameterized deep ReLU networks. *Mach. Learn.* **109** 467–492. MR4075425 https://doi.org/10.1007/s10994-019-05839-6