Exact Gap between Generalization Error and Uniform Convergence in Random Feature Models

Zitong Yang[1] Yu Bai[2] Song Mei[3]

Abstract

Recent work showed that there could be a large gap between the classical uniform convergence bound and the actual test error of zero-training-error predictors (interpolators) such as deep neural networks. To better understand this gap, we study the uniform convergence in the nonlinear random feature model and perform a precise theoretical analysis on how uniform convergence depends on the sample size and the number of parameters. We derive and prove analytical expressions for three quantities in this model: 1) classical uniform convergence over norm balls, 2) uniform convergence over interpolators in the norm ball (recently proposed by Zhou et al. (2020)), and 3) the risk of minimum norm interpolator. We show that, in the setting where the classical uniform convergence bound is vacuous (diverges to ∞), uniform convergence over the interpolators still gives a non-trivial bound of the test error of interpolating solutions. We also showcase a different setting where classical uniform convergence bound is non-vacuous, but uniform convergence over interpolators can give an improved sample complexity guarantee. Our result provides a first exact comparison between the test errors and uniform convergence bounds for interpolators beyond simple linear models.

1. Introduction

Uniform convergence—the supremum difference between the training and test errors over a certain function class—is a powerful tool in statistical learning theory for understanding the generalization performance of predictors. Bounds on uniform convergence usually take the form of $\sqrt{\text{complexity}/n}$ (Vapnik, 1995), where the numerator represents the complexity of the function class, and n is the sample size. If such a bound is tight, then the predictor is not going to generalize well whenever the function class complexity is too large.

However, it is shown in recent theoretical and empirical work that overparametized models such as deep neural networks could generalize well, even in the interpolating regime in which the model exactly memorizes the data (Zhang et al., 2016; Belkin et al., 2019a). As interpolation (especially for noisy training data) usually requires the predictor to be within a function class with high complexity, this challenges the classical methodology of using uniform convergence to bound generalization. For example, Belkin et al. (2018c) showed that interpolating noisy data with kernel machines requires exponentially large norm in fixed dimensions. The large norm would effectively make the uniform convergence bound $\sqrt{\text{complexity}/n}$ vacuous. Nagarajan & Kolter (2019a) empirically measured the spectral-norm bound in Bartlett et al. (2017) and find that for interpolators, the bound increases with n, and is thus vacuous at large sample size. Towards a more fine-grained understanding, we ask the following

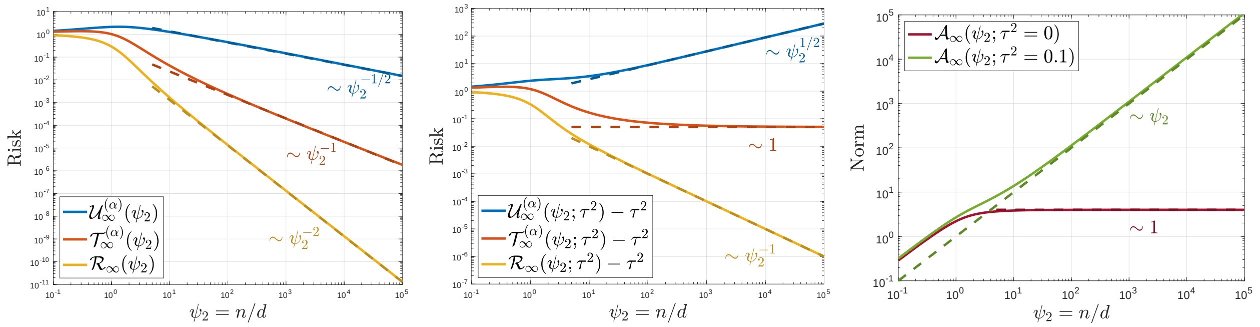
> **Question**: How large is the gap between uniform convergence and the actual generalization errors for interpolators?

In this paper, we study this gap in the random features model from Rahimi & Recht (2007). This model can be interpreted as a linearized version of two-layer neural networks (Jacot et al., 2018) and exhibit some similar properties to deep neural networks such as double descent (Belkin et al., 2019a). We consider two types of uniform convergence in this model:

- $\mathcal{U}$: The classical uniform convergence over a norm ball of radius $\sqrt{A}$.
- $\mathcal{T}$: The modified uniform convergence over the same norm ball of size $\sqrt{A}$ but only include the interpolators, proposed in Zhou et al. (2020).

[1] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. [2] Salesforce Research. [3] Department of Statistics, University of California, Berkeley. Correspondence to: Zitong Yang <zitong@berkeley.edu>, Yu Bai <yu.bai@salesforce.com>, Song Mei <songmei@berkeley.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).



(a) Noiseless response ($\tau^2 = 0$) (b) Noisy response ($\tau^2 = 0.1$) (c) Minimum norm $\mathcal{A}_\infty(\psi_2)$

Figure 1. Random feature regression with activation function $\sigma(x) = \max(0,x) - 1/\sqrt{2\pi}$, target function $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle$ with $\|\boldsymbol{\beta}\|_2^2 = 1$, and $\psi_1 = \infty$. The horizontal axes are the number of samples $\psi_2 = \lim_{d\to\infty} n/d$. The solid lines are the the algebraic expressions derived in the main theorem (Theorem 1). The dashed lines are the function ψ_2^p in the log scale. Figure 1(a) and 1(b): Comparison of the classical uniform convergence in the norm ball of size level $\alpha = 1.5$ (Eq. (17), blue curve), the uniform convergence over interpolators in the same norm ball (Eq. (18), red curve), the risk of minimum norm interpolator (Eq. (13), yellow curve). Figure 1(c): Minimum norm required to interpolate the training data (Eq. (12)).

Our main theoretical result is the exact asymptotic expressions of two versions of uniform convergence $\mathcal{U}$ and $\mathcal{T}$ in terms of the number of features, sample size, as well as other relevant parameters in the random feature model. Under some assumptions, we prove that the actual uniform convergence concentrates to these asymptotic counterparts. To further compare these uniform convergence bounds with the actual generalization error of interpolators, we adopt

- $\mathcal{R}$: the generalization error (test error) of the minimum norm interpolator.

from Mei & Montanari (2019). To make $\mathcal{U}$, $\mathcal{T}$, $\mathcal{R}$ comparable with each other, we choose the radius of the norm ball $\sqrt{A}$ to be slightly larger than the norm of the minimum norm interpolator. Our limiting $\mathcal{U}$, $\mathcal{T}$ (with norm ball of size $\sqrt{A}$ as chosen above), and $\mathcal{R}$ depend on two main variables: $\psi_1 = \lim_{d\to\infty} N/d$ representing the number of parameters, and $\psi_2 = \lim_{d\to\infty} n/d$ representing the sample size. Our formulae for $\mathcal{U}$, $\mathcal{T}$ and $\mathcal{R}$ yield three major observations.

1. **Sample Complexity in the Noisy Regime:** When the training data contains label noise (with variance τ^2), we find that the norm required to interpolate the noisy training set grows linearly with the number of samples ψ_2 (green curve in Figure 1(c)). As a result, the standard uniform convergence bound $\mathcal{U}$ grows with ψ_2 at the rate $\mathcal{U} \sim \psi_2^{1/2}$, leading to a vacuous bound on the generalization error (Figure 1(b)).

   In contrast, in the same setting, we show the uniform convergence over interpolators $\mathcal{T} \sim 1$ is a constant for large ψ_2, and is only order one larger than the actual generalization error $\mathcal{R} \sim 1$. Further, the excess versions scale as $\mathcal{T} - \tau^2 \sim 1$ and $\mathcal{R} - \tau^2 \sim \psi_2^{-1}$.

2. **Sample Complexity in the Noiseless Regime:** When the training set does not contain label noise, the generalization error $\mathcal{R}$ decays faster: $\mathcal{R} \sim \psi_2^{-2}$. In this setting, we find that the classical uniform convergence $\mathcal{U} \sim \psi_2^{-1/2}$ and the uniform convergence over interpolators $\mathcal{T} \sim \psi_2^{-1}$. This shows that, even when the classical uniform convergence already gives a non-vacuous bound, there still exists a sample complexity separation among the classical uniform convergence $\mathcal{U}$, the uniform convergence over interpolators $\mathcal{T}$, and the actual generalization error $\mathcal{R}$.

3. **Dependence on Number of Parameters:** In addition to the results on ψ_2, we find that $\mathcal{U}$, $\mathcal{T}$ and $\mathcal{R}$ decay to its limiting value at the same rate $1/\psi_1$. This shows that both $\mathcal{U}$ and $\mathcal{T}$ correctly predict that as the number of features ψ_1 grows, the risk $\mathcal{R}$ would decrease.

These results provide a more precise understanding of uniform convergence versus the actual generalization errors, under a natural model that captures a lot of essences of nonlinear overparametrized learning.

1.1. Related work

Classical theory of uniform convergence. Uniform convergence dates back to the empirical process theory of Glivenko (1933) and Cantelli (1933). Application of uniform convergence to the framework of empirical risk minimization usually proceeds through Gaussian and Rademacher complexities (Bartlett & Mendelson, 2003; Bartlett et al., 2005) or VC and fat shattering dimensions (Vapnik, 1995; Bartlett, 1998).

Modern take on uniform convergence. A large volume of recent works showed that overparametrized interpolators could generalize well (Zhang et al., 2016; Belkin et al., 2018b; Neyshabur et al., 2015a; Advani et al., 2020; Bartlett et al., 2020; Belkin et al., 2018a; 2019b; Nakkiran et al., 2020; Yang et al., 2020; Belkin et al., 2019a; Mei & Montanari, 2019; Spigler et al., 2019), suggesting that the classical uniform convergence theory may not be able to explain generalization in these settings (Zhang et al., 2016). Numerous efforts have been made to remedy the original uniform convergence theory using the Rademacher complexity (Neyshabur et al., 2015b; Golowich et al., 2018; Neyshabur et al., 2019; Zhu et al., 2009; Cao & Gu, 2019), the compression approach (Arora et al., 2018), covering numbers (Bartlett et al., 2017), derandomization (Negrea et al., 2020) and PAC-Bayes methods (Dziugaite & Roy, 2017; Neyshabur et al., 2018; Nagarajan & Kolter, 2019b). Despite the progress along this line, Nagarajan & Kolter (2019a); Bartlett & Long (2020) showed that in certain settings "any uniform convergence" bounds cannot explain generalization. Among the pessimistic results, Zhou et al. (2020) proposes that uniform convergence over interpolating norm ball could explain generalization in an overparametrized linear setting. Our results show that in the nonlinear random feature model, there is a sample complexity gap between the excess risk and uniform convergence over interpolators proposed in Zhou et al. (2020).

Random features model and kernel machines. A number of papers studied the generalization error of kernel machines (Caponnetto & De Vito, 2007; Jacot et al., 2020b; Wainwright, 2019) and random features models (Rahimi & Recht, 2009; Rudi & Rosasco, 2017; Bach, 2015; Ma et al., 2020) in the non-asymptotic settings, in which the generalization error bound depends on the RKHS norm. However, these bounds cannot characterize the generalization error for interpolating solutions. In the last three years, a few papers (Belkin et al., 2018c; Liang et al., 2020; 2019) showed that interpolating solutions of kernel ridge regression can also generalize well in high dimensions. Recently, a few papers studied the generalization error of random features model in the proportional asymptotic limit in various settings (Hastie et al., 2019; Louart et al., 2018; Mei & Montanari, 2019; Montanari et al., 2019; Gerace et al., 2020; d'Ascoli et al., 2020; Yang et al., 2020; Adlam & Pennington, 2020; Dhifallah & Lu, 2020; Hu & Lu, 2020), where they precisely characterized the asymptotic generalization error of interpolating solutions, and showed that double-descent phenomenon (Belkin et al., 2019a; Advani et al., 2020) exists in these models. A few other papers studied the generalization error of random features models in the polynomial scaling limits (Ghorbani et al., 2019; 2020; Mei et al., 2021), where other interesting behaviors were shown.

Precise asymptotics for the Rademacher complexity of some *underparameterized* learning models was calculated using statistical physics heuristics in Abbaras et al. (2020). In our work, we instead focus on the uniform convergence of *overparameterized* random features model.

2. Problem formulation

In this section, we present the background needed to understand the insights from our main result. In Section 2.1 we define the random feature regression task that this paper focuses on. In Section 2.2, we informally present the limiting regime our theory covers.

2.1. Model setup

Consider a dataset $(\boldsymbol{x}_i, y_i)_{i\in[n]}$ with n samples. Assume that the covariates follow $\boldsymbol{x}_i \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, and responses satisfy $y_i = f_d(\boldsymbol{x}_i)+\varepsilon_i$, with the noises satisfying $\varepsilon_i \sim_{iid} \mathcal{N}(0,\tau^2)$ which are independent of $(\boldsymbol{x}_i)_{i\in[n]}$. We will consider both the noisy ($\tau^2 > 0$) and noiseless ($\tau^2 = 0$) settings.

We fit the dataset using the random features model. Let $(\boldsymbol{\theta}_j)_{j\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ be the random feature vectors. Given an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, we define the random features function class $\mathcal{F}_{\mathrm{RF}}(\boldsymbol{\Theta})$ by

$$\mathcal{F}_{\mathrm{RF}}(\boldsymbol{\Theta}) \equiv \Big\{ f(\boldsymbol{x}) = \sum_{j=1}^{N} a_j \sigma\big(\langle \boldsymbol{x}, \boldsymbol{\theta}_j\rangle/\sqrt{d}\big) : \boldsymbol{a} \in \mathbb{R}^N \Big\}.$$

Generalization error of the minimum norm interpolator. Denote the population risk and the empirical risk of a predictor $\boldsymbol{a} \in \mathbb{R}^N$ by

$$R(\boldsymbol{a}) = \mathbb{E}_{\boldsymbol{x},y}\Big(y - \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{x}, \boldsymbol{\theta}_j\rangle/\sqrt{d})\Big)^2, \tag{1}$$

$$\widehat{R}_n(\boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j\rangle/\sqrt{d})\Big)^2, \tag{2}$$

and the regularized empirical risk minimizer with vanishing regularization by

$$\boldsymbol{a}_{\min} = \lim_{\lambda\to 0+} \arg\min_{\boldsymbol{a}} \Big[\widehat{R}_n(\boldsymbol{a}) + \lambda\|\boldsymbol{a}\|_2^2\Big].$$

In the overparameterized regime ($N > n$), under mild conditions, we have $\min_{\boldsymbol{a}} \widehat{R}_n(\boldsymbol{a}) = \widehat{R}_n(\boldsymbol{a}_{\min}) = 0$. In this regime, $\boldsymbol{a}_{\min}$ can be interpreted as the minimum ℓ_2 norm interpolator.

A quantity of interest is the generalization error of this predictor, which gives (with a slight abuse of notation)

$$R(N,n,d) \equiv R(\boldsymbol{a}_{\min}). \tag{3}$$

Uniform convergence bounds. We denote the uniform convergence bound over a norm ball and the uniform convergence over interpolators in the norm ball by

$$U(A, N, n, d) \equiv \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A} \Big(R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) \Big), \quad (4)$$

$$T(A, N, n, d) \equiv \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A, \widehat{R}_n(\boldsymbol{a})=0} R(\boldsymbol{a}). \quad (5)$$

Here the scaling factor N/d of the norm ball is such that the norm ball converges to a non-trivial RKHS norm ball with size $\sqrt{A}$ as $\psi_1 \to \infty$ (limit taken after $N/d \to \psi_1$). Note that in order for the maximization problem in (5) to have a non-empty feasible region, we need $\widehat{R}_n(\boldsymbol{a}_{\min}) = 0$ and need to take $A \ge (N/d)\|\boldsymbol{a}_{\min}\|_2^2$: we will show that in the region $N > n$ with sufficiently large A, this happens with high probability.

By construction, for any $A \ge (N/d)\|\boldsymbol{a}_{\min}\|_2^2$, we have $U(A) \ge T(A) \ge R(\boldsymbol{a}_{\min})$ (see Figuire 2). So a natural problem is to quantify the gap among $U(A)$, $T(A)$, and $R(\boldsymbol{a}_{\min})$, which is our goal in this paper.

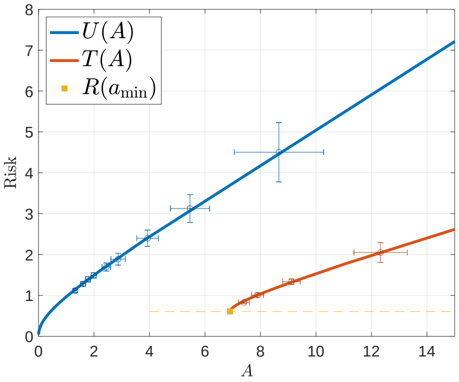


Figure 2. Illustration of uniform convergence U (c.f. eq. (4)), uniform convergence over interpolators T (c.f. eq. (5)), and minimum norm interpolator $R(\boldsymbol{a}_{\min})$. We take $y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle$ for some $\|\boldsymbol{\beta}\|_2^2 = 1$, and take the ReLU activation function $\sigma(x) = \max\{x, 0\}$. Solid lines are our theoretical predictions $\mathcal{U}$ and $\mathcal{T}$ (cf. (6) & (7)). Points with error bars are obtained from simulations with the number of features $N = 500$, number of samples $n = 300$, and covariate dimension $d = 200$. The error bar reports $1/\sqrt{20}\times$standard deviation over 20 instances. See Appendix B for details.

2.2. High dimensional regime

We approach this problem in the limit $d \to \infty$ with $N/d \to \psi_1$ and $n/d \to \psi_2$ (c.f. Assumption 3). We further assume the setting of a linear target function f_d and a nonlinear activation function σ (c.f. Assumptions 1 and 2). In this regime, our main result Theorem 1 will show that, the uniform convergence U and the uniform convergence over interpolators T will converge to deterministic functions, i.e., writing here informally,

$$U(A, N, n, d) \overset{d\to\infty}{\to} \mathcal{U}(A, \psi_1, \psi_2), \quad (6)$$

$$T(A, N, n, d) \overset{d\to\infty}{\to} \mathcal{T}(A, \psi_1, \psi_2), \quad (7)$$

where $\mathcal{U}$ and $\mathcal{T}$ will be defined in Definition 2 (which depends on the definition of some other quantities that are defined in Appendix A and heuristically presented in Remark 1). In addition to $\mathcal{U}$ and $\mathcal{T}$, Theorem 1 of Mei & Montanari (2019) implies the following convergence

$$(N/d)\|\boldsymbol{a}_{\min}\|_2^2 \overset{d\to\infty}{\to} \mathcal{A}(\psi_1, \psi_2), \quad (8)$$

$$R(\boldsymbol{a}_{\min}) \overset{d\to\infty}{\to} \mathcal{R}(\psi_1, \psi_2). \quad (9)$$

The precise algebraic expression of equation (8) and (9) was given in Definition 1 of Mei & Montanari (2019), and we include in Appendix A for completeness. We will sometimes refer to $\mathcal{U}, \mathcal{T}, \mathcal{A}, \mathcal{R}$ without explicitly mark their dependence on A, ψ_1, ψ_2 for notational simplicity.

Kernel regime. Rahimi & Recht (2007) have shown that, as $N \to \infty$, the random feature space $\mathcal{F}_{\rm RF}(\boldsymbol{\Theta})$ (equipped with proper inner product) converges to the RKHS (Reproducing Kernel Hilbert Space) induced by the kernel

$$H(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{w} \sim {\rm Unif}(\mathbb{S}^{d-1})}[\sigma(\langle \boldsymbol{x}, \boldsymbol{w} \rangle)\sigma(\langle \boldsymbol{x}', \boldsymbol{w} \rangle)].$$

We expect that, if we take limit $\psi_1 \to \infty$ after $N, d, n \to \infty$, the formula of $\mathcal{U}$ and $\mathcal{T}$ will coincide with the corresponding asymptotic limit of U and T for kernel ridge regression with the kernel H. This intuition has been mentioned in a few papers (Mei & Montanari, 2019; d'Ascoli et al., 2020; Jacot et al., 2020a). In this spirit, we denote

$$\mathcal{U}_\infty(A, \psi_2) \equiv \lim_{\psi_1 \to \infty} \mathcal{U}(A, \psi_1, \psi_2), \quad (10)$$

$$\mathcal{T}_\infty(A, \psi_2) \equiv \lim_{\psi_1 \to \infty} \mathcal{T}(A, \psi_1, \psi_2), \quad (11)$$

$$\mathcal{A}_\infty(\psi_2) \equiv \lim_{\psi_1 \to \infty} \mathcal{A}(\psi_1, \psi_2), \quad (12)$$

$$\mathcal{R}_\infty(\psi_2) \equiv \lim_{\psi_1 \to \infty} \mathcal{R}(\psi_1, \psi_2). \quad (13)$$

We will refer to the quantities $\{\mathcal{U}_\infty, \mathcal{T}_\infty, \mathcal{A}_\infty, \mathcal{R}_\infty\}$ as the {uniform convergence in norm ball, uniform convergence over interpolators in norm ball, minimum ℓ_2 norm of interpolators, and generalization error of interpolators} of kernel ridge regression.

Low norm uniform convergence bounds. There is a question of which norm A to choose in $\mathcal{U}$ and $\mathcal{T}$ to compare with $\mathcal{R}$. In order for U and T to serve as proper bounds for $R(\boldsymbol{a}_{\min})$, we need to take at least $A \ge \psi_1\|\boldsymbol{a}_{\min}\|_2^2$. Therefore, we will choose

$$A = \alpha\psi_1\|\boldsymbol{a}_{\min}\|_2^2, \quad (14)$$

for some $\alpha > 1$ (e.g., $\alpha = 1.1$). Note $\psi_1 \|\boldsymbol{a}_{\min}\|_2^2 \to \mathcal{A}(\psi_1, \psi_2)$ as $d \to \infty$. So for a fixed $\alpha > 1$, we further define

$$\mathcal{U}^{(\alpha)}(\psi_1, \psi_2) \equiv \mathcal{U}(\alpha\mathcal{A}(\psi_1, \psi_2), \psi_1, \psi_2), \tag{15}$$

$$\mathcal{T}^{(\alpha)}(\psi_1, \psi_2) \equiv \mathcal{T}(\alpha\mathcal{A}(\psi_1, \psi_2), \psi_1, \psi_2), \tag{16}$$

and their kernel version,

$$\mathcal{U}_\infty^{(\alpha)}(\psi_2) \equiv \lim_{\psi_1 \to \infty} \mathcal{U}^{(\alpha)}(\psi_1, \psi_2), \tag{17}$$

$$\mathcal{T}_\infty^{(\alpha)}(\psi_2) \equiv \lim_{\psi_1 \to \infty} \mathcal{T}^{(\alpha)}(\psi_1, \psi_2). \tag{18}$$

This definition ensures that $\mathcal{R}(\psi_1, \psi_2) \le \mathcal{T}^{(\alpha)}(\psi_1, \psi_2) \le \mathcal{U}^{(\alpha)}(\psi_1, \psi_2)$ and $\mathcal{R}_\infty(\psi_2) \le \mathcal{T}_\infty^{(\alpha)}(\psi_2) \le \mathcal{U}_\infty^{(\alpha)}(\psi_2)$.

3. Asymptotic power laws and separations

In this section, we evaluate the algebraic expressions derived in our main result (Theorem 1) as well as the quantities $\mathcal{U}^{(\alpha)}$, $\mathcal{T}^{(\alpha)}$, $\mathcal{A}$, and $\mathcal{R}$, before formally presenting the theorem. We examine their dependence with respect to the noise level τ^2, the number of features $\psi_1 = \lim_{d\to\infty} N/d$, and the sample size $\psi_2 = \lim_{d\to\infty} n/d$, and we further infer their asymptotic power laws for large ψ_1 and ψ_2.

3.1. Norm of the minimum norm interpolator

Since we are considering uniform convergence bounds over the norm ball of size α times $\mathcal{A}_\infty(\psi_2)$ (the norm of the min-norm interpolator), let's first examine how $\mathcal{A}_\infty(\psi_2)$ scale with ψ_2. As we shall see, $\mathcal{A}_\infty(\psi_2)$ behaves differently in the noiseless ($\tau^2 = 0$) and noisy ($\tau^2 > 0$) settings, so here we explicitly mark the dependence on τ^2, i.e. $\mathcal{A}_\infty(\psi_2; \tau^2)$.

The inferred asymptotic power law gives (c.f. Figure 1(c))

$$\mathcal{A}_\infty(\psi_2; \tau^2 > 0) \sim \psi_2,$$
$$\mathcal{A}_\infty(\psi_2; \tau^2 = 0) \sim 1,$$

where $X_1(\psi) \sim X_2(\psi)$ for large ψ means that

$$\lim_{\psi\to\infty} \log(X_1(\psi))/\log(X_2(\psi)) = 1.$$

In words, when there is no label noise ($\tau^2 = 0$), we can interpolate infinite data even with a finite norm. When the responses are noisy ($\tau^2 > 0$), interpolation requires a large norm that is proportional to the number of samples.

On a high level, our statement echoes the finding of Belkin et al. (2018c), where they study a binary classification problem using the kernel machine, and prove that an interpolating classifier requires RKHS norm to grow at least exponentially with $n^{1/d}$ for fixed dimension d. Here instead we consider the high dimensional setting and we show a linear grow in $\psi_2 = \lim_{d\to\infty} n/d$.

3.2. Kernel regime with noiseless data

We first look at the noiseless setting ($\tau^2 = 0$) and present the asymptotic power law for the uniform convergence $\mathcal{U}_\infty^{(\alpha)}$ over the low-norm ball, the uniform convergence over interpolators $\mathcal{T}_\infty^{(\alpha)}$ in the low norm ball, and the minimum norm risk $\mathcal{R}_\infty$ from (17) (18) (13), respectively.

In this setting, the inferred asymptotic power law of $\mathcal{U}_\infty^{(\alpha)}(\psi_2)$, $\mathcal{T}_\infty^{(\alpha)}(\psi_2)$, and $\mathcal{R}_\infty(\psi_2)$ gives (c.f. Figure 1(a))

$$\mathcal{U}_\infty^{(\alpha)}(\psi_2; \tau^2 = 0) \sim \psi_2^{-1/2},$$
$$\mathcal{T}_\infty^{(\alpha)}(\psi_2; \tau^2 = 0) \sim \psi_2^{-1},$$
$$\mathcal{R}_\infty^{(\alpha)}(\psi_2; \tau^2 = 0) \sim \psi_2^{-2}.$$

As we can see, all the three quantities converge to 0 in the large sample limit, which indicates that uniform convergence is able to explain generalization in this setting. yet uniform convergence bounds do not correctly capture the convergence rate (in terms of ψ_2) of the generalization error.

3.3. Kernel regime with noisy data

In the noisy setting (fix $\tau^2 > 0$), the Bayes risk (minimal possible risk) is τ^2. We study the excess risk and the excess version of uniform convergence bounds by subtracting the Bayes risk τ^2. The inferred asymptotic power law gives (c.f. Figure 1(b))

$$\mathcal{U}_\infty^{(\alpha)}(\psi_2; \tau^2) - \tau^2 \sim \psi_2^{1/2},$$
$$\mathcal{T}_\infty^{(\alpha)}(\psi_2; \tau^2) - \tau^2 \sim 1,$$
$$\mathcal{R}_\infty(\psi_2; \tau^2) - \tau^2 \sim \psi_2^{-1}.$$

In the presence of label noise, the excess risk $\mathcal{R}_\infty - \tau^2$ vanishes in the large sample limit. In contrast, the classical uniform convergence $\mathcal{U}_\infty$ becomes vacuous, whereas the uniform convergence over interpolators $\mathcal{T}_\infty$ converges to a constant, which gives a non-vacuous bound of $\mathcal{R}_\infty$.

The decay of the excess risk of minimum norm interpolators even in the presence of label noise is no longer a surprising phenomenon in high dimensions (Liang et al., 2019; Ghorbani et al., 2019; Bartlett et al., 2020). A simple explanation of this phenomenon is that the nonlinear part of the activation function σ has an implicit regularization effect (Mei & Montanari, 2019).

The divergence of $\mathcal{U}_\infty^{(\alpha)}$ in the presence of response noise is partly due to that $\mathcal{A}_\infty(\psi_2)$ blows up linearly in ψ_2 (c.f. Section 3.1). In fact, we can develop a heuristic intuition that $\mathcal{U}_\infty(A, \psi_2; \tau^2) \sim A/\psi_2^{1/2}$. Then the scaling $\mathcal{U}_\infty^{(\alpha)}(\psi_2; \tau^2 > 0) \sim \mathcal{A}_\infty(\psi_2; \tau^2 > 0)/\psi_2^{1/2} \sim \psi_2^{1/2}$ can be explained away by the power law of $\mathcal{A}_\infty(\psi_2; \tau^2 > 0) \sim \psi_2$. In other words, the complexity of the function space of interpolators grows faster than the sample size n, which leads to the failure of uniform convergence in explaining generalization. This echoes the findings in Nagarajan & Kolter (2019a).

To illustrate the scaling $\mathcal{U}_\infty(A, \psi_2) \sim A/\psi_2^{1/2}$. We fix all other parameters $(\mu_1, \mu_\star, \tau, F_1)$, and examine the dependence of $\mathcal{U}_\infty$ on A and ψ_2. We choose $A = A(\psi_2)$ according to different power laws $A(\psi_2) \sim \psi_2^p$ for $p = 0, 0.25, 0.5, 0.75, 1$. The inferred asymptotic power law gives $\mathcal{U}_\infty(A(\psi_2), \psi_2) \sim \psi_2^{p-0.5}$ (c.f. Figure 3). This provides an evidence for the relation $\mathcal{U}_\infty(A, \psi_2) \sim A/\psi_2^{1/2}$.

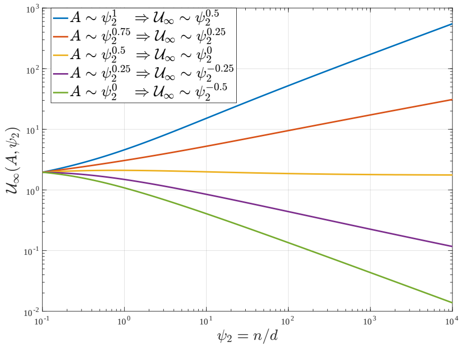


Figure 3. Uniform convergence $\mathcal{U}_\infty(A(\psi_2), \psi_2)$ over the norm ball in the kernel regime $\psi_1 \to \infty$. The size of the norm ball $A = A(\psi_2)$ is chosen according to different power laws as shown in the legend.

3.4. Finite-width regime

Here we shift attention to the dependence of $\mathcal{U}$, $\mathcal{T}$, and $\mathcal{R}$ on the number of features ψ_1. We fix the number of training samples ψ_2, noise level $\tau^2 > 0$, and norm level $\alpha > 1$ similar as before. Since $\mathcal{U}^\alpha \to \mathcal{U}_\infty^\alpha, \mathcal{T}^\alpha \to \mathcal{T}_\infty^\alpha$ and $\mathcal{R} \to \mathcal{R}_\infty$ as $\psi_1 \to \infty$, we look at the dependence of $\mathcal{U}^\alpha - \mathcal{U}_\infty^\alpha, \mathcal{T}^\alpha - \mathcal{T}_\infty^\alpha$ and $\mathcal{R}^\alpha - \mathcal{R}_\infty^\alpha$ with respect to ψ_1. The inferred asymptotic law gives (c.f. Figure 4)

$$
\begin{aligned}
\mathcal{U}^{(\alpha)}(\psi_1, \psi_2) - \mathcal{U}_\infty^{(\alpha)}(\psi_2) &\sim \psi_1^{-1}, \\
\mathcal{T}^{(\alpha)}(\psi_1, \psi_2) - \mathcal{T}_\infty^{(\alpha)}(\psi_2) &\sim \psi_1^{-1}, \\
\mathcal{R}(\psi_1, \psi_2) - \mathcal{R}_\infty(\psi_2) &\sim \psi_1^{-1}, \\
\mathcal{A}(\psi_1, \psi_2) - \mathcal{A}_\infty(\psi_2) &\sim \psi_1^{-1}.
\end{aligned}
$$

Note that large ψ_1 should be interpreted as the model being heavily overparametrized (a large width network). This asymptotic power law implies that, both uniform convergence bounds correctly predict the decay of the test error with the increase of the number of features.

Remark on power laws. For the derivation of the power laws in this section, instead of working with the analytical formula, we adopt an empirical approach: we perform linear fits with the inferred slopes, upon the numerical evaluations (of these expressions defined in Definition 2) in the log-log scale. However, these linear fits are for the analytical formulae and do not involve randomness, and thus reliably indicate the true decay rates.

4. Main theorem

In this section, we state the main theorem that presents the asymptotic expressions for the uniform convergence bounds. We will start by stating a few assumptions, which fall into two categories: Assumption 1, 2, and 3, which specify the setup for the learning task; Assumption 4 and 5, which are technical in nature.

4.1. Modeling assumptions

The three assumptions in this subsection specify the target function, the activation function, and the limiting regime.

Assumption 1 (Linear target function)**.** *We assume that $f_d \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ with $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}^{(d)}, \boldsymbol{x} \rangle$, where $\boldsymbol{\beta}^{(d)} \in \mathbb{R}^d$ and*

$$
\lim_{d \to \infty} \|\boldsymbol{\beta}^{(d)}\|_2^2 = F_1^2.
$$

We remark here that, if we are satisfied with heuristic formulae instead of rigorous results, we are able to deal with non-linear target functions, where the additional nonlinear part is effectively increasing the noise level τ^2. This intuition was first developed in (Mei & Montanari, 2019).

Assumption 2 (Activation function)**.** *Let $\sigma \in C^2(\mathbb{R})$ with $|\sigma(u)|, |\sigma'(u)|, |\sigma''(u)| \le c_0 e^{c_1|u|}$ for some constant $c_0, c_1 < \infty$. Define*

$$
\mu_0 \equiv \mathbb{E}[\sigma(G)],\ \mu_1 \equiv \mathbb{E}[G\sigma(G)],\ \mu_\star^2 \equiv \mathbb{E}[\sigma(G)^2] - \mu_0^2 - \mu_1^2,
$$

where expectation is with respect to $G \sim \mathcal{N}(0, 1)$. Assume $\mu_0 = 0$, $0 < \mu_1^2, \mu_\star^2 < \infty$.

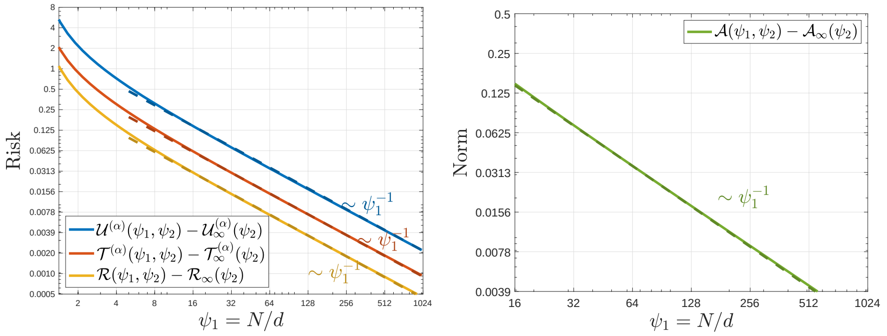
The assumption that $\mu_0 = 0$ is not essential and can be relaxed with a certain amount of additional technical work.

Assumption 3 (Proportional limit)**.** *Let $N = N(d)$ and $n = n(d)$ be sequences indexed by d. We assume that the following limits exist in $(0, \infty)$:*

$$
\lim_{d \to \infty} N(d)/d = \psi_1, \qquad \lim_{d \to \infty} n(d)/d = \psi_2.
$$

4.2. Technical assumptions

We will make some assumptions upon the properties of some random matrices that appear in the proof. These assumptions are technical and we believe they can be proved under more natural assumptions. However, proving them requires substantial technical work, and we defer them to future work. We note here that these assumptions are often implicitly required in papers that present intuitions using



(a)

(b)

Figure 4. Random feature regression with the number of sample $\psi_2 = 1.5$, activation function $\sigma(x) = \max(0, x) - 1/\sqrt{2\pi}$, target function $f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x}\rangle$ with $\|\boldsymbol{\beta}\|_2^2 = 1$, and noise level $\tau^2 = 0.1$. The horizontal axes are the number of features ψ_1. The solid lines are the the algebraic expressions derived in the main theorem (Theorem 1). The dashed lines are the function ψ_1^p in the log scale. Figure 4(a): Comparison of the classical uniform convergence in the norm ball of size level $\alpha = 1.5$ (Eq. (15), blue curve), the uniform convergence over interpolators in the same norm ball (Eq. (16), red curve), the risk of minimum norm interpolator (Eq. (9), yellow curve). Figure 4(b): Minimum norm required to interpolate the training data (Eq. (8)).

heuristic derivations. Instead, we ensure the mathematical rigor by listing them. See Section 5 for more discussions upon these assumptions.

We begin by defining some random matrices which are the key quantities that are used in the proof of our main results.

Definition 1 (Block matrix and log-determinant)**.** *Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\mathsf{T} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)^\mathsf{T} \in \mathbb{R}^{N\times d}$, where $\boldsymbol{x}_i, \boldsymbol{\theta}_a \sim_{iid}$ $\mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, as mentioned in Section 2.1. Define*

$$\boldsymbol{Z} = \frac{1}{\sqrt{d}}\sigma\left(\frac{\boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T}}{\sqrt{d}}\right), \quad \boldsymbol{Z}_1 = \frac{\mu_1}{d}\boldsymbol{X}\boldsymbol{\Theta}^\mathsf{T}, \\ \boldsymbol{Q} = \frac{\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}}{d}, \qquad \boldsymbol{H} = \frac{\boldsymbol{X}\boldsymbol{X}^\mathsf{T}}{d}, \tag{19}$$

and for $\boldsymbol{q} = (s_1, s_2, t_1, t_2, q) \in \mathbb{R}^5$, we define

$$\boldsymbol{A}(\boldsymbol{q}) \equiv \begin{bmatrix} s_1\mathbf{I}_N + s_2\boldsymbol{Q} & \boldsymbol{Z}^\mathsf{T} + p\boldsymbol{Z}_1^\mathsf{T} \\ \boldsymbol{Z} + p\boldsymbol{Z}_1 & t_1\mathbf{I}_n + t_2\boldsymbol{H} \end{bmatrix}.$$

Finally, we define the log-deteminant of $\boldsymbol{A}(\boldsymbol{q})$ by

$$G_d(\xi; \boldsymbol{q}) \equiv \frac{1}{d}\sum_{i=1}^{N+n} \mathrm{Log}\lambda_i\Big(\boldsymbol{A}(\boldsymbol{q}) - \xi\mathbf{I}_{n+N}\Big).$$

Here Log *is the complex logarithm with branch cut on the negative real axis and $\{\lambda_i(\boldsymbol{A})\}_{i\in[n+N]}$ is the set of eigenvalues of $\boldsymbol{A}$.*

The following assumption states that for properly chosen λ, some specific random matrices are well-conditioned. As we will see in the next section, this ensures that the dual problems in Eq. (20) and (21) are bounded with high probability.

Assumption 4 (Invertability)**.** *Consider the asymptotic limit as specified in Assumption 3 the activation function as in Assumption 2. We assume the following.*

- *Denote $\overline{\boldsymbol{U}}(\lambda) = \mu_1^2\boldsymbol{Q} + (\mu_\star^2 - \psi_1\lambda)\mathbf{I}_N - \psi_2^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$. There exists $\varepsilon > 0$ and $\underline{\lambda}_U = \underline{\lambda}_U(\psi_1, \psi_2, \mu_1^2, \mu_\star^2)$, such that for any fixed $\lambda \in (\underline{\lambda}_U, \infty) \equiv \Lambda_U$, with high probability, we have*
  $$\overline{\boldsymbol{U}}(\lambda) \preceq -\varepsilon\mathbf{I}_N.$$
- *Denote $\overline{\boldsymbol{T}}(\lambda) = \mathsf{P}_{\mathrm{null}}[\mu_1^2\boldsymbol{Q} + (\mu_\star^2 - \psi_1\lambda)\mathbf{I}_N]\mathsf{P}_{\mathrm{null}}$ where $\mathsf{P}_{\mathrm{null}} = \mathbf{I}_N - \boldsymbol{Z}^\dagger\boldsymbol{Z}$. There exists $\varepsilon > 0$ and $\underline{\lambda}_T = \underline{\lambda}_T(\psi_1, \psi_2, \mu_1^2, \mu_\star^2)$, such that for any fixed $\lambda \in (\underline{\lambda}_T, \infty) \equiv \Lambda_T$, with high probability we have*
  $$\overline{\boldsymbol{T}}(\lambda) \preceq -\varepsilon\mathsf{P}_{\mathrm{null}},$$
  and $\boldsymbol{Z}$ has full row rank with $\sigma_{\min}(\boldsymbol{Z}) \geq \varepsilon$ (which requires $\psi_1 > \psi_2$).

The following assumption states that the order of limits and derivatives regarding G_d can be exchanged.

Assumption 5 (Exchangeability of limits)**.** *We denote*

$$\mathcal{S}_U = \{(\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0; \psi_1, \psi_2) : \lambda \in (\underline{\lambda}_U, \infty)\}, \\ \mathcal{S}_T = \{(\mu_\star^2 - \lambda\psi_1, \mu_1^2, 0, 0, 0; \psi_1, \psi_2) : \lambda \in (\underline{\lambda}_T, \infty)\},$$

where $\underline{\lambda}_U$ and $\underline{\lambda}_T$ are given in Assumption 4 and depend on $(\psi_1, \psi_2, \mu_1^2, \mu_\star^2)$. For any fixed $(\boldsymbol{q}; \boldsymbol{\psi}) = (s_1, s_2, t_1, t_2, p; \psi_1, \psi_2) \in \mathcal{S}_U \cup \mathcal{S}_T$, in the asymptotic limit as in Assumption 3, for $k = 1, 2$, we have

$$\lim_{u\to 0_+}\lim_{d\to\infty} \mathbb{E}[\nabla_{\boldsymbol{q}}^k G_d(\boldsymbol{i}u; \boldsymbol{q})] = \\ \lim_{u\to 0_+} \nabla_{\boldsymbol{q}}^k\Big(\lim_{d\to\infty} \mathbb{E}[G_d(\boldsymbol{i}u; \boldsymbol{q})]\Big),$$

and

$$\Big\|\nabla_{\boldsymbol{q}}^k G_d(0; \boldsymbol{q}) - \lim_{u\to 0+}\lim_{d\to\infty} \mathbb{E}[\nabla_{\boldsymbol{q}}^k G_d(\boldsymbol{i}u; \boldsymbol{q})]\Big\| = o_{d,\mathbb{P}}(1),$$

where $o_{d,\mathbb{P}}(1)$ stands for convergence to 0 in probability.

4.3. From constrained forms to Lagrangian forms

Before we give the asymptotics of U and T as defined in Eq. (4) and (5), we first consider their dual forms which are more amenable in analysis. These are given by

$$\overline{U}(\lambda, N, n, d) \equiv \sup_{\boldsymbol{a}} \Big[R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a}) - \psi_1 \lambda \|\boldsymbol{a}\|_2^2 \Big], \tag{20}$$

$$\begin{aligned} \overline{T}(\lambda, N, n, d) \equiv \sup_{\boldsymbol{a}} \inf_{\boldsymbol{\mu}} \Big[& R(\boldsymbol{a}) - \lambda \psi_1 \|\boldsymbol{a}\|_2^2 \\ & + 2\langle \boldsymbol{\mu}, \boldsymbol{Z}\boldsymbol{a} - \boldsymbol{y}/\sqrt{d}\, \rangle \Big]. \end{aligned} \tag{21}$$

The proposition below shows that the strong duality holds upon the constrained forms and their dual forms.

Proposition 1 (Strong Duality)**.** *For any $A > 0$, we have*

$$U(A, N, n, d) = \inf_{\lambda \geq 0} \Big[\overline{U}(\lambda, N, n, d) + \lambda A \Big].$$

Moreover, for any $A > \psi_1 \|\boldsymbol{a}_{\min}\|_2^2$, we have

$$T(A, N, n, d) = \inf_{\lambda \geq 0} \Big[\overline{T}(\lambda, N, n, d) + \lambda A \Big].$$

The proof of Proposition 1 is based on a classical result which states that strongly duality holds for quadratic programs with single quadratic constraint (Appendix B.1 in Boyd & Vandenberghe (2004)).

4.4. Expressions of $\mathcal{U}$ and $\mathcal{T}$

Proposition 1 transforms our task from computing the asymptotics of U and T to that of $\overline{U}$ and $\overline{T}$. The latter is given by the following proposition.

Proposition 2. *Let the target function f_d satisfy Assumption 1, the activation function σ satisfy Assumption 2, and (N, n, d) satisfy Assumption 3. In addition, let Assumption 4 and 5 hold. Then for $\lambda \in \Lambda_U$, with high probability the maximizer in Eq. (20) can be achieved at a unique point $\overline{\boldsymbol{a}}_U(\lambda)$, and we have*

$$\begin{aligned} \overline{U}(\lambda, N, n, d) &= \overline{\mathcal{U}}(\lambda, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1), \\ \psi_1 \|\overline{\boldsymbol{a}}_U(\lambda)\|_2^2 &= \mathcal{A}_U(\lambda, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1). \end{aligned}$$

Moreover, for any $\lambda \in \Lambda_T$, with high probability the maximizer in Eq. (21) can be achieved at a unique point $\overline{\boldsymbol{a}}_T(\lambda)$, and we have

$$\begin{aligned} \overline{T}(\lambda, N, n, d) &= \overline{\mathcal{T}}(\lambda, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1), \\ \psi_1 \|\overline{\boldsymbol{a}}_T(\lambda)\|_2^2 &= \mathcal{A}_T(\lambda, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1). \end{aligned}$$

The functions $\overline{\mathcal{U}}, \overline{\mathcal{T}}, \mathcal{A}_U, \mathcal{A}_T$ are given in Definition 5 in Appendix A.

Remark 1. *Here we present the heuristic formulae of $\overline{\mathcal{U}}, \overline{\mathcal{T}}, \mathcal{A}_U, \mathcal{A}_T$, and defer their rigorous definition to the appendix. Define a function $g_0(\boldsymbol{q}; \boldsymbol{\psi})$ by*

$$\begin{aligned} g_0(\boldsymbol{q}; \boldsymbol{\psi}) \equiv \; & \mathrm{ext}_{z_1, z_2} \Big[\log\big((s_2 z_1 + 1)(t_2 z_2 + 1) \\ & - \mu_1^2 (1+p)^2 z_1 z_2 \big) - \mu_\star^2 z_1 z_2 + s_1 z_1 + t_1 z_2 \\ & - \psi_1 \log(z_1/\psi_1) - \psi_2 \log(z_2/\psi_2) - \psi_1 - \psi_2 \Big], \end{aligned} \tag{22}$$

where ext *stands for setting z_1 and z_2 to be stationery (which is a common symbol in statistical physics heuristics). We then take*

$$\overline{\mathcal{U}}(\lambda, \boldsymbol{\psi}) = F_1^2 (1 - \mu_1^2 \gamma_{s_2} - \gamma_p - \gamma_{t_2}) + \tau^2 (1 - \gamma_{t_1}),$$

where $\gamma_a \equiv \partial_a g_0(\boldsymbol{q}; \boldsymbol{\psi})|_{\boldsymbol{q} = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, \psi_2, 0, 0)}$ for the symbol $a \in \{s_1, s_2, t_1, t_2, p\}$, and

$$\overline{\mathcal{T}}(\lambda, \boldsymbol{\psi}) = F_1^2 (1 - \mu_1^2 \nu_{s_2} - \nu_p - \nu_{t_2}) + \tau^2 (1 - \nu_{t_1}),$$

where we define $\nu_a \equiv \partial_a g_0(\boldsymbol{q}; \boldsymbol{\psi})|_{\boldsymbol{q} = (\mu_\star^2 - \lambda\psi_1, \mu_1^2, 0, 0, 0)}$ for symbols $a \in \{s_1, s_2, t_1, t_2, p\}$. Finally $\mathcal{A}_U = -\partial_\lambda \overline{\mathcal{U}}$, $\mathcal{A}_T = -\partial_\lambda \overline{\mathcal{T}}$. By a further simplification, we can express these formulae to be rational functions of $(\mu_1^2, \mu_\star^2, \lambda, \psi_1, \psi_2, m_1, m_2)$ where (m_1, m_2) is the stationery point of the variational problem in Eq. (22) (c.f. Remark 2).

We next define $\mathcal{U}$ and $\mathcal{T}$ to be dual forms of $\overline{\mathcal{U}}$ and $\overline{\mathcal{T}}$.

Definition 2 (Formula for uniform convergence bounds)**.** *For $A \in \Gamma_U \equiv \{\mathcal{A}_U(\lambda, \psi_1, \psi_2) : \lambda \in \Lambda_U\}$, define*

$$\mathcal{U}(A, \psi_1, \psi_2) \equiv \inf_{\lambda \geq 0} \Big[\overline{\mathcal{U}}(\lambda, \psi_1, \psi_2) + \lambda A \Big].$$

For $A \in \Gamma_T \equiv \{\mathcal{A}_T(\lambda, \psi_1, \psi_2) : \lambda \in \Lambda_T\}$, define

$$\mathcal{T}(A, \psi_1, \psi_2) \equiv \inf_{\lambda \geq 0} \Big[\overline{\mathcal{T}}(\lambda, \psi_1, \psi_2) + \lambda A \Big].$$

Finally, we are ready to present the main theorem of this paper, which states that the uniform convergence bounds $U(A, N, n, d)$ and $T(A, N, n, d)$ converge to the formula presented in the definition above.

Theorem 1. *Let the same assumptions in Proposition 2 hold. For any $A \in \Gamma_U$, we have*

$$U(A, N, n, d) = \mathcal{U}(A, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1), \tag{23}$$

and for $A \in \Gamma_T$ we have

$$T(A, N, n, d) = \mathcal{T}(A, \psi_1, \psi_2) + o_{d,\mathbb{P}}(1), \tag{24}$$

where functions $\mathcal{U}$ and $\mathcal{T}$ are given in Definition 2.

The proof of this theorem is contained in Section E.

5. Discussions

In this paper, we calculated the uniform convergence bounds for random features models in the proportional scaling regime. Our results exhibit a setting in which standard uniform convergence bound is vacuous while uniform convergence over interpolators gives a non-trivial bound of the actual generalization error.

Modeling assumptions and technical assumptions. We made a few assumptions to prove the main result Theorem 1. Some of these assumptions can be relaxed. Indeed, if we assume a non-linear target function f_d instead of a linear one as in Assumption 1, the non-linear part will behave like additional noises in the proportional scaling limit. However, proving this rigorously requires substantial technical work. Similar issue exists in Mei & Montanari (2019). Moreover, it is not essential to assume vanishing μ_0^2 in Assumption 2.

Assumption 4 and 5 involve some properties of specific random matrices. We believe these assumptions can be proved under more natural assumptions on the activation function σ. However, proving these assumptions requires developing some sophisticated random matrix theory results, which could be of independent interest.

Relationship with non-asymptotic results. We hold the same opinion as in Abbaras et al. (2020): the exact formulae in the asymptotic limit can provide a complementary view to the classical theories of generalization. On the one hand, asymptotic formulae can be used to quantify the tightness of non-asymptotic bounds; on the other hand, the asymptotic formulae in many cases are comparable to non-asymptotic bounds. For example, Lemma 22 in Bartlett & Mendelson (2003) coupled with the bound of Lipschitz constant of the square loss in proper regime implies that $\mathcal{U}_\infty(A, \psi_2)$ have a non-asymptotic bound that scales linearly in A and inverse proportional to $\psi_2^{1/2}$ (c.f. Proposition 6 of E et al. (2020)). This coincides with the intuitions in Section 3.3.

Uniform convergence in other settings. A natural question is whether the power law derived in Section 3 holds for models in more general settings. One can perform a similar analysis to calculate the uniform convergence bounds in a few other settings (Montanari et al., 2019; Dhifallah & Lu, 2020; Hu & Lu, 2020). We believe the power law may be different, but the qualitative properties of uniform convergence bounds will share some similar features.

Relationship with Zhou et al. (2020). The separation of uniform convergence bounds (U and T) is first pointed out by Zhou et al. (2020), where the authors worked with the linear regression model in the "junk features" setting. We believe random features model are more natural models to illustrate the separation: in Zhou et al. (2020), there are some unnatural parameters λ_n, d_J that are hard to make connections to deep learning models, while the random features model is closely related to two-layer neural networks.

References

Abbaras, A., Aubin, B., Krzakala, F., and Zdeborová, L. Rademacher complexity and spin glasses: A link between the replica and statistical theories of learning. In *Mathematical and Scientific Machine Learning*, pp. 27–54. PMLR, 2020.

Adlam, B. and Pennington, J. Understanding double descent requires a fine-grained bias-variance decomposition. *arXiv preprint arXiv:2011.03321*, 2020.

Advani, M. S., Saxe, A. M., and Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger generalization bounds for deep nets via a compression approach. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 254–263, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/arora18b.html.

Bach, F. On the equivalence between quadrature rules and random features. *arXiv preprint arXiv:1502.06800*, pp. 135, 2015.

Bartlett, P. L. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998. doi: 10.1109/18.661502.

Bartlett, P. L. and Long, P. M. Failures of model-dependent generalization bounds for least-norm interpolation. *arXiv preprint arXiv:2010.08479*, 2020.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463–482, March 2003. ISSN 1532-4435.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 08 2005. doi: 10.1214/009053605000000282. URL https://doi.org/10.1214/009053605000000282.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 6240–6249. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/

b22b257ad0519d4500539da3c8bcf4dd-Paper.pdf.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907378117. URL https://www.pnas.org/content/early/2020/04/22/1907378117.

Belkin, M., Hsu, D. J., and Mitra, P. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 2300–2311. Curran Associates, Inc., 2018a. URL https://proceedings.neurips.cc/paper/2018/file/e22312179bf43e61576081a2f250f845-Paper.pdf.

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 541–549, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018b. PMLR. URL http://proceedings.mlr.press/v80/belkin18a.html.

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pp. 541–549. PMLR, 2018c.

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

Belkin, M., Rakhlin, A., and Tsybakov, A. B. Does data interpolation contradict statistical optimality? In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1611–1619. PMLR, 16–18 Apr 2019b. URL http://proceedings.mlr.press/v89/belkin19a.html.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441.

Cantelli, F. Sulla determinazione empirica della legge di probabilita. *Giornale dell'Istituto Italiano degli Attuari*, 38(4):421–424, 1933.

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., dtextquotesingle Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 10836–10846. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/cf9dc5e4e194fc21f397b4cac9cc3ae9-Paper.pdf.

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

Chihara, T. S. *An introduction to orthogonal polynomials*. Courier Corporation, 2011.

d'Ascoli, S., Refinetti, M., Biroli, G., and Krzakala, F. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pp. 2280–2290. PMLR, 2020.

Dhifallah, O. and Lu, Y. M. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data, 2017.

E, W., Ma, C., and Wu, L. Machine learning from a continuous viewpoint, i. *Science China Mathematics*, 63(11): 2233–2266, 2020.

Efthimiou, C. and Frye, C. *Spherical harmonics in p dimensions*. World Scientific, 2014.

El Karoui, N. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

Gerace, F., Loureiro, B., Krzakala, F., Mézard, M., and Zdeborová, L. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pp. 3452–3462. PMLR, 2020.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33, 2020.

Glivenko, V. Sulla determinazione empirica della legge di probabilita. *Giornale dell'Istituto Italiano degli Attuari*, 38(4):92–99, 1933.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 297–299. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/golowich18a.html.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Hu, H. and Lu, Y. M. Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669*, 2020.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pp. 4631–4640. PMLR, 2020a.

Jacot, A., Şimşek, B., Spadaro, F., Hongler, C., and Gabriel, F. Kernel alignment risk estimator: Risk prediction from training data. *arXiv preprint arXiv:2006.09796*, 2020b.

Liang, T., Rakhlin, A., and Zhai, X. On the risk of minimum-norm interpolants and restricted lower isometry of kernels. *arXiv:1908.10292*, 2019.

Liang, T., Rakhlin, A., et al. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48 (3):1329–1347, 2020.

Louart, C., Liao, Z., Couillet, R., et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Ma, C., Wojtowytsch, S., Wu, L., et al. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. *arXiv preprint arXiv:2009.10713*, 2020.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv e-prints*, art. arXiv:1908.05355, August 2019.

Mei, S., Misiakiewicz, T., and Montanari, A. Generalization error of random features and kernel methods: hypercontractivity and kernel matrix concentration. *arXiv preprint arXiv:2101.10588*, 2021.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlche Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 11615–11626. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper/2019/file/05e97c207235d63ceb1db43c60db7bbb-Paper.pdf.

Nagarajan, V. and Kolter, Z. Deterministic PAC-bayesian generalization bounds for deep networks via generalizing noise-resilience. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=Hygn2o0qKX.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1g5sA4twr.

Negrea, J., Dziugaite, G. K., and Roy, D. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, pp. 7263–7272. PMLR, 2020.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR (Workshop)*, 2015a. URL http://arxiv.org/abs/1412.6614.

Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. In Grünwald, P., Hazan, E., and Kale, S. (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1376–1401, Paris, France, 03–06 Jul 2015b. PMLR. URL http://proceedings.mlr.press/v40/Neyshabur15.html.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

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

ence on Learning Representations, 2019. URL https://openreview.net/forum?id=BygfghAcYX.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *NIPS*, pp. 1177–1184, 2007. URL http://papers.nips.cc/paper/3182-random-features-for-large-scale-kernel-machines.

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pp. 1313–1320, 2009.

Rudi, A. and Rosasco, L. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pp. 3215–3225, 2017.

Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 2019.

Szego, Gabor. *Orthogonal polynomials*, volume 23. American Mathematical Soc., 1939.

Vapnik, V. N. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995. ISBN 0-387-94559-8.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

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

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *0*, 2016. URL http://arxiv.org/abs/1611.03530. cite arxiv:1611.03530Comment: Published in ICLR 2017.

Zhou, L., Sutherland, D., and Srebro, N. On uniform convergence and low-norm interpolation learning. *arXiv preprint arXiv:2006.05942*, 2020.

Zhu, J., Gibson, B., and Rogers, T. T. Human rademacher complexity. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems*, volume 22, pp. 2322–2330. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf.