Uniform Convergence of Interpolators: Gaussian Width, Norm Bounds and Benign Overfitting

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider interpolation learning in high-dimensional linear regression with Gaussian data, and prove a generic uniform convergence guarantee on the generalization error of interpolators in an arbitrary hypothesis class in terms of the class's Gaussian width. Applying the generic bound to Euclidean norm balls recovers the consistency result of Bartlett et al. (2020) for minimum-norm interpolators, and confirms a prediction of Zhou et al. (2020) for near-minimal-norm interpolators. We demonstrate the generality of the bound by applying it to the simplex, obtaining a novel consistency result for minimum ℓ_1-norm interpolators (basis pursuit). Our results show how norm-based generalization bounds can explain and be used to analyze benign overfitting, at least in some settings.

1 Introduction

The traditional understanding of machine learning suggests that models with zero training error tend to overfit, and explicit regularization is often necessary to achieve good generalization. Given the empirical success of deep learning models with zero training error [37] and the (re-)discovery of the "double descent" phenomenon [4], however, it has become clear that the textbook U-shaped learning curve is only part of a larger picture: it is possible for an overparameterized model with zero training loss to achieve low population error in a noisy setting. In an effort to understand how interpolation learning occurs, recent works [e.g. 3, 11, 16, 17, 22, 30, 38] have studied linear regression with Gaussian features as a testbed problem. Significant progress has been made in this setting, including nearly-matching necessary and sufficient conditions for consistency of the minimal ℓ_2 norm interpolator [3].

Despite the fundamental role of uniform convergence in statistical learning theory, most of this line of work has used other techniques to analyze the particular minimal-norm interpolator.[1] Instead of directly analyzing the population error of a particular learning algorithm, a uniform convergence-type argument would control the worst-case generalization gap over a class of predictors containing the typical outputs of a learning rule. Typically, this is done because for many algorithms – unlike the minimal Euclidean norm interpolator – it is difficult to exactly characterize the learned predictor, but we may be able to say e.g. that its norm is not too large. Since uniform convergence does not tightly depend on a specific algorithm, the resulting analysis can highlight the key properties that lead to good generalization: it can give bounds not only for, say, the minimal-norm interpolator, but also for other interpolators with low norm [e.g. 38], increasing our confidence that low norm – and not some other property the particular minimal-norm interpolator happens to have – is key to generalization. In linear regression, practical training algorithms may not always find the exact minimal Euclidean norm solution, so it is also reassuring that all interpolators with sufficiently low Euclidean norm generalize.

[1] Negrea et al. [22] argue that Bartlett et al. [3]'s proof technique is fundamentally based on uniform convergence of a *surrogate* predictor; we discuss this in more detail in Section 4.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Nagarajan and Kolter [21], however, raised significant questions about the applicability of typical uniform convergence arguments to interpolation regimes. Following their work, several papers [2, 22, 36, 38] have demonstrated the failure of variants of uniform convergence in different setups. Despite these negative results, Zhou et al. [38] demonstrate that in one particular setting it is sufficient to consider bounds which are uniform only over predictors *with zero training error*, sidestepping all of the aforementioned lower bounds (discussed in more detail in Section 4). This weaker notion of uniform convergence has been standard in analyses of realizable settings at least since the work of Vapnik [33, Chapter 6.4] and Valiant [32]. Specifically, Zhou et al. used this notion of uniform convergence to show consistency of the minimal ℓ_2 norm interpolator in one particular learning problem, where uniform convergence arguments on sample-independent hypothesis classes cannot succeed. It remains unknown, however, whether these types of arguments can apply to more general linear regression problems and more typical asymptotic regimes, particularly showing rates of convergence rather than just consistency.

In this work, we show for the first time that uniform convergence is indeed able to explain benign overfitting in general high-dimensional Gaussian linear regression problems. Similar to the standard analysis for learning with Lipschitz losses [26], which bounds generalization gaps through Rademacher complexity, our Theorem 1 (Section 3) establishes a finite-sample high probability bound on the uniform convergence of the error of interpolating predictors in a hypothesis class, in terms of its Gaussian width. This is done through an application of the Gaussian Minimax Theorem; see the proof sketch in Section 7. Combined with an analysis of the norm of the minimal ℓ_2 norm interpolator (Theorem 2 in Section 4), our bound recovers known consistency results [3], as well as proving a conjectured upper bound for larger-norm interpolators [38].

In addition, since we do not restrict ourselves to Euclidean norm balls but instead consider interpolators in an arbitrary compact set, our results allows for a wide range of other applications. Our analysis leads to a natural extension of the consistency result and notions of effective rank of Bartlett et al. [3] for arbitrary norms (Theorem 4 in Section 5). As a demonstration of our general theory, in Section 6 we show novel consistency results for the minimal ℓ_1 norm interpolator (basis pursuit) in particular settings, which we believe are the first results of their kind.

2 Problem Formulation

Notation. We use $\|\cdot\|_p$ for the ℓ_p norm, $\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$. We always use $\max_{x\in S} f(x)$ to be $-\infty$ when S is empty, and similarly $\min_{x\in S} f(x)$ to be ∞. We use standard $O(\cdot)$ notation, and $a \lesssim b$ for inequality up to an absolute constant. For a positive semidefinite matrix A, the *Mahalanobis (semi-)norm* is $\|x\|_A^2 := \langle x, Ax\rangle$. For a matrix A and set S, AS denotes the set $\{Ax : x \in S\}$.

Data model. We assume that data (X, Y) is generated as

$$Y = Xw^* + \xi, \tag{1}$$

where $X \in \mathbb{R}^{n\times d}$ has rows $X_1, \ldots, X_n$ sampled i.i.d. from $N(0, \Sigma)$, $d \geq n$, w^* is arbitrary, and $\xi \sim N(0, \sigma^2 I_n)$ is independent of X. Though our proof techniques crucially depend on X_i being Gaussian, we can easily relax the assumption on the noise to only being sub-Gaussian; we assume Gaussian noise here for simplicity. The *empirical* and *population loss* are defined as, respectively,

$$\hat{L}(w) = \frac{1}{n}\|Y - Xw\|_2^2, \quad L(w) = \mathop{\mathbb{E}}_{(x,y)}(y - \langle w, x\rangle)^2 = \sigma^2 + \|w - w^*\|_\Sigma^2$$

where $y = \langle x, w^*\rangle + \xi_0$ with $x \sim N(0, \Sigma)$ independent of $\xi_0 \sim N(0, \sigma^2)$. For an arbitrary norm $\|\cdot\|$, the minimal norm interpolator is $\hat{w} = \arg\min_{\hat{L}(w)=0} \|w\|$. In the special case of Euclidean norm, the minimal norm interpolator can be written explicitly as $\hat{w} = X(XX^T)^{-1}Y$. If there is more than one minimal norm interpolator, all of our guarantees will hold for any minimizer $\hat{w}$.

Speculative bound. Zhou et al. [38] studied uniform convergence of low norm interpolators in a specific setting: that is, they evaluated the asymptotic limit of $\sup_{\|w\|\leq B,\, \hat{L}(w)=0} L(w) - \hat{L}(w)$. (Clearly, when $B \geq \|\hat{w}\|$, this quantity upper-bounds the population risk of $\hat{w}$.) They further speculated that a bound of the following form may hold more generally:

$$\sup_{\|w\|_2\leq B,\, \hat{L}(w)=0} L(w) - \hat{L}(w) \leq \frac{B^2\psi_n}{n} + o(1), \tag{$\star$}$$

for some appropriate choice of ψ_n based on the scale of the data. Thus, if we increase the limit on $\|w\|$ by a factor of α, ($\star$) predicts that the worst-case asymptotic generalization gap increases by a factor of α^2. The motivation of this paper is essentially to show that ($\star$) indeed holds, and to use it to show consistency of the minimal norm interpolator. Our main result (Theorem 1) can be thought of as a strengthening and significant generalization of ($\star$).

3 Generic Uniform Convergence Guarantee

To state our results, we first need to introduce some key tools.

Definition 1. The *Gaussian width* and the *radius* of a set $S \subset \mathbb{R}^d$ are

$$W(S) := \mathop{\mathbb{E}}_{g \sim N(0, I_d)} \sup_{s \in S} |\langle s, g \rangle| \quad \text{and} \quad \mathrm{rad}(S) := \sup_{s \in S} \|s\|_2.$$

The radius measures the size of a set in the Euclidean norm. The Gaussian width of a set S can be interpreted as the number of dimensions that a random projection needs to approximately preserve the norms of points in S [1, 13]. These two complexity measures are connected by Gaussian concentration: Gaussian width is the expected value of the supremum of some Gaussian process, and the radius can be considered as the typical deviation of that supremum from its expected value.

Definition 2. Given a positive semidefinite matrix $\Sigma : d \times d$ with eigenvectors $\varphi_1, \ldots, \varphi_d$ and eigenvalues $\lambda_1, \ldots, \lambda_d$, we say a pair (Σ_1, Σ_2) is a *covariance splitting* of Σ if for some $S \subseteq [d]$

$$\Sigma_1 = \sum_{i \in S} \lambda_i \varphi_i \varphi_i^T \quad \text{and} \quad \Sigma = \Sigma_1 + \Sigma_2;$$

we denote $V_1 = \mathrm{span}(\Sigma_1), V_2 = \mathrm{span}(\Sigma_2)$.

We can now state our generic bound. Section 7 sketches the proof; all full proofs are in the appendix.

Theorem 1 (Main Generalization Bound)**.** *Under the model assumptions in* (1)*, there exists an absolute constant $C_1 \leq 128$ such that the following is true. Let $\mathcal{K}$ be an arbitrary compact set, and let (Σ_1, Σ_2) be any covariance splitting of Σ. If $n \geq C_1^2(\dim(V_1) + \log(64/\delta))$, let $\beta = C_1 \left(\sqrt{\dim(V_1)} + \sqrt{\log(64/\delta)} \right) / \sqrt{n}$. With probability at least $1 - \delta$, it holds that*

$$\sup_{\substack{w \in \mathcal{K} \\ \hat{L}(w) = 0}} L(w) \leq \frac{1+\beta}{n} \left(W(\Sigma_2^{1/2}\mathcal{K}) + \mathrm{rad}(\Sigma_2^{1/2}\mathcal{K}) \sqrt{2 \log\left(\frac{16}{\delta}\right)} + \|w^*\|_{\Sigma_2} \sqrt{2 \log\left(\frac{16}{\delta}\right)} \right)^2.$$

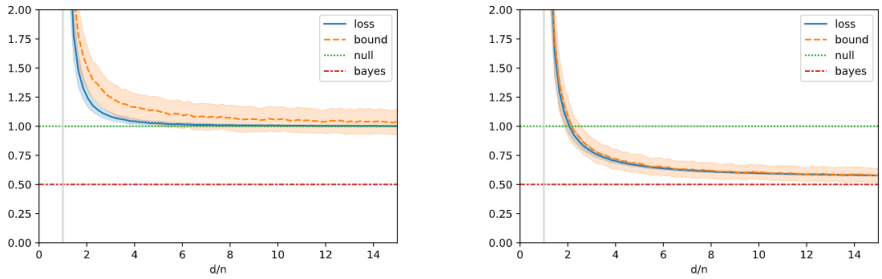
In our applications, we consider $\mathcal{K} = \{w \in \mathbb{R}^d : \|w\| \leq B\}$ for an arbitrary norm, with B based on a high probability upper bound for $\|\hat{w}\|$. Depending on the application, the dimension of V_1 will be either constant or $o(n)$, so that $\beta \to 0$. The term $\|w^*\|_{\Sigma_2}$ generally does not scale with n; hence, the $\|w^*\|_{\Sigma_2}$ term is often negligible. As hinted earlier, we can think of the Gaussian width term and the radius term as bias and variance, respectively. To recover consistency, we can expect the Gaussian width to scale like $\sigma\sqrt{n}$. This agrees with the intuition that we need increasing norm to memorize noise when the model is not realizable. The radius term requires some care in our applications, but can be handled by the covariance splitting technique. As part of the analysis in the following sections, we will rigorously show in many settings that the dominant term in the upper bound is the Gaussian width. In these cases, our upper bound is roughly $W(\Sigma_2^{1/2}\mathcal{K})^2/n$, which can be viewed as the ratio between the (probabilistic) dimension of our hypothesis class and sample size. We will also analyze the required size of $\mathcal{K}$ to contain any interpolators, allowing us to show consistency results.

4 Application: Euclidean Norm Ball

It can be easily seen that the Gaussian width of a Euclidean norm ball is nicely reduced to the product of the norm of our predictor with the typical norm of x. If $\mathcal{K} = \{w \in \mathbb{R}^d : \|w\|_2 \leq B\}$, then

$$W(\Sigma^{1/2}\mathcal{K}) = B \cdot \mathbb{E}\, \|\Sigma^{1/2} g\|_2 \leq \sqrt{B^2\, \mathbb{E}\, \|x\|_2^2}. \tag{2}$$

Therefore, it is plausible that ($\star$) holds with $\psi_n = \mathbb{E}\, \|x\|_2^2 = \mathrm{Tr}(\Sigma)$. We illustrate this generalization bound in a few simple examples motivated by [16, 38] in Figure 1. Indeed, an application of our main theorem proves that this is exactly the case.



(a) $\lambda = 1$ (isotropic) (b) $\lambda = 0.3$

Figure 1: Illustration of our generalization bound when $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & \lambda^2 I_{d-1} \end{bmatrix}$, $n = 200$, $\sigma^2 = 1/2$, $w^* = (1/\sqrt{2}, 0, \ldots, 0)$, and d is varied (x-axis). Averages (curve) and standard deviations (error bars) are estimated from 400 trials for each value of d. Here the curve marked "loss" corresponds to $L(\hat{w})$ for the minimum Euclidean norm interpolator $\hat{w}$, "bound" to $\|\hat{w}\|_2^2 \operatorname{Tr}\Sigma/n = \mathbb{E}\,\|\hat{w}\|_2^2(1+\lambda^2(d-1))/n$ which is an asymptotic bound on $L(\hat{w})$ due to Corollary 1, "null" is the loss $L(0) = 1$ of the zero estimator, and "bayes" is the Bayes-optimal error $L(w^*) = \sigma^2 = 1/2$. The vertical line is $d/n = 1$, the location of the double-descent peak; for $d/n < 1$ there are almost surely no interpolators.

Corollary 1 (Proof of Speculative Bound from [38])**.** *In the same setting as Theorem 1 with $B > \|w^*\|_2$ and $n \gtrsim \log(32/\delta)$, for some $\beta' \lesssim \sqrt[4]{\log(32/\delta)/n}$, with probability at least $1-\delta$:*

$$\sup_{\|w\|_2 \le B, \hat{L}(w)=0} L(w) \le (1+\beta')\frac{B^2 \operatorname{Tr}(\Sigma)}{n}. \tag{3}$$

The above bound is clean, but only proves a sub-optimal rate of $n^{-1/4}$. This is because the choice of covariance split used in the proof of Corollary 1 uses no information about the particular structure of Σ. Also, this bound can be slightly loose in situations where the eigenvalues of Σ decay rapidly and $\operatorname{Tr}\Sigma$ can be replaced by a smaller quantity. We next state a more precise bound on the generalization error, which requires introducing the notion of *effective rank*.

Definition 3. The effective ranks of a covariance matrix Σ are

$$r(\Sigma) = \frac{\operatorname{Tr}(\Sigma)}{\|\Sigma\|_{OP}} \quad \text{and} \quad R(\Sigma) = \frac{\operatorname{Tr}(\Sigma)^2}{\operatorname{Tr}(\Sigma^2)}.$$

The $r(\Sigma)$ rank can essentially be understood as the squared ratio between the Gaussian width and radius in our previous bound. It is related to the concentration of the ℓ_2 norm of a Gaussian vector with covariance Σ. In fact, both definitions of effective ranks can be derived by applying Bernstein's inequality to $\frac{\|x\|^2}{\mathbb{E}\,\|x\|^2}$. For more on these notions of effective rank, see [3, 18]. We will only need $r(\Sigma)$ in the generalization bound below, but we will show later, in Theorem 2, that $R(\Sigma)$ can be used to control the norm of the minimal norm interpolator $\hat{w}$.

Corollary 2. *In the same setting as Theorem 1 with $B > \|w^*\|_2$ and $n \ge C_1^2(\dim(V_1) + \log(2/\delta))$, for $\beta = C_1\left(\sqrt{\dim(V_1)} + \sqrt{\log(2/\delta)}\right)/\sqrt{n}$ and $\gamma = 4\sqrt{\frac{\log(16/\delta)}{r(\Sigma_2)}}$, with probability at least $1-\delta$,*

$$\sup_{\|w\|_2 \le B, \hat{L}(w)=0} L(w) \le (1+\beta)(1+\gamma)^2\frac{B^2 \operatorname{Tr}(\Sigma_2)}{n}. \tag{4}$$

To prove consistency with Corollary 2, we need a high probability bound for $\|\hat{w}\|_2$, the norm of the minimal norm interpolator. Such a result (Theorem 2) is obtained as a special case of our second main technical result, Theorem 3 (stated in the next section) which is a general construction of low-norm interpolators for arbitrary norms. In the Euclidean setting, Theorem 2 says that if the effective ranks $R(\Sigma_2)$ and $r(\Sigma_2)$ are large, then we can construct an interpolator with norm nearly $\|w^*\|_2 + \sigma\sqrt{n/\operatorname{Tr}(\Sigma_2)}$.

Theorem 2 (Norm bound)**.** *There exists an absolute constant $C_2 > 0$ such that the following is true. Fix any $\delta \in (0, 1/4)$. Under the model assumptions in* (1)*, for any covariance splitting (Σ_1, Σ_2) of Σ, suppose that $R(\Sigma_2) \gtrsim n \log(2/\delta) + \log^2(2/\delta)$ and $r(\Sigma_2) \gtrsim \log(16/\delta)$. Then, for $\epsilon = C_2 \left(\frac{n \log(2/\delta)}{R(\Sigma_2)} + \sqrt{\frac{\log(16/\delta)}{r(\Sigma_2)}} + \sqrt{\frac{\log(2/\delta)}{n}} \right)$, with probability at least $1 - \delta$, it holds that*

$$\|\hat{w}\|_2 \leq \|w^*\|_2 + (1 + \epsilon)\, \sigma \sqrt{\frac{n}{\mathrm{Tr}(\Sigma_2)}}. \tag{5}$$

Plugging in estimates of $\|\hat{w}\|$ to our scale-sensitive bound, we obtain a population loss guarantee for $\hat{w}$ in terms of effective ranks.

Corollary 3 (Benign overfitting)**.** *Under the model assumptions in* (1)*, for any covariance splitting (Σ_1, Σ_2), suppose that $n, r(\Sigma_2)$ and $R(\Sigma_2)$ satisfy the assumptions in Corollary 2 and Theorem 2. With the same choice of β, γ and ϵ, it holds that with probability at least $1 - \delta$,*

$$L(\hat{w}) \leq (1 + \beta)(1 + \gamma)^2 (1 + \epsilon)^2 \left(\sigma + \|w^*\|_2 \sqrt{\frac{\mathrm{Tr}(\Sigma_2)}{n}} \right)^2 \tag{6}$$

It can easily be seen that the leading terms in the upper bound besides σ^2 are $\|w^*\|_2 \sqrt{\frac{\mathrm{Tr}(\Sigma_2)}{n}}$, $\sqrt{\frac{\dim(V_1)}{n}}$, $\frac{n}{R(\Sigma_2)}$ and $\sqrt{\frac{1}{r(\Sigma_2)}}$. If all of them converge to 0, $\hat{w}$ will be consistent. Because we have from [3, Lemma 5] that $r(\Sigma_2)^2 \geq R(\Sigma_2)$, the condition $\frac{n}{R(\Sigma_2)} \to 0$ implies that $\frac{1}{r(\Sigma_2)} \to 0$, and we arrive at the following sufficient conditions:

Sufficient conditions for consistency of $\hat{w}$. As $n \to \infty$, $L(\hat{w}) - \sigma^2 \to 0$ in probability if there exists a sequence of covariance splits (Σ_1, Σ_2) of Σ such that

$$\lim_{n \to \infty} \frac{\dim(V_1)}{n} = \lim_{n \to \infty} \|w^*\|_2 \sqrt{\frac{\mathrm{Tr}\, \Sigma_2}{n}} = \lim_{n \to \infty} \frac{n}{R(\Sigma_2)} = 0. \tag{7}$$

Relationship to Bartlett et al. [3]. Our set of sufficient conditions above is almost identical to the conditions of Bartlett et al. [3], except for two differences:

1. They choose the covariance split specifically to minimize $\dim(V_1)$ such that $r(\Sigma_2) \gtrsim n$.
2. Their version of the second condition replaces $\mathrm{Tr}(\Sigma_2)$ by the larger term $\mathrm{Tr}(\Sigma)$.

From the perspective of showing $L(\hat{w}) - \sigma^2 \to 0$, the first difference is immaterial: if there exists a choice of split that satisfies our conditions, it can be shown that there exists a (possibly different) split which will also satisfy $r(\Sigma_2) \gtrsim n$ (see Appendix H). On the other hand, the second point is a genuine improvement over the consistency result of [3] when Σ has a few very large eigenvalues; this improvement in the conditions has also been implicitly obtained by Tsigler and Bartlett [30].

Regarding the rate of convergence, our additional $r(\Sigma_2)^{-1/2}$ term and the dependence on $\sqrt{\dim(V_1)/n}$ instead of $\dim(V_1)/n$ seems slightly worse at first glance than [3], but our bound can be applied for a smaller value of $\dim(V_1)$ and is better in the $\|w^*\|_2 \sqrt{\mathrm{Tr}(\Sigma_2)/n}$ term. We believe these differences are minimal in most cases, and not so important for our primary goal to showcase the power of uniform convergence.

Relationship to Negrea et al. [22]. The consistency result of Bartlett et al. [3] can also be recovered with a uniform convergence-based argument [22]. Instead of considering uniform convergence over a norm ball, Negrea et al. applied uniform convergence to a surrogate predictor, and separately showed that the minimal-norm interpolator has risk close to the surrogate. Their analysis reveals an interesting connection between realizability and interpolation learning, but it does not highlight that low norm is key to good generalization, nor does it predict the worst-case error for other low-norm interpolators.

Relationship to Bartlett and Long [2]. It was recently shown that it is impossible to find a tight excess risk bound that only depends on the learned predictor and sample size [2]. This does not,

however, rule out the speculative bound in [38], nor does it contradict our results. A closer look at the relevant lower bound construction reveals that the excess risk bounds being ruled out cannot depend on either the training error $\hat{L}$ or the population noise level σ^2. The former is crucial: their bound cannot incorporate the knowledge that the training error is small, which is the defining property of uniform convergence of interpolators. The latter point is also important; they consider excess risk $(L - \sigma^2)$, but $(\star)$ and our bounds are about the generalization gap $(L - \hat{L})$.

Relationship to Yang et al. [36]. The paper [36] gives expressions for the asymptotic generalization error of predictors in a norm ball, in a random feature model. Their model does not directly address our question because their random features are non-Gaussian, but they similarly showed that uniform convergence of interpolators can lead to a non-vacuous bound. It is unclear, though, whether uniform convergence of low-norm interpolators can yield consistency in their model: they only study sets of the form $\{\|w\| \le \alpha \|\hat{w}\|\}$ where α is taken to be a constant strictly greater than 1 – a setting where we would expect a loss of $\alpha^2\sigma^2$, i.e. would not expect consistency. They also rely on numerical methods to compare their (quite complicated) analytic expressions. It remains possible that the gap between uniform convergence of interpolators and the Bayes risk vanishes in their setting as α approaches 1.

5 General Norm Ball

All the results on the Euclidean setting are special cases of the following results for arbitrary norms.

Definition 4. The *dual norm* of a norm $\|\cdot\|$ on $\mathbb{R}^d$ is $\|u\|_* := \max_{\|v\|=1} \langle v, u\rangle$, and its sub-gradient set is $\partial\|u\|_* = \{v : \|v\| = 1, \langle v, u\rangle = \|u\|_*\}$.

Similar to the Euclidean case, the Gaussian width of a general norm ball reduces to the product of the norm of our predictor and the *dual* norm of x: $W(\Sigma^{1/2}\mathcal{K}) = B \cdot \mathbb{E}\,\|x\|_*$. We can extend the definition of one of the effective ranks by the ratio of Gaussian width and radius. The definition of the other generalized rank arises naturally from our norm bound (Theorem 3 below).

Definition 5. The effective $\|\cdot\|$-ranks of a covariance matrix Σ are

$$r_{\|\cdot\|}(\Sigma) = \left(\frac{\mathbb{E}\left\|\Sigma^{1/2}g\right\|_*}{\sup_{\|w\|\le 1}\|w\|_\Sigma}\right)^2 \quad \text{and} \quad R_{\|\cdot\|}(\Sigma) = \left(\frac{\mathbb{E}\left\|\Sigma^{1/2}g\right\|_*}{\mathbb{E}\min_{v\in\partial\left\|\Sigma^{1/2}g\right\|_*}\|v\|_\Sigma}\right)^2$$

where $g \sim N(0, I_d)$. In many examples, $\partial\|\Sigma^{1/2}g\|_*$ is a singleton set, and so v is the unique gradient.

Applying our general result, we obtain an analogue of Corollary 2.

Corollary 4. *In the same setting as Theorem 1 with $B > \|w^*\|$ and $n \ge C_1^2(\dim(V_1) + \log(2/\delta))$, for $\beta = C_1\left(\sqrt{\dim(V_1)} + \sqrt{\log(2/\delta)}\right)/\sqrt{n}$ and $\gamma = 4\sqrt{\frac{\log(16/\delta)}{r_{\|\cdot\|}(\Sigma_2)}}$, with probability at least $1-\delta$*

$$\sup_{\|w\|\le B, \hat{L}(w)=0} L(w) \le (1+\beta)(1+\gamma)^2 \frac{\left(B\cdot\mathbb{E}\|\Sigma_2^{1/2}g\|_*\right)^2}{n}. \tag{8}$$

As in the Euclidean special case, we need to combine this result with a construction of a low-norm interpolator to obtain consistency results for minimum norm interpolation. This leads us to the second main technical result of this paper, which bounds the norm of $\hat{w} = \arg\min_{Xw=Y}\|\hat{w}\|$ for an arbitrary norm. This theorem essentially says that if the effective ranks $R_{\|\cdot\|}(\Sigma_2)$ and $r_{\|\cdot\|}(\Sigma_2)$ are sufficiently large, then there exists an interpolator with norm $\|w^*\| + \sigma\sqrt{n}/\,\mathbb{E}\|\Sigma_2^{1/2}g\|_*$.

Theorem 3 (General norm bound)**.** *Fix any $\delta \in (0, 1/4)$. Under the model assumptions in* (1)*, for any covariance splitting (Σ_1, Σ_2) of Σ, suppose that $r_{\|\cdot\|}(\Sigma_2) \ge 32\log(16/\delta)$ and $\tau \le 1/4$ is chosen such that $\tau = \Omega(\sqrt{\log(2/\delta)/n})$ and*

$$\Pr_{g\sim N(0,I_d)}\left(\min_{v\in\partial\|\Sigma_2^{1/2}g\|_*}\|v\|_{\Sigma_2} \le \mathop{\mathbb{E}}_{g'\sim N(0,I_d)}\left[\min_{v\in\partial\|\Sigma_2^{1/2}g'\|_*}\|v\|_{\Sigma_2}\right]\sqrt{\frac{R_{\|\cdot\|}(\Sigma_2)}{n}}\tau\right) \ge 1 - \frac{\delta}{8}. \tag{9}$$

Then for $\epsilon = \tau + \sqrt{\frac{8\log(16/\delta)}{r_{\|\cdot\|}(\Sigma_2)}}$, with probability at least $1-\delta$, it holds that

$$\|\hat{w}\| \le \|w^*\| + (1+\epsilon)\,\sigma\frac{\sqrt{n}}{\mathbb{E}\,\|\Sigma_2^{1/2}g\|_*}. \tag{10}$$

For a specific choice of norm $\|\cdot\|$, we can verify that $\min_{v \in \partial \|\Sigma_2^{1/2} g\|_*} \|v\|_{\Sigma_2}$ concentrates about its mean; thus (9) essentially ensures that τ upper-bounds $n/R_{\|\cdot\|}(\Sigma)$. Combined with the assumption that it is at least $\sqrt{\log(2/\delta)/n}$, the definition of ϵ coincides with Theorem 2.

The general definition of $R_{\|\cdot\|}(\Sigma)$ may appear mysterious at first sight. Large effective rank means that the sub-gradient of $\|\Sigma_2^{1/2} g\|_*$ is small in the $\|\cdot\|_{\Sigma_2}$ norm. This is, in fact, closely related to the existence of low-norm interpolators. First, note that $\Sigma_2^{1/2} g$ corresponds to the small-eigenvalue components of the covariate vector. For v to be a sub-gradient means that moving the weight vector w in the direction of v is very effective at changing the prediction $\langle w, X_i \rangle$; having small $\|\cdot\|_{\Sigma_2}$ norm means that moving in this direction has a very small effect on the population loss $L(w)$. Together, this means that the sub-gradient will be a good direction for benignly overfitting the noise. The factor of n is needed because there are n training examples to benignly overfit.

Straightforwardly combining Corollary 4 and Theorem 3 yields the following theorem, which gives guarantees for minimal-norm interpolators in terms of effective rank conditions. Just as in the Euclidean case, we can extract from this result a simple set of sufficient conditions for consistency of the minimal norm interpolator.

Theorem 4 (Benign overfitting)**.** *Under the model assumptions in* (1)*, for any covariance splitting* (Σ_1, Σ_2)*, suppose that* n, τ *and* $r_{\|\cdot\|}(\Sigma_2)$ *satisfy the assumptions in Corollary 4 and Theorem 3. With the same choice of* β, γ *and* ϵ*, it holds that with probability at least* $1 - \delta$*,*

$$L(\hat{w}) \leq (1+\beta)(1+\gamma)^2(1+\epsilon)^2 \left(\sigma + \|w^*\| \frac{\mathbb{E}\, \|\Sigma_2^{1/2} g\|_*}{\sqrt{n}} \right)^2. \tag{11}$$

Sufficient conditions for consistency of $\hat{w}$. As $n \to \infty$, $L(\hat{w}) - \sigma^2 \to 0$ in probability if there exists a sequence of covariance splits (Σ_1, Σ_2) of Σ such that

$$\lim_{n\to\infty} \frac{\dim(V_1)}{n} = \lim_{n\to\infty} \frac{\|w^*\|\, \mathbb{E}\, \|\Sigma_2^{1/2} g\|_*}{\sqrt{n}} = \lim_{n\to\infty} \frac{1}{r_{\|\cdot\|}(\Sigma_2)} = \lim_{n\to\infty} \frac{n}{R_{\|\cdot\|}(\Sigma_2)} = 0. \tag{12}$$

As we see, the conditions for a minimal norm interpolator to succeed with a general norm generalize those from the Euclidean setting in a natural way. The only notable difference from the Euclidean setting is that we have two large effective dimension conditions on Σ_2 instead of a single one; in the Euclidean case, the condition on R implies the condition on r.

6 Application: ℓ_1 Norm Balls for Basis Pursuit

Unlike the minimal ℓ_2 norm interpolator, the theory for the minimal ℓ_1 norm interpolator, also known as *basis pursuit* [7] – $\hat{w}_{BP} = \arg\min_{\hat{L}(w)=0} \|w\|_1$ – is much less developed. In this section, we illustrate the consequences of our general theory for basis pursuit.

Recall that the dual of the ℓ_1 norm is the ℓ_∞ norm and $\partial \|u\|_* = \mathrm{conv}\{\mathrm{sign}(u_i)\, e_i : i \in \arg\max |u_i|\}$, where $\mathrm{conv}(S)$ denotes the convex hull of S. From the definition of sub-gradient, we observe that

$$\min_{v \in \partial \|\Sigma^{1/2} g\|_*} \|v\|_\Sigma \leq \max_{i \in [d]} \|e_i\|_\Sigma = \sqrt{\max_i \Sigma_{ii}}. \tag{13}$$

Furthermore, by convexity we have

$$\max_{\|w\|_1 \leq 1} \|w\|_\Sigma = \sqrt{\max_i \langle e_i, \Sigma e_i \rangle} = \sqrt{\max_i \Sigma_{ii}} \tag{14}$$

and so $r_{\|\cdot\|_1}(\Sigma) = \frac{\left(\mathbb{E}\, \|\Sigma^{1/2} g\|_\infty\right)^2}{\max_i (\Sigma)_{ii}} \leq R_{\|\cdot\|_1}(\Sigma)$. Therefore, we can use a single notion of effective rank. For simplicity, we denote $r_1(\Sigma) = r_{\|\cdot\|_1}(\Sigma)$. Another consequence of (13) is that we can choose $\tau = \frac{n}{r_1(\Sigma_2)} + O\left(\sqrt{\log(2/\delta)/n}\right)$ in Theorem 4 to obtain a finite-sample risk bound. Combining this discussion with (12), we obtain the following sufficient conditions for consistency of basis pursuit.

Sufficient conditions for consistency of $\hat{w}_{BP}$. As $n \to \infty$, $L(\hat{w}) - \sigma^2 \to 0$ in probability if there exists a sequence of covariance splits (Σ_1, Σ_2) of Σ such that

$$\lim_{n\to\infty} \frac{\dim(V_1)}{n} = \lim_{n\to\infty} \frac{\|w^*\|_1 \, \mathbb{E}\|\Sigma_2^{1/2} g\|_\infty}{\sqrt{n}} = \lim_{n\to\infty} \frac{n}{r_1(\Sigma_2)} = 0. \tag{15}$$

Application: Junk features. We now consider the behavior of basis pursuit in a junk feature model similar to [38]. Suppose that $\Sigma = \begin{bmatrix} \Sigma_s & 0 \\ 0 & \frac{\lambda_n}{\log(d)} I_d \end{bmatrix}$, where Σ_s is a fixed matrix and $\|w^*\|_1$ is fixed. Quite naturally, we choose the covariance splitting $\Sigma_1 = \begin{bmatrix} \Sigma_s & 0 \\ 0 & 0 \end{bmatrix}$, in which case the first sufficient condition is obviously satisfied. By standard results on the maximum of independent Gaussian variables [e.g. 34], it is routine to check that

$$\frac{\mathbb{E}\,\|\Sigma_2^{1/2} g\|_\infty}{\sqrt{n}} = \Theta\left(\sqrt{\frac{\lambda_n}{n}}\right) \quad \text{and} \quad r_1(\Sigma_2) = \Theta\left(\log(d)\right).$$

Therefore, basis pursuit interpolation will be consistent provided that $\lambda_n = o(n)$ and $d = e^{\omega(n)}$. To the best of our knowledge, the above result is the first time that basis pursuit has been shown to be consistent in any setting with Gaussian covariates and $\sigma > 0$. Although we show consistency, the rate of convergence is extremely slow because of the dependence on $n/\log(d)$ and $1/\sqrt{\log(d)}$.

Application: Isotropic features. Similar to the Euclidean case, we generally do not expect the basis pursuit interpolator be consistent when $\Sigma = I_d$ unless $w^* = 0$. However, we can expect its risk to approach the null risk $\sigma^2 + \|w^*\|^2$ if $d = e^{\omega(n)}$, and we will show this fact using uniform convergence (without covariance splitting). (Full statements and proofs can be found in the appendix.)

A direct application of Theorem 4 is not enough because the $\|w^*\|_1 \sqrt{\frac{\log(d)}{n}}$ term diverges, but we can get rid of the dependence on $\sqrt{\log(d)/n}$ with a better norm bound. Let S be the support of w^* and denote X_S as the matrix formed by selecting the columns of X in S. The key observation is that we can rewrite our model as $Y = X_{S^c} 0 + (X_S w_S^* + \xi)$, which corresponds to the case when $w^* = 0$ and the Bayes risk is $\sigma^2 + \|w^*\|_2^2$. If we interpolate using only the features in S^c, the minimal norm will be approximately upper bounded by $\sqrt{\sigma^2 + \|w^*\|_2^2} \frac{\sqrt{n}}{\mathbb{E}\,\|g\|_*}$ as long as $d - |S| = e^{\omega(n)}$, by Theorem 3. This implies the original model $\|\hat{w}_{BP}\|_1$ can also be upper bounded by the same quantity with high probability. Plugging the norm estimate in to Corollary 4 yields a risk bound of $\sigma^2 + \|w^*\|_2^2$.

Relationship to previous works. Both Ju et al. [17] and Chinot et al. [8] study the minimal ℓ_1 norm interpolator, but only in the isotropic setting. They consider a more realistic scaling where $\log(d)/n$ is not large and the target is not the null risk. The best bound in [17], their Corollary 3, is $L(\hat{w}_{BP}) \le \sigma^2(2 + 32\sqrt{14}\sqrt{s})^2$, where s is the ground truth sparsity. Note that even when $w^* = 0$, this bound does not show consistency. Similarly, Chinot et al. [8] establish sufficient conditions for $L(\hat{w}_{BP}) = O(\sigma^2)$, which is nontrivial but also does not show consistency for any $\sigma > 0$. In contrast, the constants in our result are tight enough to show $L(\hat{w}_{BP}) \to \sigma^2$ in the isotropic setting when $w^* = 0$ and in the junk feature setting when $\|w^*\|_1$ is bounded.

Like our work, the results of [8] generalize to arbitrary norms; they also consider a larger class of anti-concentrated covariate distributions than just Gaussians, as in [20]. If $\sigma = 0$ and $w^* \in \mathcal{K}$ (i.e. the model is well-specified and noiseless), their work as well as the earlier work of [20] can recover generalization bounds similar to our Corollary 4, but with a large leading constant.

7 Proof Sketches

A key ingredient in our analysis is a celebrated result from Gaussian process theory known as the Gaussian Minmax Theorem (GMT) [14, 28]. Since the seminal work of Rudelson and Vershynin [25], the GMT has seen numerous applications to problems in statistics, machine learning, and signal processing [e.g. 10, 23, 27, 29]. Most relevant to the present work is the work of Thrampoulidis et al. [28], which introduced the Convex Gaussian Minmax Theorem (CGMT) and developed a framework

for the precise analysis of regularized linear regression. Here we apply the GMT/CGMT to study uniform convergence and the norm of the minimal norm interpolator.

Proof sketch of Theorem 1. For simplicity, we first assume there is no covariance splitting. By a change of variable and introducing the Lagrangian, we can re-write the generalization gap as

$$\begin{aligned}\sup_{\substack{w\in\mathcal{K}\\ Xw=Y}} L(w) &= \sigma^2 + \sup_{\substack{w\in\Sigma^{1/2}(\mathcal{K}-w^*)\\ Zw=\xi}} \|w\|_2^2 \\ &= \sigma^2 + \sup_{w\in\Sigma^{1/2}(\mathcal{K}-w^*)} \inf_{\lambda} \langle \lambda, Zw-\xi\rangle + \|w\|_2^2\end{aligned} \tag{16}$$

where Z is a random matrix with i.i.d. standard normal entries. By GMT[2], we can control the upper tail of the max-min problem above (PO) by the auxiliary problem below (AO), with $g, h \sim N(0, I)$:

$$\sup_{w\in\Sigma^{1/2}(\mathcal{K}-w^*)} \inf_{\lambda} \|\lambda\|_2\langle g, w\rangle + \langle \lambda, h\|w\|_2 - \xi\rangle + \|w\|_2^2 = \sup_{\substack{w\in\Sigma^{1/2}(\mathcal{K}-w^*)\\ \left\|h\|w\|_2-\xi\right\|_2\le\langle g,w\rangle}} \|w\|_2^2. \tag{17}$$

By standard concentration results, we can expect $\|h\|_2^2/n \approx 1$ and $\|\xi\|_2^2/n \approx \sigma^2$, so expanding the second constraint in the AO, we obtain $\|w\|_2^2 + \sigma^2 \le |\langle g, w\rangle|^2/n$. Plugging into (16), we have essentially shown that

$$\sup_{\substack{w\in\mathcal{K}\\ Xw=Y}} L(w) \le \sup_{w\in\Sigma^{1/2}(\mathcal{K}-w^*)} \frac{|\langle g, w\rangle|^2}{n} \le \frac{\left(\sup_{w\in\Sigma^{1/2}\mathcal{K}} |\langle g, w\rangle| + |\langle g, \Sigma^{1/2}w^*\rangle|\right)^2}{n}. \tag{18}$$

Applying concentration on the right hand side concludes the proof sketch. In situations where the supremum does not sharply concentrate around its mean, we can apply GMT only to the small variance directions of Σ. This requires a slightly more general version of GMT, which we prove in the Appendix. We also show the additional terms contributed by the large variance components of X cancel out due to Wishart concentration. This is reflected in the β term of our theorem statement.

Proof sketch of Theorem 3. Since the minimal norm problem is *convex-concave*, we can apply the CGMT, which provides useful direction that GMT cannot. By the same argument as above

$$\begin{aligned}\inf_{Xw=Y} \|w\| - \|w^*\| \le \inf_{Xw=\xi} \|w\| &= \inf_{w} \sup_{\lambda} \|\Sigma^{-1/2}w\| + \langle \lambda, Zw-\xi\rangle \\ &\approx \inf_{\|w\|_2^2+\sigma^2\le|\langle g,w\rangle|^2/n} \|\Sigma^{-1/2}w\| = \inf_{\|w\|_\Sigma^2+\sigma^2\le|\langle \Sigma^{1/2}g,w\rangle|^2/n} \|w\|\end{aligned} \tag{19}$$

To upper bound the infimum, it suffices to construct a feasible w. Consider w of the form αv where $v \in \partial\|\Sigma^{1/2}g\|_*$. Plugging in the constraint, we can choose $\|w\| = \alpha = \sqrt{\sigma^2\left(\frac{\|\Sigma^{1/2}g\|_*^2}{n} - \|v\|_\Sigma^2\right)^{-1}}$. Rearranging the terms conclude the proof sketch when there is no covariance splitting. The general proof is more technical, but follows the same idea.

8 Discussion

In this work, we prove a generic generalization bound in terms of the Gaussian width and radius of a hypothesis class. We also provide a general high probability upper bound for the norm of the minimal norm interpolator. Combining these results, we recover the sufficient conditions from [3] in the ℓ_2 case and obtain novel consistency results in the ℓ_1 case. Our results provide concrete evidence that uniform convergence is indeed sufficient to explain interpolation learning, at least in some settings.

A future direction of our work is to extend the main results to settings with non-Gaussian features; this has been achieved in other applications of the GMT [29]. Another interesting problem is to study uniform convergence of low-norm *near*-interpolators, and characterize the worst-case population error as the norm and training error *both* grow. This will lead to a more precise understanding of early stopping, by connecting the optimization path with regularization path. Finally, it is unknown whether our sufficient conditions in section 5 are necessary for consistency, and it remains a challenge to apply uniform convergence of interpolators to more complex models such as deep neural networks.

[2] We ignore a compactness issue here, but this is done rigorously by a truncation argument in the Appendix.

References

[1] Afonso S. Bandeira. "Ten Lectures and Forty-Two Open Problems in the Mathematics of Data Science." Lecture notes, MIT. 2016. URL: https://people.math.ethz.ch/~abandeira/TenLecturesFortyTwoProblems.pdf.

[2] Peter L. Bartlett and Philip M. Long. "Failures of model-dependent generalization bounds for least-norm interpolation." 2020. arXiv: 2010.08479.

[3] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. "Benign overfitting in linear regression." *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30063–30070. arXiv: 1906.11300.

[4] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine learning practice and the bias-variance trade-off." *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854. arXiv: 1812.11118.

[5] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[6] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. "The convex geometry of linear inverse problems." *Foundations of Computational Mathematics* 12.6 (2012), pp. 805–849.

[7] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. "Atomic decomposition by basis pursuit." *SIAM Review* 43.1 (2001), pp. 129–159.

[8] Geoffrey Chinot, Matthias Löffler, and Sara van de Geer. "On the robustness of minimum-norm interpolators." 2021. arXiv: 2012.00807.

[9] Stefan Cobzas. *Functional analysis in asymmetric normed spaces*. Springer Science & Business Media, 2012.

[10] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. "A model of double descent for high-dimensional binary linear classification." *Information and Inference: A Journal of the IMA* (2021). arXiv: 1911.05822.

[11] Michał Dereziński, Feynman Liang, and Michael W. Mahoney. "Exact expressions for double descent and implicit regularization via surrogate random design." *Advances in Neural Information Processing Systems*. 2020. arXiv: 1912.04533.

[12] Rick Durrett. *Probability: theory and examples*. Vol. 49. Cambridge University Press, 2019.

[13] Yehoram Gordon. "On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$." *Geometric Aspects of Functional Analysis*. Vol. 1317. Lecture Notes in Mathematics. Springer, 1988, pp. 84–106.

[14] Yehoram Gordon. "Some inequalities for Gaussian processes and applications." *Israel Journal of Mathematics* 50.4 (1985), pp. 265–289.

[15] Ramon van Handel. "Probability in High Dimension." Lecture notes, Princeton University. 2014. URL: https://web.math.princeton.edu/~rvan/APC550.pdf.

[16] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. "Surprises in High-Dimensional Ridgeless Least Squares Interpolation" (2019). arXiv: 1903.08560.

[17] Peizhong Ju, Xiaojun Lin, and Jia Liu. "Overfitting Can Be Harmless for Basis Pursuit: Only to a Degree." *Advances in Neural Information Processing Systems*. 2020. arXiv: 2002.00492.

[18] Vladimir Koltchinskii and Karim Lounici. "Concentration Inequalities and Moment Bounds for Sample Covariance Operators." *Bernoulli* 23.1 (2017), pp. 110–133. arXiv: 1405.2468.

[19] Michel Ledoux. "A heat semigroup approach to concentration on the sphere and on a compact Riemannian manifold." *Geometric & Functional Analysis GAFA* 2.2 (1992), pp. 221–224.

[20] Shahar Mendelson. "Learning without concentration." *Conference on Learning Theory*. PMLR. 2014, pp. 25–39.

[21] Vaishnavh Nagarajan and J. Zico Kolter. "Uniform convergence may be unable to explain generalization in deep learning." *Advances in Neural Information Processing Systems*. 2019. arXiv: 1902.04742.

[22] Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel M. Roy. "In Defense of Uniform Convergence: Generalization via derandomization with an application to interpolating predictors." *International Conference on Machine Learning*. 2020. arXiv: 1912.04265.

[23] Samet Oymak and Babak Hassibi. "New null space results and recovery thresholds for matrix rank minimization." 2010. arXiv: 1011.6326.

[24] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[25] Mark Rudelson and Roman Vershynin. "On sparse reconstruction from Fourier and Gaussian measurements." *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 61.8 (2008), pp. 1025–1045.

[26] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[27] Mihailo Stojnic. "A framework to characterize performance of lasso algorithms." 2013. arXiv: 1303.7291.

[28] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. "Regularized linear regression: A precise analysis of the estimation error." *Conference on Learning Theory*. PMLR. 2015, pp. 1683–1709.

[29] Joel A Tropp. "Convex recovery of a structured signal from independent random linear measurements." *Sampling Theory, a Renaissance*. Springer, 2015, pp. 67–101.

[30] Alexander Tsigler and Peter L Bartlett. "Benign overfitting in ridge regression" (2020). arXiv: 2009.14286.

[31] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[32] Leslie G Valiant. "A theory of the learnable." *Communications of the ACM* 27.11 (1984), pp. 1134–1142.

[33] Vladimir Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Series in Statistics. Springer-Verlag, 1982.

[34] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge University Press, 2018.

[35] Roman Vershynin. "Introduction to the non-asymptotic analysis of random matrices" (2010). arXiv: 1011.3027.

[36] Zitong Yang, Yu Bai, and Song Mei. "Exact Gap between Generalization Error and Uniform Convergence in Random Feature Models." *International Conference on Machine Learning*. 2021. arXiv: 2103.04554.

[37] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization." *International Conference on Learning Representations*. 2017. arXiv: 1611.03530.

[38] Lijia Zhou, Danica J. Sutherland, and Nathan Srebro. "On Uniform Convergence and Low-Norm Interpolation Learning." *Advances in Neural Information Processing Systems*. 2020. arXiv: 2006.05942.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not advocate any particular algorithm or method in practice, just a style of theoretical analysis.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] The computational resources used were trivial.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]