Minimum Norm Interpolation Meets The Local Theory of Banach Spaces

Gil Kur [1] **Pedro Abdalla** [1 *] **Pierre Bizeul** [2 *] **Fanny Yang** [1]

Abstract

Minimum-norm interpolators have recently gained attention primarily as an analyzable model to shed light on the double descent phenomenon observed for neural networks. The majority of the work has focused on analyzing interpolators in Hilbert spaces, where typically an effectively low-rank structure of the feature covariance prevents a large bias. More recently, tight vanishing bounds have also been shown for isotropic high-dimensional data for ℓ_p-spaces with $p \in [1, 2)$, leveraging sparse structure of the ground truth. However, these proofs are tailored to specific settings and hard to generalize. This paper takes a first step towards establishing a general framework that connects generalization properties of the interpolators to well-known concepts from high-dimensional geometry, specifically, from the local theory of Banach spaces. In particular, we show that under 2-uniform convexity, the bias of the minimal norm solution is bounded by the Gaussian complexity of the class. We then prove a "reverse" Efron-Stein lower bound on the expected conditional variance of the minimal norm solution under cotype 2. Finally, we prove that this bound is sharp for ℓ_p-linear regression under *sub-Gaussian* covariates. [1]

1. Introduction

Experiments with neural networks have revealed a phenomenon that defies traditional statistical intuition: regularization is critical for large models when fitting noisy data. Instead, it seems that in the overparameterized regime, interpolators that achieve zero training error can still generalize well and do not profit from sacrificing datafit, or in other words; interpolation is harmless (see e.g. experiments in (Nakkiran et al., 2021)). A call to explain this counter-intuitive observation (see, e.g. (Zhang et al., 2021; Belkin et al., 2018b)) gave rise to a line of work known as *benign overfitting* that set out to prove generalization bounds for *interpolating* overparameterized models; in the case of regression, the focus of this paper, interpolation corresponds to achieving zero square loss on training data.

As there are infinitely many interpolating solutions in overparameterized regimes, the specific choice of interpolator can drastically vary generalization performance. Most commonly studied in the literature is the family of minimum-norm interpolators - a natural choice if the ground truth has a simple structure such as a small norm in a Banach space. For additional motivation, first-order methods on the square loss initialized at zero typically exhibit an implicit bias towards (i.e. converge to) such minimum-norm solutions (Gunasekar et al., 2018; Oravkin & Rebeschini, 2021; Shamir, 2022; Efron et al., 2004).

The analysis in the overparameterized regression literature has primarily focused on Hilbert spaces such as the ℓ_2-space and Reproducing Kernel Hilbert Spaces (RKHS), where the minimum-norm solution has a closed-form solution. For linear min-ℓ_2-norm interpolators in dimension d using n number of samples, the literature consists of asymptotic results in the inconsistent proportional regime where $d/n \to \gamma$ for some constant $\gamma < \infty$ (Hastie et al., 2022; Ghorbani et al., 2021; Mei & Montanari, 2022) and non-asymptotic results in the consistent regime $d/n \to \infty$. In particular, (Bartlett et al., 2020; Tsigler & Bartlett, 2023; Lecué & Shang, 2022; Chinot et al., 2020; Muthukumar et al., 2020) prove vanishing finite-sample bounds for minimum ℓ_2-interpolator when the eigenvalues of the covariance matrix of the data decay rapidly. These proofs take advantage of the inner-product structure that allows for an explicit analysis of the closed-form solution. Further, consistency heavily relies on the eigenvalue decay of the covariate distribution, as the ℓ_2-norm cannot capture structural assumptions and in general suffers from bias in the isotropic high-dimensional setting.

In the isotropic case, different structural assumptions on the ground truth are necessary to achieve consistency. However, the minimum-norm interpolators in more general Banach spaces with a corresponding inductive bias might not have

[*]Equal contribution [1]ETH Zürich [2]Technion - Israel Institute of Technology. Correspondence to: Gil Kur <gil.kur@inf.ethz.ch>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]Please consult the arXiv version of the paper for an updated (and extended) manuscript.

a closed-form. (Koehler et al., 2021) introduce a local uniform convergence framework that can be used to analyze linear interpolators without closed-form solutions. Their approach only applies to Gaussian covariates and is based on a local uniform convergence approach and crucially relies on the (Convex) Gaussian minimax theorems (Gordon, 1988; Thrampoulidis et al., 2015). Follow-up work used this technique to establish the first consistency results for the minimum-ℓ_1-norm interpolator (Wang et al., 2022) and obtained tight fast rates for minimum-ℓ_p-norm interpolators with $p \in (1, 2]$ (Donhauser et al., 2022) (see Example 1 below). However, all prior tight analysis of mininum-ℓ_p-interpolators share the deficiency that they crucially relly on the Gaussianity of the covariates.

For non-linear regression, the phenomenon of harmless interpolation or benign overfitting is much less understood. Most work so far has considered instances when linearization is a good approximation, such as specific Reproducing Kernel Hilbert Spaces (RKHS) (Liang et al., 2020; Aerni et al., 2023) or local interpolation schemes (Belkin et al., 2019; 2018a). Despite these efforts, a comprehensive theoretical understanding to handle more general non-linear models remains open to date.

In this work, we present a new geometric framework that allows us to overcome aforementioned restrictions. In particular, we relate the underlying generalization properties of minimum-norm interpolation for regression under additive Gaussian noise with (local) geometric properties of the Banach space; these include uniform convexity and smoothness as well as type and cotype, and the K-convexity (Maurey & Pisier, 1976; Pisier, 1977; 1999) that we introduce in Section 2. This theory was pioneered by Maurey and Pisier, and emerged in the study of central limit theorems in (infinite) dimensional Banach spaces, a field known nowadays as Probability in Banach spaces (Ledoux & Talagrand, 2013) Our geometric approach allows us to derive statistical bounds for a broad class of Banach spaces: The function class need not be linear and the norm may not induce an inner product space, i.e. can be "far" from a Hilbert space. This general technique also allows us to prove tight bounds for covariate distributions beyond Gaussians.

Specifically, under assumptions on the local geometric properties of the Banach space, we bound three error terms that together make up the Mean Squared Error (MSE): the bias, variance of conditional expectations and expected conditional variance. We first provide an "unlocalized" upper bound on the sum of the bias and variance of the conditional expectation in Theorem 3.1 that holds for 2-uniformly convex norms (see Definition 2.2). Remarkably and surprisingly to the authors of this paper, it aligns with the classical (unlocalized) bound for the MSE of Empirical Risk Minimization (ERM). Then, in Theorem 3.2 we show that the cotype 2 property (weaker than 2-uniformly convex) suffices to obtain a "reverse" version of the celebrated Efron-Stein inequality (Boucheron et al., 2013); providing a lower bound on the expected conditional variance of the minimum norm interpolator. Finally, in Theorem 3.3, we show that the bound of Theorem 3.2 is sharp in the case of ℓ_p-linear regression under sub-Gaussian covariates (see Example 1).

On a high level, this paper presents a new geometric perspective that allows us to analyze the behaviour of general non-linear minimum-norm interpolators under no assumptions on the covariates. We further demonstrate how these results may be used to recover tight bounds of previous works that used very specialized proof technique only applicable to Gaussian covariates. We believe that this more general approach contributes to a more fundamental understanding of benign overfitting phenomena in high-dimensional settings.

2. Preliminaries

In this section, we first introduce our regression setting and the minimum-norm interpolator in general Banach spaces. We then provide some background and introduce classical notions from high-dimensional geometry and the local theory of Banach spaces. Finally, we discuss additional structural assumptions on our model that are used in our results. We now introduce some notation that we use throughout the paper.

Notation: $C, C_1, C_2 \geq 0$ and $c_1, c_2, c_3 \in (0, 1)$ are absolute constants, and for arbitrary argument vectors a we write $C(a) \geq 0$ and $c(a) \in (0, 1)$ for constants that only depend on a. These constants *may* change from line to line. Also, for any measure $\mathbb{Q}$ on $\mathcal{X}$, we use $\|\cdot\|_{\mathbb{Q}}$ to denote the $L_2(\mathbb{Q})$-norm. The bold non-italic notation $\mathbf{f}$ refers to the vector $(f(X_1), \dots, f(X_n))$ associated with the function f. We use the standard Landau (also known as big-O) notation that hides only absolute constants that, in particular, do not depend on any other parameter of the model. Finally, $\asymp, \lesssim, \gtrsim$ are used to denote equalities and inequalities up to a multiplicative universal constant; that is, $a \lesssim b$ stands for $a = O(b)$ whereas $a \asymp b$ indicates that both $a = O(b)$ and $b = O(a)$. Finally, we use the standard notation of $\mathbb{P}_n$ to be a random uniform measure on $X_1, \dots, X_n \underset{i.i.d.}{\sim} \mathbb{P}$.

2.1. Setting

Let $\mathcal{X}$ be some domain equipped with a metric d and let $\mathbb{P}$ be a probability measure over $\mathcal{X}$. The set of all measurable functions for which $\int |f|^2 d\mathbb{P} < \infty$ is denoted by $L_2(\mathbb{P})$ and we denote by $\|\cdot\|_{\mathbb{P}}$ to be the $L_2(\mathbb{P})$ norm. Next, consider a Banach space $(\mathcal{B}(\mathcal{X}), \|\cdot\|)$, where $\mathcal{B}(\mathcal{X})$ is some linear subspace of functions on $\mathcal{X}$ that lie in $L_2(\mathbb{P})$.

For each $\mathcal{B}(\mathcal{X})$ we would like to learn a function that has bounded norm and w.l.o.g. lies in the function class

$$\mathcal{F} := \{f \in \mathcal{B}(\mathcal{X}) : \|f\| \leq 1\} \subset \mathcal{B}(\mathcal{X}). \tag{1}$$

Our main goal is to study the following well-specified regression model

$$Y = f^*(X) + \xi,$$

where $X \sim \mathbb{P}$, $\xi \sim N(0,1)$ is a standard Gaussian and $f^* \in \mathcal{F}$.

Minimum-norm interpolator For any data points $\{(x_i, y_i)\}_{i=1}^n$, the minimum norm solution that interpolates the data is defined by

$$\widehat{f}_n(\mathbf{x}, \mathbf{y}) := \operatorname*{argmin}_{f \in \mathcal{B}(\mathcal{X}): f(x_1)=y_1, \ldots, f(x_n)=y_n} \|f\|. \tag{2}$$

In words, from all possible functions in $\mathcal{B}(\mathcal{X})$ that interpolate the observations $\mathbf{y} = (y_1, \ldots, y_n)$ over the data points $\mathbf{x} := (x_1, \ldots, x_n)$, we choose the one with the smallest norm. Note how this interpolating estimator differs from the standard ERM solution that would search only in the function class $\mathcal{F}$. As common for the analysis of interpolators, we can ensure existence of $\widehat{f}_n$ by choosing $\mathcal{B}(\mathcal{X})$ to depend on the number of samples n. In general, the solution in (2) may not be unique, but under q-uniform convexity for some $q \in [2, \infty)$ (see Def. 2.2), uniqueness is also guaranteed. From a computational perspective, depending on the case, it may be computed efficiently or "is" the implicit bias of a first-order method of a convex optimization problem.

Unless stated otherwise, in the rest of the paper we write $\widehat{f}_n := \widehat{f}_n(\mathcal{D})$ where the dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ that consists of n identically distributed samples (i.i.d.) from the distribution above, and we use the notation $\mathbf{X} = (X_1, \ldots, X_n)$, $\mathbf{Y} = (Y_1, \ldots, Y_n)$.

We now introduce concepts from high-dimensional geometry and local theory of Banach spaces that we use to analyze the generalization properties of the minimum-norm interpolator (2).

2.2. Background from high-dimensional geometry

First, we recall a few classical definitions in convex geometry (cf. (Artstein-Avidan et al., 2015)). For any *symmetric* convex set K, i.e. $K = -K$ with a non-empty interior, we denote the Minkowski norm by $\|\cdot\|_K$, that is defined as

$$\|z\|_K := \inf\{r > 0 : z \in rK\}.$$

Further, consider the random coordinate projection of $\mathcal{F}$, namely the (random) convex and *symmetric* set

$$\mathcal{F}_n := \{f \in \mathcal{F} : (f(X_1), \ldots, f(X_n))\} \subset \mathbb{R}^n. \tag{3}$$

Note that for any vector $\mathbf{z} \in \mathbb{R}^n$, the following equality holds:

$$\|\widehat{f}_n(\mathbf{X}, \mathbf{z})\| = \|\mathbf{z}\|_{\mathcal{F}_n}. \tag{4}$$

Next, we define the (Gaussian) mean of a Minkowski norm $\|\cdot\|_K$ on some domain in $\mathbb{R}^n$ as

$$M(K) := \int_{\mathbb{R}^n} \|\boldsymbol{\xi}\|_K d\gamma_n,$$

where γ_n denotes the Gaussian measure on $\mathbb{R}^n$. Also, note that

$$n^{-1/2} M(K) \approx M_s(K) := \int_{\mathbb{S}^{n-1}} \|\boldsymbol{\xi}\|_K d\sigma_n,$$

where σ_n is the uniform measure on the unit sphere in $\mathbb{R}^n$, which we denote by $\mathbb{S}^{n-1}$.

For any convex body K, let

$$K^\circ := \{x \in \mathbb{R}^n : \sup_{y \in K} \langle x, y \rangle \leq 1\}$$

denote its polar body. We then write $M^*(K) = M(K^\circ)$ for the mean of the dual norm $\|\cdot\|_{K^\circ}$. The dual norm $\|\cdot\|_{K^\circ}$ is also referred to as the support function h_K of the original body K, as for a unit vector $\boldsymbol{\xi}$, it measures the distance of the supporting hyperplane of K from the origin in that direction. In our analysis below, these two quantities $M(K), M^*(K)$ and the quantity $\mathbb{E}M(\mathcal{F}_n)M^*(\mathcal{F}_n)$ (known as the MM^* estimate) play a key role in the statistical performance of the minimum norm solution.

First, it is not hard to verify that for any convex set $K \in \mathbb{R}^n$, we have that

$$1 \lesssim M_s(K) M_s^*(K).$$

At the same time, intuitively, $M_s^*(K) \cdot M_s(K) \lesssim 1$, when K does not lie in a low-dimensional subspace and is "balanced". Remarkably, the works of (Pisier, 1977; Figiel & Tomczak-Jaegermann, 1979) show that for any $K \subset \mathbb{R}^n$, there exists a linear transformation T such that

$$M_s(TK) M_s^*(TK) \lesssim \log(n).$$

Throughout this paper, we use the shorthands $M_n(\mathcal{F}) := \mathbb{E}_{\mathbf{X}} M(\mathcal{F}_n)$ for the mean norm averaged over $X_1, \ldots, X_n$, and $M_n^*(\mathcal{F}) := \mathbb{E}_{\mathbf{X}} M^*(\mathcal{F}_n)$ for the averaged dual norm, as well as $R_{MM*}(\mathcal{F}) := (M_n(\mathcal{F}) M_n^*(\mathcal{F}))/n$ for the MM^* estimate (cf. (Artstein-Avidan et al., 2015, §6.5)). As we will see below, our bounds deeply rely on these quantities.

Finally, we recall two fundamental quantities in statistical learning theory, (cf. (Bartlett & Mendelson, 2002)) the Gaussian and Rademacher complexity dependent on $(n, \mathcal{H}, \mathbb{P})$ for some $\mathcal{H} \subset \mathcal{B}(\mathcal{X})$.

Definition 2.1. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$, $\boldsymbol{\xi} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_n)$ and $\mathcal{H} \subset L_2(\mathbb{P})$. Then the *Gaussian complexity* of $\mathcal{H}$ is defined via

$$\mathcal{G}_n(\mathcal{H}) := \mathbb{E}_{\mathbf{X},\boldsymbol{\xi}} \sup_{h \in \mathcal{H}} | \langle \mathbf{h}, \boldsymbol{\xi} \rangle_n |$$

where $\langle \boldsymbol{\xi}, \mathbf{h} \rangle_n := n^{-1} \sum_{i=1}^n h(X_i)\xi_i$ is the Euclidean inner product in $\mathbb{R}^n$ scaled by $1/n$. The *Rademacher complexity* $\mathcal{R}_n(\mathcal{H})$ is defined analogously with $\boldsymbol{\xi} \sim U(\{-1, 1\}^n)$.

Finally, note that as $\mathcal{F}$ is symmetric, i.e. $\mathcal{F} = -\mathcal{F}$, $M_n^*(\mathcal{F})$ equals to the Gaussian complexity up to a scaling of $1/n$, i.e. $\mathcal{G}_n(\mathcal{F}) = \frac{M_n^*(\mathcal{F})}{n}$.

2.3. Curvature notions in Banach Spaces

We now recall some basic concepts of non-linear functional analysis, specifically from the local theory of Banach spaces (cf. (Artstein-Avidan et al., 2022, Chp. 5)). First, we introduce the definitions of uniformly convex (*UC*) and uniformly smooth (*US*) norms (cf. (Lindenstrauss & Tzafriri, 2013) and (Pisier, 2016, Thms 10.1 10.25)).

Our paper focuses on $UC(2)$ norms, yet to provide a full picture, we define $UC(q)$ norm for $q \in [2, \infty)$.

Definition 2.2 ($UC(q)$ norm)**.** A Banach space $(\|\cdot\|, \mathcal{B}(\mathcal{X}))$ is $q \in [2, \infty)$-uniformly convex, or, equivalently, the norm $\|\cdot\|$ is $UC(q)$ with constant $t > 0$, if for all $f, g \in \mathcal{B}(\mathcal{X})$

$$\|f\|^q + t\|g\|^q \leq \frac{\|f+g\|^q + \|f-g\|^q}{2}.$$

To give some intuition on the $UC(2)$-norm, one should think of $UC(2)$ as a lower bound on the minimal singular value of the Hessian on the unit ball $\mathcal{F}$ induced by this norm. Furthermore, 2-uniform convexity plays a key role in high dimensional geometry, in the sense that implies concentration of Lipschitz functionals, see the influential works of (Gromov & Milman, 1983; 1987).

Next, we define the "dual" notion of uniform convexity which is the uniform smoothness.

Definition 2.3 ($US(p)$ norm)**.** A Banach space $(\|\cdot\|, \mathcal{B}(\mathcal{X}))$ is $p \in (1, 2]$-uniformly smooth with constant $s > 0$, if for all $f, g \in \mathcal{B}(\mathcal{X})$

$$\|f\|^p + s\|g\|^p \geq \frac{\|f+g\|^p + \|f-g\|^p}{2}.$$

If a norm is $UC(q)$ with constant t, then its dual norm is $US(p)$ with constant $s = \Theta(t)$, where $1/p = 1 - 1/q$ (see (Lindenstrauss & Tzafriri, 2013)); and if a norm is$US(p)$, then its dual norm is $US(q)$ with constant $t = \Theta(s)$.

Finally, we remark that the celebrated Dvoretzky's theorem states that any unit ball $\mathcal{F}$ of an m-finite dimensional normed space has a $\Omega(\log(\epsilon \cdot m))$ section that is ϵ-close to a Euclidean ball (in terms of Banach-Mazur distance) – e.g., to a Hilbert space. As a consequence, the uniform convexity of any norm (or equivalently a Banach space) is at most with $q = 2$ and $t \leq 1/8 + o(1)$; and similarly $US(2)$ with parameter $s \geq 1/8 + o(1)$. To see this, any Hilbert space is $UC(2)$ and $US(2)$ with constants $1/8$ by the parallelogram law.

As uniform convexity and uniform smoothness are considered to be quite strong notions of curvature (for example, the ℓ_1-norm itself does not satisfy them); we also consider weaker notions of curvature in Banach spaces, which are known as (Gaussian) *type* $p \in [1, 2]$ and *cotype* $q \in [2, \infty)$ (Pisier, 2016; Ledoux & Talagrand, 2013) (see also Remark B.1 below).

In the following definitions, $\xi_1, \ldots, \xi_m$ are i.i.d. standard Gaussian random variables:

Definition 2.4 ($CO(q)$ norm)**.** A Banach space $(\|\cdot\|, \mathcal{B}(\mathcal{X}))$ is cotype $q \in [2, \infty]$- with constant $t > 0$, if for all $m \geq 1$ and $f_1, \ldots, f_m \in \mathcal{B}(\mathcal{X})$

$$t^q \cdot \sum_{i=1}^m \|f_i\|^q \leq \mathbb{E}_{\boldsymbol{\xi}} \| \sum_{i=1}^m \xi_i f_i \|^q.$$

Definition 2.5 ($T(p)$ norm)**.** A Banach space $(\|\cdot\|, \mathcal{B}(\mathcal{X}))$ is type $p \in [1, 2]$- with constant $s > 0$, if for all $m \geq 1$ and $f_1, \ldots, f_m \in \mathcal{B}(\mathcal{X})$

$$s^p \cdot \sum_{i=1}^m \|f_i\|^p \geq \mathbb{E}_{\boldsymbol{\xi}} \| \sum_{i=1}^m \xi_i f_i \|^p.$$

Hilbert spaces are both $T(2)$ and $CO(2)$ with constant 1, as they are 2-uniformly convex and smooth. However, the converse is also true, as Kapwien's theorem (cf. (Artstein-Avidan et al., 2022)) implies that

$$d_{BM}(\|\cdot\|, \ell_2) \lesssim st,$$

where $d_{BM}(\|\cdot\|, \ell_2)$ is the Banach-Mazur distance to a Hilbert space. Therefore, a space that is both $T(2)$ and $CO(2)$ with parameters $s, t > 0$ is *norm equivalent* to a Hilbert space.

We remark that if a norm is $T(p)$ with constant $s \geq 0$, then the dual norm is $CO(q)$ with constant $\Theta(s)$ (where $1/p + 1/q = 1$). However, and differently from uniform convexity, the converse is not true, i.e if a norm is $CO(q)$ with constant t, then the dual norm is not necessarily $T(p)$ with constant $\Theta(t)$.

The seminal work (Pisier, 1981) introduced the K-convexity constant, which we denote by $K_{\|\cdot\|}$ (see §B.1 for a formal definition).

Definition 2.6 (K-convexity constant, informal)**.** The constant $K_{\|\cdot\|}$ is the smallest constant that upper bounds the following: If a norm is $CO(q)$ with constant $t \geq 0$, then its dual norm is $T(q - 1/q)$ with constant smaller than $O(t \cdot K_{\|\cdot\|})$.

Finally, the following facts appear in (Klartag & Milman, 2008) and references within:

Fact 1. *If $(\|\cdot\|, \mathcal{B}(\mathcal{X}))$ is $UC(2)$ with constant $t \geq 0$, then it is also $CO(2)$ with constant $1/\sqrt{t}$.*

Fact 2. *If $(\|\cdot\|, \mathcal{B}(\mathcal{X}))$ be a $UC(2)$ with constant $t \geq 0$, then it is $T(1 + c(t))$ with constant $1/\sqrt{t}$.*

2.4. Assumptions

In this section, we introduce basic assumptions on our function class and covariate distributions, as well as the ground truth f^*. First, we make the following two assumptions on the ground truth:

Assumption 1. *The norm of f^* in both $L_2(\mathbb{P})$ and $\|\cdot\|$ is of order one, i.e. $\|f^*\|_{\mathbb{P}} \asymp \|f^*\| \asymp 1$. Furthermore, there exists a set $A \subset \mathcal{X}$ (that depends on f^*) of measure $1-n^{-2}$, such that $\|f^* \cdot \mathbb{1}_A\|_\infty \lesssim \Gamma \lesssim \log(n)$.*

This assumption rules out the (pathological) case when the norm of f^* (which we refer to as "true signal") vanishes with the number of samples n. In the high-dimensional setting, for example, such an assumption is crucial. Furthermore, we assume that in most of the space $(\mathcal{X}, \mathbb{P})$, the sup-norm of f^* equals its $L_2(\mathbb{P})$-norm (up to a $\log(n)$ factor).

Assumption 2 ("inductive" bias)**.** *For $f^* \in \mathcal{F}$ that satisfies Assumption 1, there exist absolute constants $c, c_1 \in (0, 1)$, such that with probability at least $1 - n^{-2}$ it holds that*

$$c_1 \leq \|\widehat{f}_n(\mathbf{X}, \mathbf{f^*})\| \leq c\|\widehat{f}_n(\mathbf{X}, \boldsymbol{\xi})\|. \tag{5}$$

This assumption captures the (necessary) properties for a minimum-norm interpolator to generalize well. First, the second inequality of (5) indicates that the norm has an "inductive bias" towards the ground truth f^*, in that it requires a larger norm to interpolate pure noise rather than the underlying "true signal" $f^* \in \mathcal{F}$. The first inequality (5) ensures that the minimum-norm solution of noiseless samples has a positive norm: without such an assumption, the minimal norm solution would not generalize well even on noiseless samples.

Finally, we introduce a few regularity assumptions on both the function class and our distribution.

We first assume that the space $(\mathcal{F}, \mathbb{P})$ satisfies a small-ball property (Mendelson, 2014). This assumption ensures that the $L_1(\mathbb{P})$ and $L_2(\mathbb{P})$ error of the minimum-norm solution are similar (up to an absolute multiplicative constant) so that we can study the Mean Squared Error (MSE) of the minimum-norm interpolator, defined as $\mathbb{E}\|\widehat{f}_n - f^*\|_{\mathbb{P}}^2$, rather than the $L_1(\mathbb{P})$ error.

Assumption 3. *There exist universal constants $c_1, c_2 \in (0, 1)$ such that:*

$$\forall f, g \in \mathcal{F} \quad \mathbb{P}_X(|f(X) - g(X)| \geq c_1\|f - g\|_{\mathbb{P}}) \geq c_2.$$

Note that the class of linear functions with sub-Gaussian covariates satisfies this property (see (Mendelson, 2017) for further details).

The second assumption ensures that the geometry of $\mathcal{F}_n$ is similar with high probability, which is weaker than assuming concentration. This additional assumption is essential, as we do not assume that $\mathcal{F}$ is uniformly bounded or has a bounded envelope or any strong structure on the covariates.

Assumption 4. *With probability at least $1 - n^{-2}$ over $X_1, \ldots, X_n$, the following inequalities hold:*

$$\sup_{f \in \mathcal{F}} \int f d(\mathbb{P}_n - \mathbb{P}) \asymp \mathbb{E}_{\mathbf{X}} \sup_{f \in \mathcal{F}} \int f d(\mathbb{P}_n - \mathbb{P}),$$

and the random set $\mathcal{F}_n$ satisfies

$$M(\mathcal{F}_n) \asymp M_n(\mathcal{F}) \quad \text{and } M^*(\mathcal{F}_n) \asymp M_n^*(\mathcal{F}).$$

2.5. Two Classical Examples

For concreteness, we now provide two classical examples that satisfy all of the above assumptions and that could be analyzed with our framework (we refer to (Ledoux & Talagrand, 2013; Pisier, 1999) for more examples).

The first example is the standard linear regression model with respect to general ℓ_p norms for $p \in (1, 2]$.

Example 1 (Linear regression in terms of ℓ_p norm)**.** Consider the setting where $X_1, \ldots, X_d$ are i.i.d. sub-Gaussian random variables with zero mean, variance one and constant $L > 0$, i.e. for all $t \geq 0$

$$\mathbb{P}(|X| \geq t) \leq \exp(-t^2/L^2),$$

and X has a continuous density upper bounded by M. We consider the Banach spaces of linear functions equipped with ℓ_p-norms

$$\mathcal{F} = \ell_{p,d} := \{w \in \mathbb{R}^d : \|w\|_p \leq 1\}.$$

When $p \in (1, 2]$, the space $\ell_{p,d}$ is $UC(2)$ with constant $t = (p-1)/8$ and $US(p)$ with constant $s = 1/p$. Next, when $p \in (2, \infty)$, the $\ell_{p,d}$ is $UC(p)$ with constant $t = 2^p/p$, and $US(2)$ with $s = (p-1)/2$. Finally, when $p \in [1, 2]$, $\ell_{p,d}$ is $CO(2)$ with an absolute constant and $T(p)$, and for $p \in [2, \infty]$, $\ell_{p,d}$ is $T(2)$ with constant of order $\min(p, \log(d))$ and $CO(p)$.

The second example is Sobolev spaces that play a key role in non-parametric statistics (cf. (Tsybakov, 2003; Giné & Nickl, 2021)); in particular, it illustrates that our model can be applied to a space of non-linear functions.

Example 2 (Sobolev spaces)**.** Fix a domain $\Omega \subset \mathbb{R}^d$ and let $\mathbb{P} = \mathrm{Unif}(\Omega)$, $p \in [1, \infty]$, and $k \in \mathbb{N}$. For a fixed multi-index $|\alpha| \leq k$, we define $D^\alpha(f)$ to be the mixed partial derivative in terms of the multi-index α, that is $D^\alpha(f) = \frac{D^{|\alpha|} f}{D^{\alpha_1} f \ldots D^{\alpha_k} f}$. The space $W^{k,p}(\Omega)$, is defined as the space of all functions that have k partial (weak) derivatives that lie in $L_p(\mathbb{P})$, also known as a Sobolev space. Note that this is a Banach space with respect to the norm

$$\|f\|_{k,p} := \left(\sum_{|\alpha| \leq k} \|D^\alpha(f)\|^p_{L_p(\mathbb{P})} \right)^{1/p}.$$

Clearly, $(\|\cdot\|_{k,p}, W^{k,p}(\Omega))$ is *isometric* to $(\ell_p, \mathbb{R}^{\sum_{i=0}^k \binom{d}{i}})$. Therefore, its q-uniform convexity and p-smoothness (and similarly its p-type and q-cotype) are preserved under isometry, they are equal to the ones of $(\ell_p, \mathbb{R}^{\sum_{i=0}^k \binom{k}{d}})$ (see Example 1 above), and therefore it is captured by our model.

3. Main Results

In this section, we state the main results of this manuscript. In the first theorem, the expectation is conditioned over an event (that is defined as the intersection of the events of our assumptions) that holds with probability at least $1 - n^{-2}$.

For any regression estimate, we typically aim to establish bounds on its risk or MSE. A standard approach to analyze this quantity is via a decomposition, by observing

$$\begin{aligned} \mathbb{E}_{\mathcal{D}} \|\widehat{f}_n - f^*\|^2_{\mathbb{P}} &= \underbrace{\|\mathbb{E}\widehat{f}_n - f^*\|^2_{\mathbb{P}}}_{B^2(\widehat{f}_n)} + \underbrace{\mathbb{E}\|\widehat{f}_n - \mathbb{E}\widehat{f}_n\|^2_{\mathbb{P}}}_{V(\widehat{f}_n)} \quad (6) \\ &= \underbrace{B^2(\widehat{f}_n) + \mathrm{Var}\left[\mathbb{E}\left[\widehat{f}_n|\mathbf{X}\right]\right]}_{T_1} + \underbrace{\mathbb{E}\mathrm{Var}\left[\widehat{f}_n|\mathbf{X}\right]}_{T_2}. \end{aligned}$$

For regularized estimators (or in the underparameterized case), it is more natural to analyze the bias and the variance separately to bound the MSE. Roughly speaking, in that case, the bias can be interpreted as an approximation error of f^* induced by the function space and the variance as the effect of noise. However, in overparameterized models, i.e. when $\widehat{f}_n$ interpolates the observations $\mathbf{Y}$, it is more natural to split the variance term and analyze the last term separately from the first two terms. In particular, the first term T_1 can be interpreted as characterizing the "structural" error, while T_2 captures the "noise effect" error. To see this, we can rewrite T_1 in (6) as follows

$$\mathbb{E}\|\mathcal{P}_{f^*}(\mathbb{E}_{\boldsymbol{\xi}}\widehat{f}_n|\mathbf{X}) - f^*\|^2_{\mathbb{P}} + \mathbb{E}\|\mathcal{P}_{(f^*)^\perp}(\mathbb{E}_{\boldsymbol{\xi}}\widehat{f}_n|\mathbf{X})\|^2_{\mathbb{P}}, \quad (7)$$

where $\mathcal{P}_{f^*}$ is the projection on f^* in term of $L_2(\mathbb{P})$. Let us briefly explain why this new decomposition (7) is useful for analyzing the minimum-norm solution; the first term in Eq. (7) measures how much "energy" minimum norm interpolator retains from the original signal f^*; i.e , it measures the shrinkage of the signal f^* due to undersampling. In overparameterized models, this term is typically non-zero. Then, as $\mathbb{E}_{\boldsymbol{\xi}}\widehat{\mathbf{f}}_\mathbf{n}|\mathbf{X} = \mathbf{f^*}$, some of the energy must have emerged from an uncorrelated function $f_\mathbf{X} \perp f^*$, and the second term in Eq. (7) measures how much energy was added from $f_\mathbf{X}$ (in expectation) to the minimum norm solution. Finally, T_2, measures the amount of energy that was added to the minimum norm solution due to the noise in the observations.

We now present our main results that separately control the "structural" and "noise" effect error.

3.1. Upper bound on the "structural" error

The following theorem is our "unlocalized" upper bound for $UC(2)$-norms on T_1:

Theorem 3.1. *Assume that $(\|\cdot\|, \mathcal{B}(\mathcal{X}))$ is $UC(2)$ with constant $t > 0$. Then, under Assumptions 1-4, the minimum-norm solution $\widehat{f}_n$ satisfies*[2]

$$\mathrm{Var}\left[\mathbb{E}\widehat{f}_n|\mathbf{X}\right] + B^2(\widehat{f}_n) \lesssim \frac{R_{MM^*} \cdot \mathcal{G}_n(\mathcal{F})}{t}. \quad (8)$$

Furthermore, if $(\|\cdot\|, \mathcal{B}(\mathcal{X}))$ is $US(p)$ for some $p \in (1, 2]$ with constant $s > 0$, the upper bound can be improved to

$$\frac{R^{2-p}_{MM^*} \cdot C_{SB} \cdot \mathcal{G}_n(\mathcal{F})^p}{t}.$$

Note that under 2-uniform convexity, Theorem 3.1 aligns with the standard result based on the "unlocalized" risk bound on (cf. (van de Geer, 2000; Chatterjee, 2014)) for the MSE of Empirical Risk Minimization (ERM) with squared loss over the function class $\mathcal{F}$, i.e

$$\sup_{f^* \in \mathcal{F}} \mathbb{E}_{\mathcal{D}} \|\bar{f}_n - f^*\|^2_{\mathbb{P}} \lesssim \mathcal{G}_n(\mathcal{F})$$

where $\bar{f}_n$ is given by

$$\bar{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n (Y_i - f(X_i))^2.$$

We refer to (Bartlett et al., 2005; van de Geer, 2000) for more information on localized bounds for ERM. Further, we *conjecture* that the $UC(2)$ assumption in Theorem 3.1 *can be relaxed to* $CO(2)$ – we believe that improving to $CO(2)$ will require a major technical advances, which will

[2] up to a multiplicative constant that depends on the constants that appear in the assumptions above.

be of independent interest. We discuss this in further detail in the extended version of this manuscript. Moreover, we believe that without additional regularity assumptions, the upper bound of Theorem 3.1 is sharp.

Finally, we would like to emphasize that although it seems that the smoothness of the function class $\mathcal{F}$ *reduces* T_1, actually, the opposite is correct, as higher smoothness also implies higher Gaussian complexity (see also §4.1).

On a Matching Lower Bound on T_1: In the extended version of this paper, we provide a localized version of Theorem 3.1. For example, under additional regularity assumption, an improved bound on the error can be of order (up to a multiplicative constant that depends on t, s, R_{MM^*})

$$\max\{\mathcal{G}_n(\mathcal{F})^{2p}, 1/n, \mathbb{E}\|\widehat{f}_n(\mathbf{X}, \mathbf{f^*}) - f^*\|_{\mathbb{P}}^2\}. \quad (9)$$

In the case of Example 1, this bound can be attained. In (Donhauser et al., 2022), it was shown that when $w^* \equiv (1, 0, \ldots, 0)$, it holds that

$$\mathrm{Var}\left[\mathbb{E}\widehat{f}_n|\mathbf{X}\right] + B^2(\widehat{f}_n) \lesssim \tilde{O}(d^{2p-2}/n^{2p}).$$

As

$$M_n(\ell_{p,d}) \lesssim \sqrt{n}d^{1/p-1} \text{ and } R_{MM^*}(\ell_{p,d}) \lesssim \log(d),$$

it aligns with our bound. Finally, we believe that (9) is a sharp lower bound on T_1 under nice enough regularity assumptions.

3.2. Lower bound on the "noise effect" error

For the next theorem, recall that $K_{\|\cdot\|}$ is the K-convexity constant of $(\|\cdot\|, \mathcal{B}(\mathcal{X}))$ as informally defined in 2.6 (see §B.1 for the formal definition). We emphasize that in our setting, this constant is *at most* logarithmic in the number of samples. Further, for $UC(2)$-norms for example, it is at most $O(1/\sqrt{t})$, and under mild assumptions, logarithmic in the "intrinsic" dimension of the class $\mathcal{F}$. To state the theorem, for any fixed vector in $\mathbf{v} \in \mathbb{R}^n$ and $r \geq 0$, we define

$$\Psi_n(\mathbf{v}, r) := \mathrm{Med}\left[\min_{\{f \in r\mathcal{F}: \mathbf{f}=\mathbf{v}\}} \|f\|_{\mathbb{P}}\right],$$

which is the *median* (over $\mathbf{X}$) of the $L_2(\mathbb{P})$ minimum norm solution in $r \cdot \mathcal{F}$ that interpolates $\mathbf{v}$ on $\mathbf{X}$, and use the shorthand $\mathbf{e}_1 = (1, 0, \ldots, 0) \in \mathbb{R}^n$.

Theorem 3.2 (Reverse Efron-Stein's for $CO(2)$ norms)**.** *Assume that $(\|\cdot\|, \mathcal{B}(\mathcal{X}))$ is $CO(2)$ with constant $t > 0$. Then, under Assumptions 1-2, the minimum-norm solution $\widehat{f}_n$ satisfies the following:*

$$\mathbb{E}\mathrm{Var}\left[\widehat{f}_n|\mathbf{X}\right] \gtrsim n \cdot \Psi_n\left(\mathbf{e_1}, C\frac{K_{\|\cdot\|} \cdot M_n(\mathcal{F})}{t\sqrt{n}}\right)^2, \quad (10)$$

where $K_{\|\cdot\|}$ is the K-convexity constant, and $C > 0$ is an absolute constant.

This result can be interpreted as a "reverse-type" Efron-Stein inequality for minimum norm interpolators as it lower bounds the variance in terms of the marginal contribution of each point X_i. Intuitively, it reflects how the expected conditional variance (in terms of $L_2(\mathbb{P})$) is always lower bounded by the $L_2(\mathbb{P})$ norm required to interpolate "appropriately scaled" 1-"spikes".

3.2.1. Tightness for ℓ_p-linear regression

In this part, we prove that under the model of ℓ_p-regression with $p \in [1, 2]$ in Example 1 above, Theorem 3.2 provides a tight bound. We remark that under the additional assumption of Gaussian covariates, Example 1 is the only known model (that is not a Hilbert space) for which sharp bounds on the MSE of the minimum norm interpolator exist (see (Donhauser et al., 2022)).[3]

First, let us apply Theorem 3.2 to Example 1 with isotropic sub-Gaussian covariates, and for simplicity, consider the case of $w^* \equiv 0$. Note that as with high probability $(X_1, \ldots, X_d)$ is dense, we have for any interpolator[4] w of $\mathbf{e}_1$, that $\|w\|_2^2 \gtrsim 1/d$. Therefore, it holds for any $1 \leq p \leq 2$ that

$$\mathbb{E}\mathrm{Var}\left[\widehat{w}_p|\mathbf{X}\right] \gtrsim n/d,$$

aligning with the lower bound in (Muthukumar et al., 2020).

(Donhauser et al., 2022) *implicitly* proved the converse *under* Gaussian covariates. Our next result extends their bound to cover sub-Gaussian covariates, emphasizing that the bound of Theorem 3.2 can be sharp for $UC(2)$-norms, and that it is not specific to the Gaussianity of the data. As such, it is the first step to overcoming a fundamental limitation of some existing literature on benign overfitting that heavily uses the Gaussianity of the covariates.

Theorem 3.3 (Expected Conditional Variance in Linear Models)**.** *Consider the model of Example 1. Then, when $p \in [1 + \frac{C \log\log\log(d)}{\log\log(d)}, 2]$, and $d \gtrsim n \cdot \log(n)$, the following holds when $w^* \equiv 0$:*

$$\mathbb{E}\,\mathrm{Var}\left[\widehat{w}_p|\mathbf{X}\right] \lesssim \log(d)^2 \cdot \frac{n}{d}.$$

Note that when $w^* \equiv 0$, we have

$$\mathbb{E}\,\mathrm{Var}\left[\widehat{w}_p|\mathbf{X}\right] = \mathbb{E}\|\widehat{w}_p - w^*\|_2^2 \lesssim \frac{\log(d)^2 n}{d}.$$

[3] In the extended version of this manuscript, we prove this for i.i.d. sub-Gaussian entries with bounded density from above and below.

[4] The K-convexity constant does not play a key role as we are in a linear setting. In the case of Example 2, the $K_{\|\cdot\|}$ plays a key role in this bound – as there is a sequence of functions f_n such that $\|f_n\| \to \infty$ that interpolate $(1, 0, \ldots, 0)$ that also $\|f_n\|_{\mathbb{P}} \to 0$.

On a matching upper bound on T_2 : As we see in Theorem 3.3, Theorem 3.2 is tight in the case of Example 1, when $p \geq 1 + \frac{C \log\log\log d}{\log\log d}$. However, for $p = 1$ and $w^* \equiv 0$, we know from (Wang et al., 2022) that

$$\mathbb{E}\,\mathrm{Var}\left[\widehat{w}_1|\mathbf{X}\right] \asymp \frac{1}{\log(d/n)}.$$

On the other hand, applying Theorem 3.2 directly only provides a lower bound $\Omega(n/d)$. Therefore, one may wonder what the tight upper bound for the term T_2 should be. We conjecture that *in contrast* to the first term T_1, 2-uniform convexity (or a similar property) is essential for the bound of Theorem 3.2 to be sharp. We believe that under sufficient regularity assumptions (additionally to $UC(2)$), Theorem 3.2 is tight. We discuss this question in further detail in the extended version of this manuscript.

Before we end this section, we would like discuss the importance of the Gaussian noise assumption for our theorems.

Remark 3.1 (Beyond Gaussian noise). We would like to emphasize the Gaussian noise (or very well structured noise) assumption is essential in all our results. In the Theorem 3.1, the R_{MM^*} ratio is tightly connected to Gaussian, or at least Rademacher noise. The K-convexity argument in Theorem 3.2 is only valid for Gaussian or Rademacher noise. Finally, the analysis of Theorem 3.3 is only valid for Gaussian noise; since we do not assume Gaussian covariates, it is essential that the noise vector is rotationally invariant, and since the model assumes i.i.d. noise, it must be Gaussian.

4. Discussion

In this part, we discuss the consequences and subtleties that are revealed by our analysis.

4.1. On the Tradeoff Between Type p and Gaussian Complexity

As briefly discussed in §3.1, Theorem 3.1 may suggest that the upper bound is improving under $US(p)$ with $p > 1$ with constant s. We now provide more details on why this is not necessarily true. First, we argue that when the class contains m "well-separated" signals in $L_2(\mathbb{P})$, there is a price to be paid, as the Gaussian complexity of the class is at least of order $(s^{-2}m)^{1-1/p}/\sqrt{n}$. Therefore, larger p does not automatically lead to an improved upper bound in Theorem 3.1.

First, let m be the maximal number of elements $f_1, \ldots, f_m \in \mathcal{F}$ that are orthogonal and well-separated, namely

$$\forall 1 \leq i < j \leq m \quad \langle f_i, f_j \rangle_{\mathbb{P}} = 0 \quad \|f_i - f_i\|_{\mathbb{P}} \gtrsim 1,$$

where $\langle \cdot, \cdot \rangle_{\mathbb{P}}$ denotes the inner product with respect to $L_2(\mathbb{P})$. Then, if the underlying Banach space is type $p \geq 1$, then the Gaussian complexity is lower bounded by

$$\mathcal{G}_n(\mathcal{F}) \gtrsim \sqrt{\frac{\max\{(s^{-1} \cdot m)^{2-2/p}, \log(m)\}}{n}}. \tag{11}$$

In particular, for type 2 spaces, we have $\mathcal{G}_n(\mathcal{F}) \gtrsim \sqrt{m/n}$. In contrast, for a space that is both cotype 2 and type 1, such as the ℓ_1-ball, we only pay a logarithmic price in the number of different signals which is tight in Example 1.

To simplify the presentation, we now provide the argument for type 2 spaces, while the lower bound (11) for general type p follows analogously. In fact, in a $T(2)$ space with constant s, we have

$$\left(\mathbb{E}\left\|\frac{\sum_{i=1}^m \xi_i f_i}{\sqrt{m}}\right\|\right)^2 \asymp \frac{\mathbb{E}\|\sum_{i=1}^m \xi_i f_i\|^2}{m} \lesssim s^2,$$

where we used Kahane's inequality (cf. (Milman & Schechtman, 1986)) and that $\|f_i\| \leq 1$ for all $1 \leq i \leq m$. Moreover, by the orthogonality of $\{f_1, \ldots, f_m\}$, we have that

$$\mathbb{E}\|\frac{1}{\sqrt{m}}\sum_{i=1}^m \xi_i f_i\|_{\mathbb{P}} \asymp s.$$

Clearly, it implies that the function class $\mathcal{F}$ contains a set of $2^{\Omega(m)}$ functions, constructed by choosing random signs in the last equality, that are $\Theta(s^{-1})$ far from each other (in terms of $L_2(\mathbb{P})$). Hence, due to the small ball assumption (see Assumption 3), we know that for each two functions $f, g \in \mathcal{F}$ that

$$\|f - g\|_{\mathbb{P}_n} \geq c_4 \|f - g\|_{\mathbb{P}} \gtrsim s^{-1}$$

with probability of at least $1 - \exp(-cn)$. By the union bound, we obtain that there are $2^{\Omega(m)}$ mixture signals that are well-separated and belong to $\mathcal{F}_n$ with high probability. Equivalently, $\mathcal{F}_n$ contains a well-separated set with cardinality at least $2^{\Omega(m)}$. By Sudakov's minoration inequality (Ledoux & Talagrand, 2013), it immediately follows that

$$\mathcal{G}_n(\mathcal{F}) \gtrsim s^{-1} \cdot \sqrt{\frac{m}{n}}.$$

Therefore, we conclude that additional smoothness increases Gaussian complexity, and reduces the possibility of the benign over-fitting way for *rich* function classes.

4.2. On the role of the covariate distribution

Note that our bounds are general in their nature, as they hold for arbitrary i.i.d. covariates. In contrast, many previous works (see, e.g., (Bartlett et al., 2020; Tsigler & Bartlett, 2023; Koehler et al., 2021; Donhauser et al., 2022; Liang & Rakhlin, 2020; Liang et al., 2020)) deeply rely on additional

structure on the covariates such as isotropic Gaussianity (cf. (Donhauser et al., 2022)), or low-stable rank covariance matrix (cf. (Bartlett et al., 2020)).

We now briefly discuss how additional assumptions on covariate structure affect the final bounds obtained through our general results in Theorem 3.1. For one, the Gaussian complexity may depend on the covariates. Further, the covariate distribution affects the MM^*-estimate (i.e, R_{MM^*}). Moreover, additional structural assumptions on covariates may significantly *improve* the uniform convexity and smoothness constants of a "typical" norm induced by the set $\mathcal{F}_n$. Therefore, since the geometry of the minimum norm interpolator deeply depends on the norm induced by $\mathcal{F}_n \subset \mathbb{R}^n$, special structure of the covariates could also significantly reduce the MSE.

A notable example is high-dimensional linear regression with *isotropic* Gaussian covariates. In such a model, one should view $\mathcal{F}_n$ as

$$\mathcal{F}_n \approx \sqrt{d} \cdot P(\mathcal{F})$$

where $\mathcal{F} \subset \mathbb{R}^d$ is convex and symmetric, and $P \sim \mathrm{Unif}(\mathrm{Gr}(n,d))$ is a random projection from $\mathbb{R}^d$ to $\mathbb{R}^n$ (here $\mathrm{Gr}(n,d)$ denotes the Grassmanian manifold).

When $n \ll d$, the classical Dvoretzky-Milman's theorem and its many variants (cf. (Vershynin, 2018) and (Artstein-Avidan et al., 2015, Chps. 5,7)) show that such random projections usually have a small R_{MM^*} estimate. Furthermore, the norm induced by $\mathcal{F}_n$ may have better uniform smoothness and convexity constant than the norm induced by $\mathcal{F}$. However, without special structural assumptions, such improvements are not possible.

4.3. On the role of 2-uniform convexity and the MM^* estimate

Finally, we would like to provide some rough intuition of our approach that led to the results in §3. Consider the symmetric convex set in $\mathcal{B}(\mathcal{X})$ defined via

$$\mathcal{F}^\perp := \mathcal{P}_{(f^*)^\perp}\mathcal{F} \tag{12}$$

that contains functions from $\mathcal{F}$ orthogonal to f^* in terms of $L_2(\mathbb{P})$, and denote the projected set $\mathcal{F}^\perp$ on $X_1,\ldots,X_n$ by $\mathcal{F}_n^\perp$, and its Minkowski norm on $\mathbb{R}^n$ by $\|\cdot\|_{\mathcal{F}_n^\perp}$. Consider $\delta \cdot f^*$, and some $\delta \le c_1$. Then, by Anderson's lemma (cf. (Wainwright, 2019)), it holds that

$$\mathbb{E}\|\boldsymbol{\xi} + \delta \cdot \mathbf{f^*}\|_{\mathcal{F}_n^\perp} \ge \mathbb{E}\|\boldsymbol{\xi}\|_{\mathcal{F}_n^\perp},$$

where $\boldsymbol{\xi}$ is an isotropic Gaussian. Clearly, we would like the last inequality to be strict, as we would hope that this random norm is sensitive to the fact that there is an additional signal rather than pure noise. One hopes that the sampled signal f^* that is uncorrelated to all functions in $\mathcal{F}^\perp$, would imply that $\|\boldsymbol{\xi}+\delta\cdot\mathbf{f^*}\|_{\mathcal{F}_n^\perp}$ behaves as $\|\boldsymbol{\xi}'\|$, where, $\boldsymbol{\xi}' \sim N(0,1+\delta^2)$. In other words, $\mathbf{f^*}$ "behaves" as additional independent noise.

Let us consider the ℓ_p-linear regression setting with isotropic Gaussian covariates and $w^* = (1,0,\ldots,0)$ (and note that $\mathbf{f} = (w^*X_1,\ldots,w^*X_n)$. Then, one can then verify that for any $\delta \ge 0$

$$\mathbb{E}_{\boldsymbol{\xi},\mathbf{X}}\|\boldsymbol{\xi} + \delta\cdot\mathbf{f^*}\|^2_{\mathcal{F}_n^\perp} \ge (1+c_1\delta^2)\mathbb{E}\|\boldsymbol{\xi}\|^2_{\mathcal{F}_n^\perp},$$

where we used that $\mathbf{f^*} \sim N(0,\delta^2 I_n)$ is independent from both $\boldsymbol{\xi}$ and the norm $\|\cdot\|_{\mathcal{F}_n^\perp}$. Remarkably, this fact is not special to linear models. The proof of Theorem 3.1 shows that such a property follows from 2-uniform convexity. Specifically, it holds in the general case that

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\xi}}\|\boldsymbol{\xi} + \delta\cdot\mathbf{f^*}\|^2_{\mathcal{F}_n^\perp} \ge \left(1 + \frac{c\delta^2 R_{MM^*}^{-2}}{t}\right)\mathbb{E}\|\boldsymbol{\xi}\|^2_{\mathcal{F}_n^\perp}.$$

Therefore, if R_{MM^*} is bounded, we have a similar behavior as in the linear setting with Gaussian covariates.

5. Conclusions

In this work, we study the behavior of minimum norm interpolators using a novel approach based on the local theory of Banach spaces that does not rely on the linearity of the function class. Instead, we show how such geometric properties of the underlying space are tightly connected to the phenomenon of benign overfitting. We use 2-uniform convexity of norms to provide insights on the structural and noise-induced error of minimum norm interpolators.

In a future extended version of this work, we aim to provide a more complete picture of our framework than is possible in this short conference version. For one, we will prove a "localized" version of Theorem 3.1 that yields improved bound under additional regularity assumptions that align with the bounds on ℓ_p-linear regression. Further, we will provide a full picture of ℓ_p linear regression with sub-Gaussian covariates. Finally, we aim to derive sharp bounds for the minimum-norm interpolator in Sobolev spaces for $1 \le p \le 2$.

Impact Statement

This paper presents work whose goal is to advance the field of mathematical statistics and and statistical learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

GK is supported by the Computer Science Department at ETH Zürich. GK acknowledges Inga Miller for the inspiration for this paper; secondly his close friend Konstantin Donhauser for exposing him to the limitation of Gaussian covariates in the benign overfitting literature and the many useful discussions; and finally for Eli Putterman, Sven Wang, Rafal Latala, and especially Beatrice-Helen Vritsiou for the useful discussions.

PB's research of leading to these results is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 101001677-ISOPERIMETRY).

We also acknowledge Julia Kostin for improving the final version of this manuscript, and to Maud Szusterman for introducing GK and PB to each other.

References

Aerni, M., Milanta, M., Donhauser, K., and Yang, F. Strong inductive biases provably prevent harmless interpolation. 2023.

Artstein-Avidan, S., Giannopoulos, A., and Milman, V. D. *Asymptotic geometric analysis, Part I*, volume 202 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2015.

Artstein-Avidan, S., Giannopoulos, A., and Milman, V. D. *Asymptotic Geometric Analysis, Part II*, volume 261 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, 2022. ISBN 9781470467777.

Aubrun, G. and Szarek, S. J. *Alice and Bob meet Banach*, volume 223. American Mathematical Soc., 2017.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local Rademacher complexities. *The Annals of Statistics*, 33 (4):1497–1537, 2005.

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

Belkin, M., Hsu, D. J., and Mitra, P. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31, 2018a.

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, 2018b.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1611–1619. PMLR, 2019.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Brazitikos, S., Giannopoulos, A., Valettas, P., and Vritsiou, B.-H. *Geometry of isotropic convex bodies*, volume 196. American Mathematical Soc., 2014.

Chatterjee, S. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42(6):2340–2381, 2014.

Chinot, G., Löffler, M., and van de Geer, S. On the robustness of minimum norm interpolators and regularized empirical risk minimizers. *arXiv preprint arXiv:2012.00807*, 2020.

Donhauser, K., Ruggeri, N., Stojanovic, S., and Yang, F. Fast rates for noisy interpolation require rethinking the effects of inductive bias. *arXiv preprint arXiv:2203.03597*, 2022.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004.

Enflo, P. On the nonexistence of uniform homeomorphisms between l p-spaces. *Arkiv för matematik*, 8(2):103–105, 1970a.

Enflo, P. Uniform structures and square roots in topological groups. *Israel Journal of Mathematics*, 8:230–252, 1970b.

Figiel, T. and Tomczak-Jaegermann, N. Projections onto hilbertian subspaces of banach spaces. *Israel Journal of Mathematics*, 33:155–171, 1979.

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

Giné, E. and Nickl, R. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

Gordon, Y. On milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1986–87*, pp. 84–106. Springer, 1988.

Gromov, M. and Milman, V. D. A topological application of the isoperimetric inequality. *American Journal of Mathematics*, 105(4):843–854, 1983.

Gromov, M. and Milman, V. D. Generalization of the spherical isoperimetric inequality to uniformly convex banach spaces. *Compositio mathematica*, 62(3):263–282, 1987.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018.

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

Ivanisvili, P., van Handel, R., and Volberg, A. Rademacher type and enflo type coincide. *Annals of mathematics*, 192 (2):665–678, 2020.

Klartag, B. and Milman, E. On volume distribution in 2-convex bodies. *Israel Journal of Mathematics*, 164: 221–249, 2008.

Koehler, F., Zhou, L., Sutherland, D. J., and Srebro, N. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. *Advances in Neural Information Processing Systems*, 34:20657–20668, 2021.

Koltchinskii, V. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Lecué, G. and Shang, Z. A geometrical viewpoint on the benign overfitting property of the minimum ℓ_2-norm interpolant estimator. *arXiv preprint arXiv:2203.05873*, 2022.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

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

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

Lindenstrauss, J. and Tzafriri, L. *Classical Banach spaces II: function spaces*, volume 97. Springer Science & Business Media, 2013.

Maurey, B. and Pisier, G. Séries de variables aléatoires vectorielles indépendantes et propriétés géométriques des espaces de banach. *Studia Mathematica*, 58(1):45–90, 1976.

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

Mendel, M. and Naor, A. Metric cotype. *Annals of Mathematics*, pp. 247–298, 2008.

Mendelson, S. Learning without concentration. In *Conference on Learning Theory*, pp. 25–39, 2014.

Mendelson, S. Extending the scope of the small-ball method. *arXiv preprint arXiv:1709.00843*, 2017.

Milman, V. D. and Schechtman, G. *Asymptotic theory of finite dimensional normed spaces: Isoperimetric inequalities in riemannian manifolds*, volume 1200. Springer Science & Business Media, 1986.

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

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

Oravkin, E. and Rebeschini, P. On optimal interpolation in linear regression. *Advances in Neural Information Processing Systems*, 34:29116–29128, 2021.

Pisier, G. Un nouveau théorème de factorisation. *CR Acad. Sci. Paris*, 285:715–718, 1977.

Pisier, G. On the duality between type and cotype. In *Martingale Theory in Harmonic Analysis and Banach Spaces: Proceedings of the NSF-CBMS Conference Held at the Cleveland State University, Cleveland, Ohio, July 13–17, 1981*, pp. 131–144. Springer, 1981.

Pisier, G. *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press, 1999.

Pisier, G. Probabilistic methods in the geometry of banach spaces. In *Probability and Analysis: Lectures given at the 1st 1985 Session of the Centro Internazionale Matematico Estivo (CIME) held at Varenna (Como), Italy May 31–June 8, 1985*, pp. 167–241. Springer, 2006.

Pisier, G. *Martingales in Banach spaces*, volume 155. Cambridge University Press, 2016.

Rudelson, M. and Vershynin, R. Small ball probabilities for linear images of high-dimensional distributions. *International Mathematics Research Notices*, 2015(19):9594–9617, 2015.

Schütt, C. Entropy numbers of diagonal operators between symmetric banach spaces. *Journal of approximation theory*, 40(2):121–128, 1984.

Shamir, O. The implicit bias of benign overfitting. In *Conference on Learning Theory*, pp. 448–478. PMLR, 2022.

Szarek, S. J. Spaces with large distance to ℓ_∞^n and random matrices. *American Journal of Mathematics*, 112(6):899–942, 1990.

Thrampoulidis, C., Oymak, S., and Hassibi, B. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pp. 1683–1709. PMLR, 2015.

Tsigler, A. and Bartlett, P. L. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24 (123):1–76, 2023.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer, 2003.

van de Geer, S. A. *Empirical Processes in M-estimation*, volume 6. Cambridge University Press, 2000.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

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

Wang, G., Donhauser, K., and Yang, F. Tight bounds for minimum in lone-norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics*, pp. 10572–10602. PMLR, 2022.

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

A. Proofs

A.1. Proofs of Theorem 3.1

Recall that for abbreviation, we use the notation $\widehat{f}_n(\mathbf{z}) := \widehat{f}_n(\mathbf{X}, \mathbf{z})$, for every $\mathbf{z} = (z_1, \ldots, z_n)$. For a convex set with a non-empty interior (convex body) K, $\|\cdot\|_K$ denotes the Minkowski functional.

Next, recall the definition of $\mathcal{F}_n$ in (3), and notice that on each realization of the data $\mathbf{X}$ (and therefore of $\mathcal{F}_n$), for any $\mathbf{z} \in \mathbb{R}^n$, we have that

$$\|\widehat{f}_n(\mathbf{z})\| = \|\mathbf{z}\|_{\mathcal{F}_n} := \|\mathbf{z}\|_n.$$

Also, for simplicity, we assume that $\Gamma \leq C_1$, for some absolute constant $C_1 \geq 0$, and that $\|f^*\|_{\mathbb{P}} = 1$.

Step I: Symetrization Consider the operator $\mathcal{P}_{(f^*)^\perp} : L_2(\mathbb{P}) \to L_2(\mathbb{P})$, defined via

$$f \mapsto f - \langle f^*, f\rangle_{\mathbb{P}} \cdot f^*.$$

Then, we proof the following simple and useful lemma:

Lemma A.1. *With high probability the following holds for $\mathcal{F}_* := \mathcal{P}_{(f^*)^\perp}\mathcal{F}$:*

$$\sup_{f \in \mathcal{F}_*} \langle f, f^*\rangle_n \lesssim \mathcal{G}_n(\mathcal{F}).$$

Proof. Note that for every $f \in \mathcal{F}_*$, we have by definition that

$$\langle f, f^*\rangle_{\mathbb{P}} = 0.$$

Therefore, by the Gine-Zinn symetrization (Koltchinskii, 2011, Thms 2.1 and 2.2) and $\mathcal{F}^* = -\mathcal{F}^*$, we obtain that

$$\begin{aligned}\mathbb{E} \sup_{f \in \mathcal{F}_*} \langle \mathbf{f^*}, \mathbf{f}\rangle_n &= \mathbb{E} \sup_{f \in \mathcal{F}_*} \langle \mathbf{f^*}, \mathbf{f}\rangle_n - \langle f, f^*\rangle_{\mathbb{P}} = \mathbb{E} \sup_{f' \in \mathcal{F}_* \cdot f^*} \left| \int f' d(\mathbb{P}_n - \mathbb{P}) \right| \\ &= \mathbb{E} \sup_{f' \in \mathcal{F}_* \cdot f^*} \int f' d(\mathbb{P}_n - \mathbb{P}) \leq 2 \cdot \mathrm{Rad}_n(f^* \cdot \mathcal{F}_*) \lesssim \mathrm{Rad}_n(\mathcal{F}_*),\end{aligned}$$

where $\mathrm{Rad}_n(\mathcal{F}_*)$ are the Rademacher averages of $\mathcal{F}_*$, and in the last inequality we used that with a probability of $1 - n^{-1}$ it holds that $f^*(X_i) \leq C_1$. Using that

$$\mathrm{Rad}_n(\mathcal{F}_*) \lesssim \mathcal{G}_n(\mathcal{F}_*),$$

the claim follows from

$$\mathcal{G}_n(\mathcal{F}_*) \lesssim \mathcal{G}_n(\mathcal{F}), \tag{13}$$

as projections reduces the Gaussian complexity (up to a multiplicative absolute constant). □

Step II: Controlling the Inflation Factor In this part, we show that

$$\mathbb{E}_{\boldsymbol{\xi}} \|\widehat{f}_n(\boldsymbol{\xi} + \mathbf{f^*})\|^2 - \mathbb{E}_{\boldsymbol{\xi}} \|\widehat{f}_n(\boldsymbol{\xi})\|^2 \leq \left(\mathbb{E}_{\boldsymbol{\xi}} \|\widehat{f}_n(\boldsymbol{\xi})\|\right)^{2-p} = M(\mathcal{F}_n)^{2-p}, \tag{14}$$

To this end, recall the definition of $US(p)$ with parameter $s \geq 0$ and note that

$$\begin{aligned}\|\boldsymbol{\xi}\|_n^p + s\|\mathbf{f^*}\|_n^p &= \|\widehat{f}_n(\boldsymbol{\xi})\|^p + s\|\widehat{f}_n(\mathbf{f^*})\|^p \\ &\geq \frac{\|\widehat{f}_n(\boldsymbol{\xi}) + \widehat{f}_n(\mathbf{f^*})\|^p}{2} + \frac{\|\widehat{f}_n(-\boldsymbol{\xi}) + \widehat{f}_n(\mathbf{f^*})\|^p}{2} \\ &\geq \frac{\|\widehat{f}_n(\boldsymbol{\xi} + \mathbf{f^*})\|^p}{2} + \frac{\|\widehat{f}_n(-\boldsymbol{\xi} + \mathbf{f^*})\|^p}{2} \\ &= \frac{\|\boldsymbol{\xi} + \mathbf{f^*}\|_n^p}{2} + \frac{\|-\boldsymbol{\xi} + \mathbf{f^*}\|_n^p}{2} \\ &= \mathbb{E}_{\pm\boldsymbol{\xi}} \|\boldsymbol{\xi} + \mathbf{f^*}\|_n^p.\end{aligned}$$

Therefore, we obtain that

$$\mathbb{E}_{\pm\boldsymbol{\xi}}\|\boldsymbol{\xi}+\mathbf{f}^*\|_n^2=\mathbb{E}_{\pm\boldsymbol{\xi}}(\|\boldsymbol{\xi}+\mathbf{f}^*\|_n^p)^{2/p}\leq\mathbb{E}_{\pm\boldsymbol{\xi}}(\|\boldsymbol{\xi}\|_n^p+s\|\mathbf{f}^*\|_n^p)^{2/p}=\mathbb{E}_{\pm\boldsymbol{\xi}}\left[\|\boldsymbol{\xi}\|_n^2\left(1+s\left(\frac{\|\mathbf{f}^*\|_n}{\|\boldsymbol{\xi}\|_n}\right)^p\right)^{2/p}\right].$$

Using that with high probability $\|\boldsymbol{\xi}\|_n\geq C$, and the assumption of $\|\mathbf{f}^*\|\leq 1$, and the identity $(1+x)^{2/p}\approx 1+2x/p$, we obtain that

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{\xi}+\mathbf{f}^*\|_n^2-\|\boldsymbol{\xi}\|_n^2\right]\lesssim\mathbb{E}_{\boldsymbol{\xi}}\|\boldsymbol{\xi}\|_n^{2-p}$$

and by Borrel's lemma (cf. (Brazitikos et al., 2014, Thm 2.4.6)), we know that $\mathbb{E}_{\boldsymbol{\xi}}\|\boldsymbol{\xi}\|_n^{2-p}\lesssim\left(\mathbb{E}_{\boldsymbol{\xi}}\|\boldsymbol{\xi}\|_n\right)^{2-p}$. Therefore, we proved (14), and the claim follows.

Step III: Applying Uniform Convexity In this part, we prove the following equation:

$$\mathbb{E}\|\mathcal{P}_{f^*}\mathbb{E}_{\boldsymbol{\xi}}\widehat{f}_n|\mathbf{X}-f^*\|_{\mathbb{P}}^2\lesssim\frac{sM_n(\mathcal{F})^{2-p}\cdot M_n^*(\mathcal{F})^2}{tn^2}\tag{15}$$

Proof. First, by the $UC(2)$ property of $\mathcal{F}$, and the fact that f^* is independent of $\boldsymbol{\xi}$, we obtain that over each realization of $\boldsymbol{\xi}$ (and with high probability over $\mathbf{X}$) (here the expectation is over two points of $\boldsymbol{\xi},-\boldsymbol{\xi}$)

$$\begin{aligned}\mathbb{E}\|\widehat{f}_n(\boldsymbol{\xi}+\mathbf{f}^*)\|^2=&\mathbb{E}\left[\frac{\|\widehat{f}_n(\boldsymbol{\xi}+\mathbf{f}^*)\|^2}{2}+\frac{\|\widehat{f}_n(-\boldsymbol{\xi}+\mathbf{f}^*)\|^2}{2}\right]\\&\geq\mathbb{E}\left\|\frac{\widehat{f}_n(\boldsymbol{\xi}+\mathbf{f}^*)-\widehat{f}_n(-\boldsymbol{\xi}+\mathbf{f}^*)}{2}\right\|^2+t\mathbb{E}\left\|\frac{\widehat{f}_n(\boldsymbol{\xi}+\mathbf{f}^*)+\widehat{f}_n(-\boldsymbol{\xi}+\mathbf{f}^*)}{2}\right\|^2\\&\geq\mathbb{E}\|\widehat{f}_n(\boldsymbol{\xi})\|^2+t\mathbb{E}\left\|\frac{\widehat{f}_n(\boldsymbol{\xi}+\mathbf{f}^*)+\widehat{f}_n(-\boldsymbol{\xi}+\mathbf{f}^*)}{2}\right\|^2\end{aligned}\tag{16}$$

where the last inequality follows from the definition of the minimal norm solution. Now, as the noise is symmetric, and by Jensen's inequality, on the last term we obtain that

$$\mathbb{E}_{\boldsymbol{\xi}}\|\widehat{f}_n(\boldsymbol{\xi}+\mathbf{f}^*)\|^2\geq\mathbb{E}_{\boldsymbol{\xi}}\|\boldsymbol{\xi}\|_n^2+t\mathbb{E}_{\boldsymbol{\xi}}\left\|\frac{\widehat{f}_n(\boldsymbol{\xi}+\mathbf{f}^*)+\widehat{f}_n(-\boldsymbol{\xi}+\mathbf{f}^*)}{2}\right\|^2=\mathbb{E}_{\boldsymbol{\xi}}\|\boldsymbol{\xi}\|_n^2+t\left\|\mathbb{E}_{\boldsymbol{\xi}}\left[\widehat{f}_n(\boldsymbol{\xi}+\mathbf{f}^*)\right]\right\|^2,$$

where we used that these two terms have the same expectaion. Now, as the function

$$f_{\mathbf{X}}:=\mathbb{E}_{\boldsymbol{\xi}}\left[\widehat{f}_n(\boldsymbol{\xi}+\mathbf{f}^*)\right]\in\mathcal{B}(\mathcal{X})$$

interpolates $\mathbf{f}^*$, it may be decomposed into

$$f_{\mathbf{X}}=\mathcal{P}_{f^*}f_{\mathbf{X}}+\mathcal{P}_{(f^*)^\perp}f_{\mathbf{X}}=(1-\gamma_{\mathbf{X}})f^*+\underbrace{\mathcal{P}_{(f^*)^\perp}f_{\mathbf{X}}}_{:=f'_{\mathbf{X}}},$$

where $\gamma_{\mathbf{X}}=1-\left\langle\mathbb{E}_{\boldsymbol{\xi}}\mathcal{P}_{f^*}\widehat{f}_n|\mathbf{X},f^*\right\rangle_{\mathbb{P}}$. Note that the following holds: First,

$$\boldsymbol{\mathcal{P}}_{(\mathbf{f}^*)^\perp}\mathbf{f}_{\mathbf{X}}=\gamma_{\mathbf{X}}\mathbf{f}^*,$$

secondly, as with high probability it holds $\|f^*\|_{\mathbb{P}_n}\asymp\|f^*\|_{\mathbb{P}},$ (and recall the notation $\|\cdot\|_{\mathbb{P}_n}=\|\cdot\|_{L_2(\mathbb{P}_n)}$)

$$\gamma_{\mathbf{X}}^2\asymp\|f'_{\mathbf{X}}\|_{\mathbb{P}_n}^2$$

Next, by Step I, with high probability, it holds:

$$h_{\mathcal{F}_n^*}(\mathbf{f}^*)=\sup_{f\in\mathcal{F}^*}\langle\mathbf{f}^*,\mathbf{f}\rangle_n\lesssim\mathcal{G}_n(\mathcal{F})\|f^*\|_{\mathbb{P}_n}\asymp\mathcal{G}_n(\mathcal{F}).$$

Now, note that

$$\|f'_{\mathbf{X}}\|_{\mathcal{F}^*} \gtrsim \|f'_{\mathbf{X}}\|_{\mathbb{P}_n}\|\mathbf{f^*}\|_{\mathcal{F}_n^*} \geq \frac{\gamma_{\mathbf{X}}}{h_{\mathcal{F}_n^*}(\mathbf{f^*})},$$

and by using the last equation, and that $\|f'_{\mathbf{X}}\| \geq \|f'_{\mathbf{X}}\|_{\mathcal{F}^*}$ (as $f'_{\mathbf{X}} \perp f^*$), we obtain that

$$\frac{\|f'_{\mathbf{X}}\|}{\gamma_{\mathbf{X}}} \geq \frac{\|f'_{\mathbf{X}}\|_{\mathcal{F}^*}}{\gamma_{\mathbf{X}}} \gtrsim \|\mathbf{f^*}\|_{\mathcal{F}_n^*} \geq \frac{n}{\sup_{f\in\mathcal{F}_*}\langle \mathbf{f}, \mathbf{f^*}\rangle_n} \gtrsim \frac{n}{M_n^*(\mathcal{F})}. \tag{17}$$

Therefore, under $\mathcal{E}_1$, we have that

$$\|\mathbb{E}_{\boldsymbol{\xi}}\left[\widehat{f}_n(\boldsymbol{\xi}+\mathbf{f^*})\right]\| \gtrsim \max\{\frac{n\gamma_{\mathbf{X}}}{M_n^*(\mathcal{F})}, 1\}$$

where we used our assumption of $\|f^*\| \gtrsim 1$; and hence,

$$\mathbb{E}_{\boldsymbol{\xi}}\|\widehat{f}_n(\boldsymbol{\xi}+\mathbf{f^*})\|^2 \geq \mathbb{E}\|\widehat{f}_n(\boldsymbol{\xi})\|^2 + c_1 t\cdot\max\{1, \frac{\gamma_{\mathbf{X}}^2 n^2}{M_n^*(\mathcal{F})^2}\}, \tag{18}$$

By using (14) and the last inequality, we obtain that

$$\frac{tn^2}{M_n^*(\mathcal{F})^2}\cdot\gamma_{\mathbf{X}}^2 \lesssim \mathbb{E}\left[\|\widehat{f}_n(\boldsymbol{\xi}+\mathbf{f^*})\|^2\right] - \mathbb{E}\left[\|\widehat{f}_n(\boldsymbol{\xi})\|^2\right] \lesssim \frac{s}{\mathbb{E}\|\boldsymbol{\xi}\|_n^p}\mathbb{E}\|\boldsymbol{\xi}\|_n^2 \lesssim sM_n(\mathcal{F})^{2-p}.$$

Or equivalently (under the expectation of the $1-n^{-2}$ event of $\mathcal{E}_1$), we obtain that

$$\gamma_{\mathbf{X}}^2 \lesssim \frac{sM_n^*(\mathcal{F})^2 M_n(\mathcal{F})^{2-p}}{tn^2}.$$

Therefore, we conclude

$$\mathbb{E}_{\mathbf{X}}\left\|\mathcal{P}_{f^*}(\mathbb{E}_{\boldsymbol{\xi}}\widehat{f}_n|\mathbf{X}) - f^*\right\|_{\mathbb{P}}^2 \asymp \mathbb{E}\gamma_{\mathbf{X}}^2 \lesssim \frac{sM_n^*(\mathcal{F})^2 M_n(\mathcal{F})^{2-p}}{tn^2},$$

and (15) follows. □

Step IV: Concluding Theorem 3.1 It remains to prove that

$$\mathbb{E}\left\|\mathcal{P}_{(f^*)^\perp}\mathbb{E}_{\boldsymbol{\xi}}\left[\widehat{f}_n|\mathbf{X}\right]\right\|_{\mathbb{P}}^2 \lesssim \frac{sM_n(\mathcal{F})^{2-p}\cdot M_n^*(\mathcal{F})^2}{tn^2}. \tag{19}$$

as we know that $\|f^*\|_{\mathbb{P}_n} \asymp \|f^*\| \asymp 1$, under the event of the last the last step, we have that

$$\left\|\mathcal{P}_{(f^*)^\perp}\mathbb{E}_{\boldsymbol{\xi}}\left[\widehat{f}_n|\mathbf{X}\right]\right\|_{\mathbb{P}_n}^2 \lesssim \frac{sM_n(\mathcal{F})^{2-p}\cdot M_n^*(\mathcal{F})^2}{tn^2}.$$

Also, note that one can easily show that

$$\mathcal{P}_{(f^*)^\perp}\mathbb{E}_{\boldsymbol{\xi}}\widehat{f}_n|\mathbf{X} \in \gamma_*\cdot\mathcal{F},$$

where $\gamma_* \lesssim \sqrt{s/t}\cdot M(\mathcal{F}_n)^{1-p/2}$.

Therefore, using the small-ball assumption (Mendelson, 2014), which implies that for all $f\in\mathcal{F}$ with high probablity over $X_1,\ldots,X_n$ (denote this event by $\mathcal{E}_2$)

$$\|f\|_{\mathbb{P}}^2 \leq 2\|f\|_{\mathbb{P}_n}^2 + C_{SB}\cdot\mathcal{G}_n(\mathcal{F})^2,$$

where $C_{SB}\geq 0$ is the small ball constant that emerges from the constants appearing in Assumption 3. Therefore, under $\mathcal{E}_1\cap\mathcal{E}_2$, and using the last two equations

$$\begin{aligned}\|\mathbb{E}_{\boldsymbol{\xi}}\mathcal{P}_{(f^*)^\perp}\widehat{f}_n|\mathbf{X}\|_{\mathbb{P}}^2 &\lesssim \frac{sM_n(\mathcal{F})^{2-p}\cdot M_n^*(\mathcal{F})^2}{tn^2} + \gamma_*^2\cdot C_{SB}\cdot\mathcal{G}_n(\mathcal{F})^2 \\ &\lesssim \frac{C_{SB}\cdot s\cdot M_n(\mathcal{F})^{2-p}\cdot M_n^*(\mathcal{F})^2}{tn^2}.\end{aligned} \tag{20}$$

Therefore, Theorem 3.1 follows by taking expectation on high probablity event $\mathcal{E}_1\cap\mathcal{E}_2$, as

$$\mathrm{Var}(\mathbb{E}\left[\widehat{f}_n|\mathbf{X}\right]) + B^2(\widehat{f}_n) = \mathbb{E}\|\mathcal{P}_{f^*}\mathbb{E}\left[\widehat{f}_n|\mathbf{X}\right]\|_{\mathbb{P}}^2 + \mathbb{E}\|\mathcal{P}_{(f^*)^\perp}\mathbb{E}\left[\widehat{f}_n|\mathbf{X}\right]\|_{\mathbb{P}}^2$$

B. Proof of Theorem 3.2

In order to prove our lower bound over the conditional variance, we first study the "intrinsic" variance of $\widehat{f}_n$, defined via

$$\mathrm{Var}_{\|\cdot\|}\left[\widehat{f}_n\right] := \mathbb{E}_{\mathcal{D}}\|\widehat{f}_n(\mathcal{D}) - \mathbb{E}\widehat{f}_n(\mathcal{D})\|^2,$$

under the assumption that the underlying space is $CO(2)$.

Theorem B.1 (Intrinsic Variance Under $CO(2)$ – Reverse Efron-Stein)**.** *Assume that $(\|\cdot\|, \mathcal{B}(\mathcal{X}))$ is $CO(2)$ with constant $t > 0$. Then, under Assumptions 1-2, the minimum norm solution $\widehat{f}_n$ satisfies that*

$$n \cdot t^2 \cdot K_{\|\cdot\|}^{-2} \cdot \mathbb{E}\|\widehat{f}_X\|^2 \lesssim \mathrm{Var}_{\|\cdot\|}\left[\widehat{f}_n\right] \asymp M_n(\mathcal{F})^2, \tag{21}$$

where $\widehat{f}_X$ is minimum norm interpolator of the samples $\{(X_1, 1), (X_2, 0), \ldots, (X_{n-1}, 0), (X_n, 0)\}$.

B.1. Preliminaries on K-convexity

First, we present a few facts and further details on K-convexity constant (cf. (Artstein-Avidan et al., 2015, Cpt. 6)), for the reader's convenience. Let us define the K-convexity in a proper way (with respect to the Gaussian measure). Recall that γ_n is the Gaussian measure over $\mathbb{R}^n$, and consider a map $F : (\mathbb{R}^n, \gamma_n) \to (\mathcal{B}(\mathcal{X}), \|\cdot\|)$, and its so-called ℓ-norm is defined via

$$\ell(F) := \sqrt{\mathbb{E}_{\boldsymbol{\xi}}\|F(\boldsymbol{\xi})\|^2},$$

and the linearization of F is defined as

$$L_F := \sum_{i=1}^{n} \alpha_i \boldsymbol{\xi}_i + \int_{\mathbb{R}^n} F(\boldsymbol{\xi}) d\gamma_n(\boldsymbol{\xi}),$$

where $\alpha_i = \int_{\mathbb{R}^d} F(\boldsymbol{\xi}_+^{(i)}) - F(\boldsymbol{\xi}_-^{(i)}) d\gamma_n(\boldsymbol{\xi}),$, and $\boldsymbol{\xi}_\pm^{(i)} = (\xi_1, \ldots, \pm|\xi_i|, \ldots, \xi_n)$. In other words, we project F to the subspace spanned by Hermite polynomials of degree one.

We endow the Banach space of maps (and its subspace of linear maps) from $(\mathbb{R}^n, \gamma_n) \to (\|\cdot\|, \mathcal{B}(\mathcal{X}))$ with the ℓ-norm that is defined above. Next, we denote by $\mathcal{L}$ to be the linear operator that maps any centered F ($\int_{\mathbb{R}^n} F(\boldsymbol{\xi}) d\gamma_n(\boldsymbol{\xi}) = 0$) to its linearization L_F and the (Gaussian) K-convexity constant is defined as the operator norm of $\mathcal{L}$,

$$K_{\|\cdot\|} := \|\mathcal{L}\|_{OP}, \tag{22}$$

or equivalently for any zero mean map $F : (\mathbb{R}^n, \gamma_n) \to (\|\cdot\|, \mathcal{B}(\mathcal{X}))$ it holds that

$$\ell(F) \gtrsim K_{\|\cdot\|}^{-1} \cdot \ell(L_F).$$

We conclude the preliminaries with a few fundamental facts on K-convexity.

1. In general, any norm $\|\cdot\|$ in $\mathbb{R}^n$ (for which the unit ball is a symmetric convex body $K \subset \mathbb{R}^n$) satisfies that

   $$K_{\|\cdot\|} \lesssim \log(1 + d_{BM}(K, B_2^n)),$$

   where d_{BM} denotes the Banach-Mazur distance and B_2^n is the Euclidean ball in $\mathbb{R}^n$. Therefore, in every regression task with respect to any norm in $\mathbb{R}^d$, it follows that $K_{\|\cdot\|} \lesssim \log d$.

2. When $\|\cdot\|$ is type $p > 1$ with a constant, $s \geq 0$, one can show that $K_{\|\cdot\|} \lesssim C(s, p)$ and furthermore, by Fact 2 for any $UC(2)$ it holds holds that $K_{\|\cdot\|} \lesssim 1/\sqrt{t}$.

Remark B.1. A delicate point that we would like to mention is that the definition of type and cotype is classically defined with respect to the Rademacher random variables rather than Gaussian. However, the remarkable result of Maurey and Pisier (cf. (Artstein-Avidan et al., 2022, Theorem 5.4.1)) implies that for cotype $q \in (2, \infty)$ spaces, these definitions are equivalent up to a multiplicative constant that only depends on the cotype constant t.

B.2. Proof of Theorem B.1

Proof. First, recall (4) above. We consider the *non*-linear map $M_{\mathbf{X}} : (\mathbb{R}^n, \gamma_n) \to (\mathcal{B}(\mathcal{X}), \|\cdot\|)$ which is defined via

$$(\widehat{f}_n(\mathcal{D})|\mathbf{X}) - (\mathbb{E}_{\boldsymbol{\xi}}\widehat{f}_n(\mathcal{D})|\mathbf{X}), \tag{23}$$

and exploit the linear structure behind it.

For a realization of the vector $\mathbf{X}$, consider the linearization of $M_{\mathbf{X}}$ with respect to the span of the Hermite polynomials of degree one, precisely

$$L_{\mathbf{X}}(\boldsymbol{\xi}) := \sum_{i=1}^{n} \alpha_i \xi_i,$$

where $\alpha_1, \ldots, \alpha_n \in \mathcal{B}(\mathcal{X})$ are defined as

$$\alpha_i := \int_{\mathbb{R}^n} \xi_i \cdot (M_{\mathbf{X}}(\boldsymbol{\xi}))d\gamma_n(\boldsymbol{\xi}) = \int_{\mathbb{R}^n} \frac{|\xi_i|}{2} \cdot (M_{\mathbf{X}}(\boldsymbol{\xi}^i_+)) - M_{\mathbf{X}}(\boldsymbol{\xi}^i_-))d\gamma_n(\boldsymbol{\xi})$$

and $\boldsymbol{\xi}^i_{\pm} = (\xi_1, \ldots, \pm|\xi_i|, \ldots, \xi_n)$. Note that each α_i is an interpolator to

$$(0, \ldots, \mathbb{E}|\xi|^2, \ldots, 0) = (0, \ldots, 1, \ldots, 0),$$

as we average interpolators with respect to the Gaussian measure. Now, since our space is of cotype 2, we apply the Pisier-Maurey result (see §B.1 above) on K-convexity bound which guarantees that

$$\ell(M_{\mathbf{X}})^2 \gtrsim K^{-2}_{\|\cdot\|} \cdot \ell(L_{\mathbf{X}})^2,$$

or equivalently,

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\left\|\widehat{f}_n - \mathbb{E}_{\boldsymbol{\xi}}\left[\widehat{f}_n|\mathbf{X}\right]\right\|^2\right] \gtrsim K^{-2}_{\|\cdot\|} \cdot \mathbb{E}_{\boldsymbol{\xi}} \sum_{i=1}^{n} \|\alpha_i\xi_i\|^2.$$

Taking expectation over $\mathbf{X}$ the left hand side gives the conditional variance. Next, using the cotype 2-property and Jensen's inequality, we obtain that

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\left\|\widehat{f}_n - \mathbb{E}_{\boldsymbol{\xi}}\left[\widehat{f}_n|\mathbf{X}\right]\right\|^2\right] \gtrsim t^2 \cdot K^{-2}_{\|\cdot\|} \cdot \sum_{i=1}^{n} \|\alpha_i\|^2.$$

As, $\{\alpha_i\}_{i=1}^n$ depend on $\mathbf{X}$, we take expectation over $\mathbf{X}$, we know that

$$\mathbb{E}\|\alpha_1\|^2 = \ldots = \mathbb{E}\|\alpha_i\|^2 = \ldots = \mathbb{E}\|\alpha_n\|^2$$

and therefore by Jensen's inequality

$$\mathrm{Var}_{\|\cdot\|}\left[\widehat{f}_n\right] \geq \mathrm{Var}_{\|\cdot\|}\left[\widehat{f}_n|\mathbf{X}\right] \gtrsim n \cdot K^{-2}_{\|\cdot\|} \cdot t^2 \cdot \mathbb{E}_{\mathbf{X}}\|\alpha_1\|^2 \geq n \cdot K^{-2}_{\|\cdot\|} \cdot t^2 \cdot \mathbb{E}_{\mathbf{X}}\|\widehat{f}_X\|^2,$$

where in the last inequality we used the definition of the minimum norm solution.

Finally, we prove the right hand side of the bound . First, we consider by $f^* \equiv 0$, and using Borell's lemma (cf. (Artstein-Avidan et al., 2015) or (Brazitikos et al., 2014))

$$\mathrm{Var}_{\|\cdot\|}\left[\widehat{f}_n\right] = \mathbb{E}_{\mathbf{X}}\ell^2(\widehat{f}_n) = \mathbb{E}_{\mathbf{X}}\int \|\boldsymbol{\xi}\|_n^2 d\gamma_n \lesssim \mathbb{E}_{\mathbf{X}}\left(\int \|\boldsymbol{\xi}\|_n d\gamma_n\right)^2 \lesssim M_n(\mathcal{F})^2,$$

and the claim follows when $f^* \equiv 0$. Finally, under Assumption 2, we conclude that

$$\mathbb{E}_{\boldsymbol{\xi}}\|\widehat{f}_n(\mathbf{f}^* + \boldsymbol{\xi})\|^2 \lesssim \mathbb{E}\|\boldsymbol{\xi}\|_n^2$$

and the claim follows as $\|\mathbb{E}_{\boldsymbol{\xi}}\widehat{f}_n(\mathbf{f}^* + \boldsymbol{\xi})\| \lesssim M(\mathcal{F}_n)$ by simply using Jensen's, and taking additional expectation over $\mathbf{X}$. □

B.3. Proof of Theorem 3.2

In what follows, recall that $\|\cdot\|_{\mathbb{P}}$ denotes the $L_2(\mathbb{P})$ norm. Here, we lower bound the expected conditional variance of $\widehat{f}_n$, that is

$$\mathbb{E}_{\mathbf{X}}\mathrm{Var}(\widehat{f}_n|\mathbf{X}).$$

Following the ideas from Theorem B.1, we consider an operator that is defined for every realization of $\mathbf{X}$, denoted by $M_{\mathbf{X}}$, defined via

$$\boldsymbol{\xi} \mapsto \widehat{f}_n|\mathbf{X} - \mathbb{E}\widehat{f}_n|\mathbf{X}.$$

And we will inspect it with respect to different two norms $\|\cdot\|$, $\|\cdot\|_{\mathbb{P}}$ and lower bound the expected conditional variance. More accurately, first we linearize the operator $M_{\mathbf{X}}$, to obtain $\alpha_1,\ldots,\alpha_i,\ldots,\alpha_n$ (that depend on $\mathbf{X}$) defined by

$$\alpha_i := \int_{\mathbb{R}^n} \frac{|\xi_i|}{2}\cdot(\widetilde{f_n}(\mathbf{f^*}+\boldsymbol{\xi}^i_+)) - \widetilde{f_n}(\mathbf{f^*}+\boldsymbol{\xi}^i_-))d\gamma_n(\boldsymbol{\xi}),$$

where $\boldsymbol{\xi}^i_{\pm} = (\xi_1,\ldots,\pm|\xi_i|,\ldots,\xi_n)$. Using Theorem B.1, it is easy to see that

$$\mathbb{E}\|\alpha_i\|^2 \lesssim (t^{-1}/\sqrt{n})^2\cdot K_{\|\cdot\|}\cdot M_n(\mathcal{F})^2$$

Then, by Jensen's inequality, we obtain that

$$\mathbb{E}\|\alpha_i\| \lesssim \frac{t^{-1}\cdot K_{\|\cdot\|}}{\sqrt{n}}\cdot M_n(\mathcal{F}) := \frac{t^{-1}_{\|\cdot\|}}{\sqrt{n}}\cdot M_n(\mathcal{F}).$$

The linearization of $M_{\mathbf{X}}$ writes

$$\widetilde{L}_n := \sum_{i=1}^n \alpha_i\xi_i$$

and note that for each realization of $\mathbf{X}$ and $\boldsymbol{\xi}$, it holds $\widetilde{\mathbf{L_n}}|\mathbf{X}=\boldsymbol{\xi}$.

As $K_{\|\cdot\|_{\mathbb{P}}}=1$ (since it is Hilbert space), we obtain that

$$\mathbb{E}\,\mathrm{Var}(\widehat{f}_n|\mathbf{X}) = \mathbb{E}_{\boldsymbol{\xi}}\|\widehat{f}_n - \mathbb{E}_{\boldsymbol{\xi}}\widehat{f}_n(\mathcal{D})\|^2_{\mathbb{P}} \geq \mathbb{E}\|\sum_{i=1}^n \alpha_i\xi_i\|^2_{\mathbb{P}} = \sum_{i=1}^n \mathbb{E}_{\mathbf{X}}\|\alpha_i\|^2_{\mathbb{P}} = n\cdot\mathbb{E}\|\alpha_1\|^2_{\mathbb{P}}.$$

where in the last inequality we used that n samples are (X_i,ξ_i) are independent and identically distributed.

Now, recall that for every realization $\mathbf{X}$, we have that $\boldsymbol{\alpha}_i = (0,\ldots,1,\ldots,0)$, and that $\mathbb{E}\|\alpha_i\| \lesssim \frac{t^{-1}_{\|\cdot\|}\cdot M(\mathcal{F}_n)}{\sqrt{n}}$. Therefore, by Markov's inequality, $\alpha_1 \in r\cdot\mathcal{F}$ with probability at least $3/4$ for $r = c\frac{t^{-1}_{\|\cdot\|}\cdot M(\mathcal{F}_n)}{\sqrt{n}}$ for an appropriate $c>0$. Thus, with probability at least $1/4$, $\alpha_1\in r\cdot\mathcal{F}$ and $\Psi^X_n(\mathbf{e_1},r)\geq\Psi^X_n(\mathbf{e_1},r)$, which implies that:

$$\mathbb{E}\|\alpha_1\|^2_{\mathbb{P}} \gtrsim \Psi_n\left(\mathbf{e_1}, C\cdot\frac{t^{-1}_{\|\cdot\|}M(\mathcal{F}_n)}{\sqrt{n}}\right)^2.$$

By combining all we obtained that

$$\mathbb{E}\,\mathrm{Var}(\widehat{f}_n|\mathbf{X}) \gtrsim n\cdot\Psi_n\left(\mathbf{e_1}, C\cdot\frac{K_{\|\cdot\|}\cdot t^{-1}\cdot M_n(\mathcal{F})}{\sqrt{n}}\right)^2,$$

and the claim follows. We conclude the proof with two remarks. □

B.4. Concluding remarks

Remark B.2 (The K-convexity and minimum norm interpolators). Following the notation of the proof, note that by definition it holds that

$$\|L_{\mathbf{X}}(\widehat{f}_n(\boldsymbol{\xi}))\| \leq \|\widehat{f}_n(\boldsymbol{\xi})\|,$$

in particular for the ℓ-norm, we have that

$$\ell(\widehat{f}_n) \leq \ell(L_{\mathbf{X}}(\widehat{f}_n)).$$

When $\|\cdot\|$ is Hilbert, the converse holds due to the Representer theorem (cf. (Vershynin, 2018)). Remarkably, the K-convexity implies that in some sense the converse is true, meaning that

$$K_{\|\cdot\|}^{-1} \cdot \ell(L_{\mathbf{X}}) \lesssim \ell(\widehat{f}_n) \lesssim \ell(L_{\mathbf{X}}).$$

Namely, the minimal norm interpolator is norm equivalent *in terms of ℓ-norm* to an ellipsoid. This equivalence holds up to a price that is at most logarithmic in the Banach Mazur distance to a Hilbert space. The core point that we highlight is that if we have a non-trivial type and cotype (i.e. $1+\delta$ with some constant $t, s > 0$ that are independent of the dimension and n) then we pay a price only in terms of t, s. It follows from the K-convexity argument that this ellipsoid is induced by averaging from the local behavior of the boundary of $\mathcal{F}$, as the coefficients α_i given by

$$\alpha_i \approx \int_{\mathbb{S}^{n-1}} F(\sqrt{n}\cdot\boldsymbol{\xi}_i^+) - F(\sqrt{n}\cdot\boldsymbol{\xi}_i^0) + F(\sqrt{n}\cdot\boldsymbol{\xi}_i^0) - F(\sqrt{n}\cdot\boldsymbol{\xi}_i^-) dU(\boldsymbol{\xi}),$$

where $\boldsymbol{\xi}_i^0 := (\xi_1, , \ldots, \xi_{i-1}, 0, \xi_{i+1}, \ldots, \xi_n)$. Remarkably, it almost captures (on average) the global behavior of $\mathcal{F}$.

Remark B.3. [On Efron Stien's type bounds in Banach Spaces] We begin by reminding the Efron-Stein's inequality (see (Boucheron et al., 2013)) for a function $f : \mathbb{R}^n \to \mathbb{R}$ with random input $X_1, \ldots, X_n$.

$$\mathrm{Var}(F(X_1, \ldots, X_n)) \leq 2\sum_{i=1}^{n} \mathbb{E}\left[(F(X_{-i}, X_i) - F(X_{-i}, X_i'))^2\right],$$

where X_i' is independent copy of X_i.

Using the classical observations of Enflo's (Enflo, 1970a;b), known in these days as the Enflo's type property, he showed that for any non-linear operator $F : \{-1, 1\}^n \to (\mathcal{Z}, \|\cdot\|_{\mathcal{H}})$, where $\|\cdot\|_{\mathcal{H}}$ is a Hilbert space that

$$\mathrm{Var}_{\|\cdot\|_{\mathcal{H}}}(F) \lesssim \sum_{i=1}^{n} \mathbb{E}\left[\|(F(\epsilon_{-i}, \epsilon_i) - F(\epsilon_{-i}, -\epsilon_i))^2\|_{\mathcal{H}}^2\right]$$

here randomness is with respect to the uniform measure over the hypercube. Later, (Pisier, 2006) extended this to any type 2 norm – with a price of $\log(n)$-factor. Finally, in the recent paper of (Ivanisvili et al., 2020), they removed the $\log(n)$-factor by showing that for any type 2 norm

$$\mathbb{E}\|F(\epsilon) - F(-\epsilon)\|^2 \lesssim \sum_{i=1}^{n} \mathbb{E}\left[\|(F(\epsilon_{-i}, \epsilon_i) - F(\epsilon_{-i}, -\epsilon_i))^2\|^2\right].$$

Unfortunately, the most similar version for such bound for $CO(2)$-spaces (Mendel & Naor, 2008), is not applicable in our setting. Yet, Pisier's K-convexity constant (for Rademacher or Gaussian noise) is the closest equivalent version to a "reverse Efron Stien" inequality as it implies that

$$\log(1 + d_{BM}(\|\cdot\|, \|\cdot\|_{\mathcal{H}}))^{-1} \cdot \sum_{i=1}^{n} \|\mathbb{E}_{-i}\left[F(\epsilon_{-i}, \epsilon_i) - F(\epsilon_{-i}, -\epsilon_i)\right]\|^2 \lesssim \mathrm{Var}_{\|\cdot\|}(F).$$

Remark B.4. From a statistical point of view, the first term of our bound has an intuitive interpretation. From the sub-class of the functions that interpolate $n-1$ i.i.d. data points with zero, we choose one that can interpolate on a fresh data point X_n with the value of 1, which aligns to the lower bound on Theorem B.1.

B.5. Proof of Theorem 3.3

In this part, we assume that $\boldsymbol{\xi} \sim U(\mathbb{S}^{n-1})$ and use the notation of $P := \frac{1}{\sqrt{d}}\mathbf{X}$. Throughout this proof, we will use the notation of $\sigma \in [d]$, i.e. a subset of $1, \ldots, d$. Recall that $p \in (1+\eta(d), 2]$, where

$$\eta(d) := C \cdot \frac{\log\log\log(d)}{\log\log(d)}.$$

Also, denote by

$$\widehat{f}_n := \operatorname*{argmin}_{Pv=\boldsymbol{\xi}} \|v\|_p.$$

Recall the definition of the $\ell_{p,d} := \ell_p(\mathbb{R}^d)$ as the ℓ_p ball in $\mathbb{R}^d$, and denote by $\ell'_{p,d} := d^{1/p-1/2} \cdot \ell_{p,d}$, namely $\ell'_{p,d}$ is in John's position (cf. (Artstein-Avidan et al., 2015)).

Theorem B.2. *Let $p \in (1+\eta(d), 2]$ an let $d^{\frac{1}{p}-\frac{1}{2}} \lesssim n \lesssim d/\log(d)$. Then, with probability of at least $1-d^{-1}$ it holds that*

$$1 \lesssim \|\widehat{f}_n\|_2 \lesssim \log(d).$$

Consider Example 1 when $w^* \equiv 0$, then the last theorem implies that

$$\mathrm{Var}(\widehat{w}_p) = \mathbb{E}\|\widehat{w}_p\|_2^2 \lesssim \frac{\log(d)n}{d},$$

to see this, apply the last theorem with $\boldsymbol{\xi}' \sim \sqrt{n} \cdot U(\mathbb{S}^{n-1})$ and the definition of $\widehat{w}_p = \mathrm{argmin}_{\mathbf{X}w=\boldsymbol{\xi}} \|w\|_p$ implies that $\mathbb{E}\|\widehat{w}_p\|_2^2 \approx \frac{n}{d}\mathbb{E}\|\widehat{f}_n\|_2^2$, and the claim follows by the homogeneity of $\widehat{w}_p$.

B.5.1. Preliminaries

First, we state a lemma known as Kashin's theorem cf. (Artstein-Avidan et al., 2015, Thm 5.5.3) and (Szarek, 1990). For completeness, we provide a simple proof in §B.5.2 below

Lemma B.3. *Let $\mathcal{C}_d := [-1/\sqrt{d}, 1/\sqrt{d}]^d$, then when $d \geq Cn$ (for sufficiently large $C > 0$), and under the $\mathbf{X}$ that satisfies Example 1, the following holds:*

$$c\epsilon \cdot B_n \subset \frac{1}{\sqrt{d}}\mathbf{X}(\mathcal{C}_d) = P(\mathcal{C}_d)$$

with probability of at least $1 - \exp(-c\log(1/\epsilon)d)$, and c depends on the sub-Gaussianity and the maximum of the density of X (see Example 1).

In other words, when d is large enough compared to n, the projection of a cube contains a ball.

Next, we state another classical bound (cf. (Aubrun & Szarek, 2017)) on the singular values of $\mathbf{X}$.

Lemma B.4. *Under our assumptions*

$$\Pr\left(\sigma_{\min}(\mathbf{X}) \asymp \sigma_{\max}(\mathbf{X}) \asymp \sqrt{d}\right) \geq 1 - \exp(-cd). \tag{24}$$

Proof of Theorem B.2

Without loss of generality, we set $\ell_{p,d} := \ell'_{p,d}$, and we only prove this theorem for $p = 1+\eta(d)$ (for sufficiently large $C \geq 0$) and for $n \asymp \tilde{O}(\sqrt{d})$, the other regimes in theorem follow from similar arguments.

As $\ell_{p,d}$ is in John's position, it follows from a classical result (Schütt, 1984, Thm 1) that

$$\log \underbrace{\mathcal{N}\left(\lambda_k, \ell_{p,d}, \|\cdot\|_2\right)}_{:=\mathcal{N}_k} = k.$$

where $\lambda_k \asymp \left(\frac{k}{d\log(d/k)}\right)^{1/2-1/p}$ when $\log(d) \le \lambda_k \le d$, and $\lambda_k \asymp d^{1/p-1/2}$, when $1 \le k \le \log(d)$. We will use its explicit construction and slightly modify it. The net $\mathcal{N}_k$ is composed by a $O((k/d)^C)$-net (in terms of ℓ_2) of the sets

$$A_\sigma := \left(\frac{d\log(d/k)}{k}\right)^{1/p-1/2} \cdot \frac{1}{\sqrt{|\sigma|}} B_\infty^\sigma \subset \ell_{p,d},$$

for all $\sigma \subset [d]$ such that $C \cdot k/\log(d/k) \le |\sigma| \le 2C \cdot k/\log(d/k)$, and when $|\sigma| \le \log(d)$, we set

$$A_\sigma := \left(\frac{d}{k}\right)^{1/p-1/2} \cdot \frac{1}{\sqrt{|\sigma|}} B_\infty^\sigma \subset \ell_{p,d}$$

To see that $\log \mathcal{N}_k \lesssim k$, we define for every fixed k,

$$\mathcal{V}_k := \{\bigcup_\sigma A_\sigma : \sigma \subset [d], k/(2\log(d/k)) \le |\sigma| \le k/\log(d/k), v \in \mathbb{R}^\sigma, v \in A_\sigma\} \tag{25}$$

and note that

$$\log \mathcal{N}((k/d)^C, \mathcal{V}_k, \|\cdot\|_2) \le \log\left(\binom{d}{k/\log(d/k)} \cdot \left(\frac{(d/k)^{1/p-1/2}}{(k/d)^C}\right)^{k/\log(d/k)}\right) \lesssim k$$

where we used that

$$\mathcal{N}(\epsilon, \frac{R}{\sqrt{|\sigma|}} B_\infty^\sigma, \|\cdot\|_2) \le \mathcal{N}(\epsilon, RB_2^\sigma, \|\cdot\|_2) \le (1+2R/\epsilon)^{|\sigma|}$$

and that $\binom{d}{l} \le (ed/l)^l$. Furthermore, observe for every $w \in \ell_{p,d}$, can be decomposed as follows:

$$w = \sum_{k \in 2,4,\ldots,d/\log(d)} \delta_k f_k + \delta_0 f_{ns}$$

where $f_k \in \mathcal{V}_k$, $\sum \delta_i^p \le 1$, $f_{ns} \in C\log(d) \cdot \mathcal{C}_d$, $\|f_k\|_p \asymp d^{1/p-1/2}$. To see this, sort the entries of w by decreasing magnitude coordinates, and use the definition of the ℓ_p norm.

In order to prove the theorem, we need to show that with high probability for all $k \lesssim d/\log(d)$, it holds that $\delta_k = \tilde{O}((k/d)^{1/p})$, then we know that up to a $log(d)$-factors, we have that for each $k \lesssim d/\log(d)$

$$\|\delta_k f_k\|_\infty \lesssim (d/k)^{-1/2} \cdot k^{-1/2} \lesssim d^{-1/2}.$$

Therefore, under such an event, $\widehat{f}_n \in C\log(d) \cdot \mathcal{C}_d$, and in particular $\|\widehat{f}_n\|_2 \lesssim \log(d)$. The rest of the proof is dedicated to proving this claim.

Now, recall that for any vector $v \in \mathbb{R}^n$, it holds that (cf. (Vershynin, 2018) on the JL lemma):

$$\Pr_P\left(\|Pv\|_2 > (1+\epsilon)\sqrt{\frac{n}{d}}\|v\|_2\right) \le \exp(-cn\epsilon^2), \tag{26}$$

where we used the fact that $\mathbb{E}\|Pv\|_2^2 = \frac{n}{d}\|v\|_2^2$, and therefore $\mathbb{E}\|Pv\|_2 = (1+O(1/\sqrt{n}))\sqrt{n/d}\|v\|_2$.

Now, by applying (26) (with $\epsilon \asymp \max\{\sqrt{k/n}, 1\}$) over $\mathcal{N}_k$ for any $1 \le k \le d/(C_2\log(d))$, it holds that

$$\Pr\left(\forall v \in \mathcal{V}_k : \|Pv\|_2 \le \underbrace{C_1 \log(d/k)^{1/p-1/2}\left(\frac{k}{d}\right)^{1-1/p}\|v\|_{p'}}_{\text{uniform deviation (*)}} + \underbrace{2\sqrt{\frac{n}{d}}\left(\frac{k}{\log(d/k)d}\right)^{1/2-1/p}\|v\|_{p'}}_{\text{expectation (**)}}\right)$$
$$\ge 1 - \exp(-ck) \ge 1 - \exp(-c_1 n),$$

where $\|v\|_{p'} = \frac{\|v\|_p}{d^{1/p-1/2}}$ (note that $\mathcal{V}_k$ is not a finite a set, however by using that $\sigma_{\max}(\mathbf{X}) \lesssim \sqrt{d}$ with high probability, and the $\mathcal{N}_k$ is $O((k/d)^C)$-net of $\mathcal{V}_k$, the last equation is valid).

Also, note that by Kashin theorem's and that $\sigma_{\max}(\mathbf{X}) \lesssim \sqrt{d}$, imply that

$$\Pr\left(cB_n \subset P(\ell_{p,d} \cap C \cdot \log(d) \cdot \mathcal{C}_d) \subset C_1 \log(d) B_n\right) \geq 1 - \exp(-Cd). \tag{27}$$

Let us explain why we assume that $p \geq 1 + \eta(d)$. First, we set $k_0 \asymp n/\log(n)$. Note that (26) above implies that when $k_0 \leq k \leq d/\log(d)$, it holds for any $f_k \in \mathcal{V}_k$ that

$$(*) \lesssim \max\{(\log\log(d))^{1/2} \log(d)^{1/p-1}, \log(d)^{-1}\} \lesssim \log(d)^{-C_1/\log\log(d)} \leq c(C_1), \tag{28}$$

where $c(t) > 0$ is a decreasing function of t that vanishes at infinity. So we may choose C_1 to be large enough such that $c(C_1) < 0.1$. Next, note that for this k_0 and $p \geq 1 + C\log\log(d)/\log(d)$, we have that for $f_{k_0} \in \mathcal{V}_{k_0}$ that

$$(**) \leq \sqrt{\frac{n}{d}} \cdot \left(\frac{d\log(d/k_0)}{k_0}\right)^{1/p-1/2} \lesssim \log(d) \cdot (n/d)^{\frac{C_1 \log\log n}{\log n}} \lesssim \log(d) \cdot (C_2)^{-\log\log d} \lesssim \log(d)^{-2}, \tag{29}$$

where we set $C > 0$ to be large enough.

Now, we would like to find the k such that the term $(*)$ is more dominant then $(**)$. Indeed, when $p > 1 + C/\log\log(d)$, it holds

$$\log(d/k)^{1/2-1/p} \left(\frac{k}{d}\right)^{1-1/p} \gtrsim \sqrt{\frac{n}{d}} \left(\frac{k}{\log(d/k)d}\right)^{1/2-1/p} \iff \sqrt{k/d} \asymp \sqrt{n/d} \iff k \gtrsim n$$

Therefore, when $k \leq n$, the expectation term is more dominant, while for $k \geq n$ the uniform deviation is more dominant.

Therefore, (26) implies the following:

$$\begin{aligned}\Pr\Bigg(\forall k \in \mathcal{R}, v_k \in \bigcup_{k\in\mathcal{R}} \mathcal{V}_k \ : \ \|Pv_k\|_2 \lesssim \log(d/k)^{1/p-1/2} \left(\frac{k}{d}\right)^{1-1/p} \|v_k\|_{p'}(1 + \log(n) \cdot \mathbb{1}_{k_0 \leq k \leq n})\Bigg) \\ \geq 1 - \exp(-cn),\end{aligned} \tag{30}$$

where

$$\mathcal{R} := \{k_0, 2k_0, \ldots, d/\log(d)\},$$

and

$$\mathcal{R}' = \{1, 2, 4, \ldots, k_0/4, k_0/2\}.$$

Also, note that under the events of lemmas B.4 and (26) it holds that

$$\Pr_{\boldsymbol{\xi}}\left(\forall k \in \mathcal{R}', \sup_{f_k \in \mathcal{V}_k} \langle \boldsymbol{\xi}, Pf_k\rangle \lesssim \left(\frac{k}{d}\right)^{1-1/p} \|v_k\|_{p'}(1 + \sqrt{\log(d)}\mathbb{1}_{k \leq \log(d)})\right) \geq 1 - n^{-100}. \tag{31}$$

To see this, note that for a fix $\mathcal{C}_\sigma$, it holds by Lipschitz concentration that

$$\begin{aligned}\sup_{f_k \in A_\sigma} \left\langle \boldsymbol{\xi}, \frac{Pf_k}{\|v_k\|_{p'}} \right\rangle &\lesssim \sqrt{\frac{\max\{k/\log(d/k), \log(d)\}}{n}} \left(\frac{d(1 + \log(d/k)\mathbb{1}_{k \geq \log(d)})}{k}\right)^{1/p-1/2} \\ &\lesssim \left(\frac{\max\{k, \sqrt{\log(d)}\}}{d}\right)^{1-1/p},\end{aligned}$$

with a probability of $1 - \exp(-C\max\{k, \log(n)\})$, and the claim follows by taking a union bound over $Ck/\log(d/k) \leq |\sigma| \leq 2Ck/\log(d/k)$, and then on $k \in \mathcal{R}'$. Note that the last inequality follows from our assumption on p.

Finally, for each $k \in 1, \dots, d/(C_2 \log(d))$, we define $f_k = \Pi_k(\widehat{f}_n) - \Pi_{k/2}(\widehat{f}_n)$, where Π_k outputs the vector with the largest k entries (in terms of absolute value); and we decompose $\widehat{f}_n$ as follows:

$$\widehat{f}_n = f_{ns} + \sum_{k \in \mathcal{R}} \delta_k f_k + \sum_{k \in \mathcal{R}'} \delta_k f_k, \tag{32}$$

where $f_{ns} \in C \log(d) \mathcal{C}_d$, for all $k \in 1, \dots, d/(C_2 \log(d))$, $f_k \in \mathcal{V}_k$, and $\delta_k \geq 0$, and we may assume without loss of generality that $\|f_k\|_p \asymp d^{1/p-1/2}$, and $\sum_{k \in \mathcal{R}'} \delta_k^p \leq 1$. Note that Hölder's inequality implies that

$$\sup_{\widehat{f}_n} \| \sum_{k \in \mathcal{R}} \delta_k P(f_k)\|_2 \leq \sum_{k \in \mathcal{R}} \delta_k \underbrace{\| \sup_{f_v \in \mathcal{V}_k} P(f_k)\|_2}_{f_k} \leq \|\boldsymbol{\delta}\|_p \|\mathbf{f}\|_q \leq C_2 \cdot c(C) \tag{33}$$

where $c(C)$ is defined in (28).

Before moving to the next step, we claim that

$$\log(d)^{-1} \lesssim \mathbb{E}\|\widehat{f}_n\|_{p'} \lesssim 1. \tag{34}$$

To see this, note that with high probability (over $\mathbf{X}$) Gaussian width of

$$P(\mathcal{S}) := \bigcup_{1 \leq k \leq k_0} \{ \sum_{k \leq k_0} \delta_k P(f_k) : f_k \in \mathcal{V}_k, \sum_{k \leq k_0} \delta_k^p \leq 1\} \subset \mathbb{S}^{n-1}$$

to see this let $\alpha_k := \log(n)^{1/p-1/2} \left(\frac{k}{d}\right)^{1-1/p}$, and recall that

$$\Pr(\forall k \in \mathcal{R}', \sup_{f_k \in \mathcal{V}_k} \langle \mathbf{f}, \boldsymbol{\xi} \rangle \lesssim \alpha_k) \geq 1 - n^{-100}.$$

Therefore, we have that

$$\mathbb{E}_{\boldsymbol{\xi}} \sup_{\mathbf{f} \in P(\mathcal{S})} \langle \mathbf{f}, \boldsymbol{\xi} \rangle \leq \|\boldsymbol{\delta}\|_p \|\boldsymbol{\alpha}\|_q \lesssim \log(d)^{-2}.$$

However, and we showed above that $\sup_{\widehat{f}_n} \| \sum_{k \in \mathcal{R}} \delta_k P(f_k)\|_2 \leq 0.01$. Therefore, by the additive of the mean width, with high probability

$$\mathbb{E} \sup_{\widehat{f}_n - f_{ns}} \left\langle \boldsymbol{\xi}, \widehat{f}_n - f_{ns} \right\rangle \leq 0.02$$

and by definition we have that $\mathbb{E} \sup_{\widehat{f}_n} \left\langle \boldsymbol{\xi}, \widehat{f}_n - f_{ns} \right\rangle = 1$. Hence, $\widehat{f}_n$ must contain a component from $C \log(d) \mathcal{C}_d \cap \ell_{p,d}$, and as

$$cB_n \subset P(C \log(d) \mathcal{C}_d \cap \ell_{p,d}) \subset C \log(d) B_n$$

it must hold that

$$\log(d)^{-1} \lesssim \mathbb{E}\|f_{ns}\|_{p'} \lesssim 1.$$

Finally, as $cB_2^n \subset P(\ell_{p,d})$, we have that

$$\Pr(|\frac{\|\boldsymbol{\xi}\|_n}{\mathbb{E}\|\boldsymbol{\xi}\|_n} - 1| \geq \epsilon) \leq 2 \exp(-cn(t/\log(d))^2),$$

by choosing $t \asymp \mathbb{E}\|\boldsymbol{\xi}\|_n$, we have that

$$\Pr(\|\widehat{f}_n\|_p \asymp \mathbb{E}\|f_{ns}\|_p) \geq 1 - \exp(-c_1 n/\log(d))$$

Finally, by using Markov's inequality and that $\|f_{ns}\|_1 \asymp d^{1-1/p}\|f_{ns}\|_p \asymp d^{1/2}$, must implies that there exists a σ of cardinality $|\sigma| \gtrsim d/\log(d)$, such that $|(f_{ns})_i| \gtrsim 1/(\log(d)\sqrt{d})$ for all $i \in \sigma$.

We summarize the last claim in the following lemma:

Lemma B.5. *Let* $\mathbf{X}$ *that lies in the events of* (30) *and Lemma B.4 that holds with a probability of at least* $1-\exp(-cn)$. *Then, there exists an event* $A_{\mathbf{X}} \subset \mathbb{S}^{n-1}$ *such that* $\sigma(A_{\mathbf{X}}) \geq 1-\exp(-cn/\log(n))$ *such that* $\|\widehat{f}_n\|_p \asymp \mathbb{E}\|\widehat{f}_n\|_p$, *and in particular, it holds that*

$$\widehat{f}_n = f_{ns} + \sum_{k\in\mathcal{R}} \delta_k f_k + \sum_{k\in\mathcal{R}'} \delta_k f_k, \tag{35}$$

where $\delta_k \geq 0$, $\sum_k \delta_k^p \leq 1$, *and* $\|f_{ns}\|_p \asymp \mathbb{E}\|\widehat{f}_n\|_p$, *and* $\|f_{ns}\|_\infty \lesssim \log(d)/\sqrt{d}$. *Furthermore,* f_{ns} *has* $d/\log(d)$ *(denoted by* σ*) such that* $|f_i| \gtrsim 1/(\log(d)\sqrt{d})$ *for all* $i \in \sigma$, $f_k \in \mathcal{V}_k$, *and* $\|f_k\|_p \asymp d^{1/p-1/2}$.

Step I: In this *step*, we show that for $k \in \mathcal{R}$ it holds that

$$\delta_k \lesssim (k/d)^{\frac{1}{p}} \log(d/k)^{\frac{1}{2(p-1)}} \tag{36}$$

where $k \in \mathcal{R}$, and $f_k \in \mathcal{V}_k$, and $\|f_k\|_p \asymp d^{1/p-1/2}$. Now, as f_k and f_{ns} lie in different coordinates, it must hold that

$$\|\widehat{f}_n - \delta_k f_k\|_p \leq (1-c_1\delta_k^p)\|\widehat{f}_n\|_p, \tag{37}$$

where we used $\|\widehat{f}_n\|_p \asymp \|f_k\|_p$. Recall that $\mathcal{C}_d \subset B_d \subset \ell_{p,d}$ (as it is in John's position), and by Kashin's theorem we that that

$$cB_n \subset P(\ell_{p,d})$$

with probability of $1-\exp(-cd)$ when $d \geq Cn$. The latter implies that we may find an interpolator $\tilde{f}_k$, to $\frac{\delta_k P(f_k)}{\|P(f_k)\|_2}$ (i.e. $P\tilde{f}_k = \frac{\delta_k P(f_k)}{\|P(f_k)\|_2}$) that has an ℓ_p norm of at most $O(\delta_k d^{1/p-1/2})$. Consider

$$\bar{f}_n := \widehat{f}_n - \delta_k \cdot f_k + \delta_k \|Pf_k\|_2 \cdot \tilde{f}_k,$$

which interpolates $\boldsymbol{\xi}$, and by triangle inequality, we know that

$$\|\bar{f}_n\|_p \leq \|\widehat{f}_n\|_p + d^{1/p-1/2}\cdot(-c\delta_k^p + C\delta_k\|Pf_k\|_2) \leq (1-c_2\delta_k^p + C_1\delta_k\|Pf_k\|_2)\|\widehat{f}_n\|_p, \tag{38}$$

where we used that $\|\widehat{f}_n\|_p \asymp d^{1/p-1/2}$. We know that $\widehat{f}_n$ is the minimal norm solution, and hence we have that

$$-c_2\delta_k^p + C_1\delta_k\|Pf_k\|_2 < 0.$$

Therefore, we balance the following:

$$\begin{aligned}\delta_k^p \asymp \delta_k\|Pf_k\|_2 &\iff \delta_k^{p-1} \asymp \log(d/k)^{1/2}(k/d)^{1-1/p} \iff \delta_k^{p-1} \asymp (k/d)^{\frac{p-1}{p}}\log(d/k)^{1/2} \\ &\iff \delta_k \asymp (k/d)^{\frac{1}{p}}\log(d/k)^{\frac{1}{2(p-1)}}\end{aligned} \tag{39}$$

Therefore, by the definition of $\mathcal{V}_k$, we obtain that

$$\delta_k\|f_k\|_2 \lesssim \delta_k\lambda_k \lesssim \left(\frac{k}{d}\right)^{1/2}\log(d/k)^{\frac{1}{2(p-1)}+\frac{1}{p}-\frac{1}{2}},$$

and clearly this term maximized $k = d/(C_2\log(d))$, and when $p = 1 + C\frac{\log\log\log(d)}{\log\log(d)}$, we obtain that

$$\delta_{d/\log d}\|f_{d/\log(d)}\|_2 \lesssim \log(d)^{-1/2}(\log\log(d))^{c_2(C)\frac{\log\log(d)}{\log\log\log(d)}+\frac{1}{2}} \lesssim \log(d)^{-1/4},$$

where we set $C > 0$ to be large enough. hence the claim follows for all $k \in \mathcal{R}$. Note that *uniformly* the entries of $\delta_k f_k$ are bounded by $O(\log(d)/\sqrt{d})$. To see this, note that

$$\delta_k f_k \in \left(\frac{k\log(d/k)}{d}\right)^{1/p-1/2}\frac{1}{\sqrt{|\sigma|}}B_\infty^\sigma$$

where $|\sigma| \asymp k/\log(d/k)$. Note that Step I holds for *every* realization of $\boldsymbol{\xi}$, under a high probability event over P, random projection the claim follows.

Before we move to the next step, we may assume the following decomposition over $\widehat{f}_n$:

$$\widehat{f}_n = f_{ns} + \sum_{k \in \mathcal{R}'} \delta_k f_k, \tag{40}$$

where $\delta_k \geq 0$, $\|f_{ns}\|_p \asymp d^{1/p-1/2}$, and $\|f_{ns}\|_\infty \lesssim \log(d)/\sqrt{d}$. Furthermore, f_{ns} has $d/\log(d)$ (denoted by σ) such that $|f_i| \gtrsim 1/(\log(d)\sqrt{d})$ for all $i \in \sigma$, $f_k \in \mathcal{V}_k$, and $\|f_k\|_p \asymp d^{1/p-1/2}$.

Step II: In this regime, we will have to use a different argument for two reasons. First, as $k_0 < d$, we are "under" parameterized, and also when $k \leq k_0$, the expectation is the dominant term.

Our proof boils down to the following lemma:

Lemma B.6. *For each $k \in 1, 2, 4 \ldots, k_0$, and $\delta < c(\delta_k, p, d)$, there exists an interpolator $\bar{f}_n$ to $\boldsymbol{\xi}$, i.e. ($P\bar{f}_n = \boldsymbol{\xi}$) such that*

$$\|\bar{f}_n\|_p^p \leq \left(1 + C\delta \cdot (\log(d)^{C_1} \cdot \|Pf_k\|_2 \cdot \delta_k \cdot \sqrt{\frac{k}{n}} - c_1\delta_k^p)\right) \cdot \|\widehat{f}_n\|_p^p,$$

for all $\delta \in (0, c_3(d, n, k))$, where $C_1 \leq 10$.

We will prove this lemma below as it is quite technical in its nature. But let us provide a "wrong" explanation (e.g., a physics proof) of how this lemma is proven.

Consider a fixed $u \in \mathbb{S}^{n-1}$, then we find an interpolator, denoted by f_u, to u (i.e $Pf_u = u$), that lies in

$$\mathcal{C}_d := [-1/\sqrt{d}, 1/\sqrt{d}]^d \subset \ell_{p,d}.$$

and it always exists by Kashin's theorem for sub-Gaussian projections (see Lemma B.3), and we will study the norm of

$$\|\widehat{f}_n + tf_u\|_p,$$

for t sufficiently small. First, we show that in expectation

$$\mathbb{E}\|\widehat{f}_n + tf_u\|_p = (1 + \tilde{\Theta}(t^2))\mathbb{E}\|\widehat{f}_n\|_p$$

Then, we roughly show that

$$|\|\widehat{f}_n + t \cdot f_u\|_p - \|\widehat{f}_n\|_p|$$

is $t\mathbb{E}\|\widehat{f}_n\|_p$ is Lipcshitz in $\boldsymbol{\xi}$ for every $\mathbf{X}$ that lies in a high probablity event. Then, by applying chaining over $P(\mathcal{V}_k)$ (in the sense of Dudley's), we show that for all $f_u = P(f_k)/\|P(f_k)\|_2$ and $t = \delta \cdot \delta_k \cdot \|Pf_k\|_2$ that

$$\|\widehat{f}_n + tf_u\|_p^p \leq \left(1 + t \cdot (C\log(d) \cdot \sqrt{\frac{k}{n}}) + (\delta t)^2\right) \cdot \|\widehat{f}_n\|_p^p,$$

and we may choose sufficiently small $\delta > 0$ to obtain

$$\|\widehat{f}_n + tf_u\|_p^p \leq \left(1 + C \cdot \delta \cdot \delta_k \log(d) \cdot \|Pf_k\|_2 \cdot \sqrt{\frac{k}{n}}\right) \cdot \|\widehat{f}_n\|_p^p$$

and the claim follows by considering

$$\bar{f}_n := \widehat{f}_n + t \cdot f_u - (1 - \delta + O(\delta^2))\delta_k f_k$$

that interpolates $\boldsymbol{\xi}$, as it holds

$$\|\bar{f}_n\|_p^p = (1-\delta)^p\delta_k^p\|f_k\|_p^p + \delta_k^p\|f_{ns} + \delta_k\delta\|_p^p \leq (1 - C\delta \cdot (\delta_k^p - \log(d)\|Pf_k\|_2 \cdot \delta_k \cdot \sqrt{\frac{k}{n}}))\|\widehat{f}_n\|_p.$$

After providing the idea of the proof of the lemma above, we see how this lemma implies the theorem. Note that the last equation *must* be positive for every $\delta > 0$, and hence we balance the following:

$$\log(d)^{C_1} \cdot \sqrt{\frac{k}{n}} \cdot \|Pf_k\|_2 \cdot \delta_k \asymp \delta_k^p \iff \log(d)^{C_1} \cdot \left(\frac{d}{k}\right)^{1-1/p} \delta_k \asymp \delta_k^p$$
$$\iff \log(d)^{\frac{C_1}{(p-1)}} (k/d)^{1/p} \asymp \delta_k$$

which corresponds to the same stationary point of (39). Therefore, it is not hard to verify that

$$\delta_k \|f_k\|_2 \lesssim (k/d)^{1/2} \log(d)^{\frac{C_1}{(p-1)}} \lesssim (k/d)^{1/2} \log(d)^{C_2 \frac{\log\log(d)}{\log\log\log(d)}}$$

and the claim follows. As

$$\|\sum_{k\in\mathcal{R}'} \delta_k f_k\| \lesssim (n/d)^{1/2} \log(d)^{-1/2} \log(d)^{C_1 \frac{\log\log(d)}{\log\log\log(d)}} \lesssim (1/\sqrt{d})^{0.49}.$$

It remains to prove Lemma B.6.

Proof of Lemma B.6. First, we fix an $\mathbf{X}$ that lies in the event of Lemma B.5, and furthermore we assume that $\mathbf{X}$ lies in the event of

$$c_3 \log(d)^{-1} B_2^n \subset \frac{1}{\sqrt{|\sigma|}} \mathbf{X}_\sigma[\mathcal{C}_\sigma] := P_\sigma(\mathcal{C}_\sigma)$$

for all subsets of $\sigma = s$ such that $s = cd/\log(d)$. Note that this event holds with probability of $1 - \exp(-cd/\log(d))$. To see this, note that

$$\binom{d}{c_2 d/\log(d)} \leq \exp(\log(c_2 d/\log(d)) c_2 d/\log(d))$$

by Kashin's theorem (see Lemma B.3), we may choose a small enough $c_3 > 0$ such that

$$c_3 \log(d)^{-2} \cdot B_2^n \subset P_\sigma(\mathcal{C}_\sigma)$$

with probability of

$$1 - \binom{d}{c_2 d/\log(d)} \exp(c_4 \log(\epsilon) d/\log(d)) \geq 1 - \exp(-c_4 d/\log(d))$$

when we choose $\epsilon \simeq \log(d)^{-1}$.

Next, for every such $\mathbf{X}$, we denote the event of $\widehat{f_n}$ has $\Omega(d/\log(d))$-entries that their magnitude is at least $\Omega(1/\log(d)\sqrt{d})$, which holds with a probability of $1 - \exp(-cn)$ (see Lemma B.5 above). We denote it by $A = A_{\mathbf{X}} \subset \mathbb{S}^{n-1}$, and note that $A = -A$, throughout this proof we condition on this event.

Clearly, with high probability event of $\mathbf{X}$, under the event $A = A_{\mathbf{X}}$, for every fixed $u \in \mathbb{S}^{n-1}$, we may find an interpolator to u which we denote by f_u (i.e $Pf_u = u$), that lies in $C\log(d)^2 \cdot \mathcal{C}_\sigma$, and in particular its $\|\cdot\|_{p'}$-norm is at-most $\log(d)$, and note that each entire of f_u is bounded from above by $O(\log(d) \cdot d)$.

Now, for every fixed $\mathbf{X}$ and $tu \in t \cdot \mathbb{S}^{n-1}$, we define the map

$$F_{\mathbf{X},tu} : A_{\mathbf{X}} \subset \mathbb{S}^{n-1} \to \mathbb{R}$$

as follows:

$$\boldsymbol{\xi} \mapsto \|\widehat{f}_n(\boldsymbol{\xi} + tu)\|_p^p = \operatorname*{argmin}_{\{f\in\mathbb{R}^d : \mathbf{f} = \boldsymbol{\xi} + tu\}} \|f\|_p^p - \|\widehat{f}_n(\boldsymbol{\xi})\|_p^p$$

First, we prove the following claim:

Claim 1. *Let $tu \in t \cdot \mathbb{S}^{n-1}$, then under the event of Lemma of (40). There exists a set $A_{tu} \subset A_{\mathbf{X}}$ of measure $0.5 - \exp(-cn/\log(n)$ such that*

$$F_{\mathbf{X},tu}(\boldsymbol{\xi}) \lesssim t^2 \log(d).$$

Proof. Let f_u be the interpolator of u that lies in the cube $\mathcal{C}_{\sigma_{\boldsymbol{\xi}}}$, where $\sigma_{\boldsymbol{\xi}}$ are the s-entries with the largest magnitude of f_{ns} (in terms of our decomposition of $\widehat{f}_n$). Note that entries of f_u depend on $\widehat{f}_n$, and yet we may first consider the conditional expectation over the two points $\boldsymbol{\xi}$,$-\boldsymbol{\xi}$, as the same f_u is chosen by our decision rule.

Note that when t small enough, it holds that $t\|f_u\|_\infty \le c_1(\widehat{f}_n)_i$, for all $i \in \sigma = \sigma_{\boldsymbol{\xi}}$, then we obtain by Taylor's expansion (or using the identity of $(1\pm x)^p \le 1 \pm px + O(p^2x^2)$) that

$$\begin{aligned}
\frac{\min\left\{\|\widehat{f}_n(\pm\boldsymbol{\xi}) + tf_u\|_p^p\right\} - \|\widehat{f}_n(\boldsymbol{\xi})\|_p^p}{d^{1-p/2}} &\le \frac{\mathbb{E}\left[\|\widehat{f}_n + tf_u\|_p^p - \|\widehat{f}_n\|_p^p\right]}{d^{1-p/2}} \\
&\lesssim d^{p/2}\cdot\mathbb{E}\left[\|(\widehat{f}_n + tf_u)_1\|_p^p - \|(\widehat{f}_n)_1\|_p^p\right] \\
&\lesssim d^{p/2}\cdot\left(\mathbb{E}\left[x|(\widehat{f}_n)_1|^{p-1}\operatorname{sign}((\widehat{f}_n)_1)\cdot(f_u)_1 t + |(\widehat{f}_n)_1|^{p-2}\cdot|(f_u)_1|^2\cdot t^2\right]\right) \\
&= d^{p/2}\cdot\mathbb{E}\left[|(\widehat{f}_n)_1|^{p-2}\cdot|(f_u)_1|^2\cdot t^2\right] \\
&\lesssim d^{p/2}\cdot|\frac{1}{\sqrt{d}}|^{p-2}\cdot d^{-1}\cdot t^2 \\
&\lesssim \log(d)\cdot t^2
\end{aligned} \tag{41}$$

and the last inequality for *all* d entries of $\widehat{f}_n$ are greater than $\Omega(1/\sqrt{d\log(d)})$. Therefore, we have that for $\Pr(A_u) \ge \Pr(A_{\mathbf{X}})/ \ge 1/2 - \exp(-cn/\log(n))$ that

$$F_{\mathbf{X},tu}(\boldsymbol{\xi}) \lesssim t^2\log(d).$$

and the claim follows. □

Next, we prove

Claim 2. *Consider* $A_s := \{\boldsymbol{\xi}\in\mathbb{S}^{n-1} : d(\boldsymbol{\xi}, A_{tu}) \le s\}$*, then the following holds for all* $\boldsymbol{\xi}\in A_s$

$$F_{\mathbf{X},tu}(\boldsymbol{\xi}) \lesssim \log(d)\cdot(st + t^2)$$

Using the isoperimetry of the noise implies that by setting $s = \sqrt{\max\{k,\log(d)\}/n}$, we obtain that for every fixed $u\in A_s$

$$F_{\mathbf{X},tu}(\boldsymbol{\xi}) \lesssim \log(d)\cdot\sqrt{\max\{k,\log d\}/n}t + t^2,$$

with probability of at least $1-\exp(-c_2\max\{k,\log(d)\})$. Let $\mathcal{R}(P(\mathcal{V}_k))$ be the radial projection of $\mathcal{V}_k$, and recall the definition of $\mathcal{V}_k$ of (25), and by applying union bound over a net of $\mathcal{V}_k$ with radius $O(n^{-1})$, we obtain that uniformly over $\mathcal{V}_k$ that

$$\begin{aligned}
\sup_{u\in\mathcal{R}(P(\mathcal{V}_k))} F_{\mathbf{X},tu}(\boldsymbol{\xi}) &\le \mathbb{E}F_{\mathbf{X},tu}(\boldsymbol{\xi}) + C\log(d)\cdot\sqrt{k/n}\cdot t \\
&\le \left(1 + C\log(d)\cdot\sqrt{k/n}\cdot t\right)\|\widehat{f}_n\|_p^p
\end{aligned}$$

with probability of $1-\exp(-c_1\max\{k, C\log d\}) \ge 1 - d^{-10}$,

Now, consider the interpolator $\bar{f}_n$ to $\boldsymbol{\xi}$ (i.e. $P\bar{f}_n = \boldsymbol{\xi}$) defined via

$$\bar{f}_n = \widehat{f}_n - \delta\delta_k f_k + \delta\delta_k\|P(f_k)\|_2 f_{Pf_k/\|Pf_k\|}.$$

and note that

$$\|\bar{f}_n\|_p^p \le (1 - \delta\delta_k^p + C\log(d)^{C_3}\cdot\delta\delta_k\cdot\sqrt{k/n}\cdot\|Pf_k\|_2)\|\widehat{f}_n\|_p^p,$$

where $C_3 \le 10$, and the lemma follows. It remains to prove Claim 2 □

Proof of Claim 2. Fix $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ such that $d(\boldsymbol{\xi}', \boldsymbol{\xi}) = s$, and let $\sigma_{\boldsymbol{\xi}} \subset [d]$ to be all the entries of $\widehat{f}_n(\boldsymbol{\xi})$, such that $|(\widehat{f}_n)_i| \gtrsim 1/\sqrt{d\log(d)}$ (clearly σ depends on $\boldsymbol{\xi}$). Now, recall that

$$\frac{\|\widehat{f}_n(\boldsymbol{\xi}) - \widehat{f}_n(\boldsymbol{\xi}')\|_p^p}{d^{1-p/2}} \leq \frac{\|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_n^p}{d^{1-p/2}} \lesssim \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2^p \lesssim s^p.$$

Then, by Markov's inequality, there is $\sigma'_{\boldsymbol{\xi}} \subset \sigma_{\boldsymbol{\xi}}$, such that for all $i \in \sigma'_{\boldsymbol{\xi}}$

$$|\widehat{f}_n(\boldsymbol{\xi}')_i| \gtrsim 1/\sqrt{\log(d)d},$$

and $|\sigma'_{\boldsymbol{\xi}}| \geq (1 - 2s^p)|\sigma|$. Therefore, as $|\widehat{f}_n(\boldsymbol{\xi}')_i| \lesssim 1/\sqrt{d\log(d)}$ over $\sigma_{\boldsymbol{\xi}} \setminus \sigma_{\boldsymbol{\xi}'}$, we obtain

$$\frac{\|(\widehat{f}_n(\boldsymbol{\xi}') + tf_u)1_{\sigma_{\boldsymbol{\xi}} \setminus \sigma'_{\boldsymbol{\xi}}}\|_p^p - \|\widehat{f}_n(\boldsymbol{\xi}')1_{\sigma_{\boldsymbol{\xi}} \setminus \sigma'_{\boldsymbol{\xi}}}\|_p^p}{d^{1-p/2}} \lesssim s^p t d^{p/2} \cdot \max_{i \in \sigma_{\boldsymbol{\xi}} \setminus \sigma'_{\boldsymbol{\xi}}} |\widehat{f}_n(\boldsymbol{\xi}')_i|^{p-1}(f_u)_i \lesssim s^p t \log(d),$$

and therefore

$$F_{\mathbf{X},tu}(\boldsymbol{\xi}') \leq \frac{\left|\|\widehat{f}_n(\boldsymbol{\xi}) + tf_u\|_p^p - \|\widehat{f}_n(\boldsymbol{\xi})\|_p^p - (\|\widehat{f}_n(\boldsymbol{\xi}') + tf_u\|_p^p - \|\widehat{f}_n(\boldsymbol{\xi}')\|_p^p)\right|}{d^{1-p/2}} + C\log(d)(t^2 + s^p t),$$

where we used the last equation. Next, by using the identities of $(1 \pm x)^l = 1 \pm lx + O(l^2x^2)$, we obtain that for $\sigma := \sigma_{\boldsymbol{\xi}'}$, and $(*) = (\left|\|\widehat{f}_n(\boldsymbol{\xi}) + tf_u\|_p^p - \|\widehat{f}_n(\boldsymbol{\xi})\|_p^p - (\|\widehat{f}_n(\boldsymbol{\xi}') + tf_u\|_p^p - \|\widehat{f}_n(\boldsymbol{\xi}')\|_p^p\right|)/d^{1-p/2}$ that

$$\begin{aligned}
d^{1-p/2} \cdot (*) &\lesssim t \cdot \sum_{i \in \sigma} \left(|\widehat{f}_n(\boldsymbol{\xi})_i|^{p-1} - |\widehat{f}_n(\boldsymbol{\xi}')_i|^{p-1}\right) \operatorname{sign}(\widehat{f}_n(\boldsymbol{\xi})_i)u_i + t^2 \sum_{i \in \sigma} |\widehat{f}_n(\boldsymbol{\xi})_i|^{p-2} - |\widehat{f}_n(\boldsymbol{\xi}')_i|^{p-2}u_i^2 \\
&\lesssim t \cdot \sum_{i \in \sigma} \left(|\widehat{f}_n(\boldsymbol{\xi})_i|^{p-1} - |\widehat{f}_n(\boldsymbol{\xi}')_i|^{p-1}\right) \operatorname{sign}(\widehat{f}_n(\boldsymbol{\xi})_i)u_i + O(t^2\log(d)),
\end{aligned}$$

where the last inequality follows from the analysis of (41), and by choosing t small enough (as it is allowed), it is enough to bound the first term. Then, by using that

$$\begin{aligned}
(*) &\lesssim \frac{t}{d^{1-p/2}} \sum_{i \in \sigma} \left||\widehat{f}_n(\boldsymbol{\xi})|_i^{p-1}(f_u)_i - |\widehat{f}_n(\boldsymbol{\xi}')|_i^{p-1}(f_u)_i\right| \\
&\lesssim \frac{t}{\sqrt{d} \cdot d^{1-p/2}} \sum_{i \in \sigma} \left||\widehat{f}_n(\boldsymbol{\xi})|_i^{p-1} - |\widehat{f}_n(\boldsymbol{\xi}')|_i^{p-1}\right| \\
&\lesssim \frac{t}{\sqrt{d} \cdot d^{1-p/2}} \sum_{i \in \sigma} \min\{|\widehat{f}_n(\boldsymbol{\xi})_i|, |\widehat{f}_n(\boldsymbol{\xi}')_i|\}^{p-2}|\widehat{f}_n(\boldsymbol{\xi})_i - \widehat{f}_n(\boldsymbol{\xi}')_i| \\
&\lesssim \frac{t}{\sqrt{d} \cdot d^{1-p/2}} \sum_{i \in \sigma} \left(\sqrt{\frac{1}{\log(d)d}}\right)^{p-2} |\widehat{f}_n(\boldsymbol{\xi})_i - \widehat{f}_n(\boldsymbol{\xi}')_i| \\
&\lesssim \frac{t\log(d)}{\sqrt{d} \cdot d^{1-p/2}} \|\frac{\mathbf{1}}{\mathbf{d^{p/2-1}}}\|_q \|\widehat{f}_n(\boldsymbol{\xi}) - \widehat{f}_n(\boldsymbol{\xi}')\|_p \\
&\lesssim t\log(d)d^{1/2-1/p}\|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_p \\
&\lesssim t\log(d)d^{1/2-1/p} \cdot d^{1/p-1/2}\|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2 \\
&\lesssim \log(d)t\|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2
\end{aligned} \tag{42}$$

where we used that $\|\widehat{f}_n(\boldsymbol{\xi}) - \widehat{f}_n(\boldsymbol{\xi}')\|_p = \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_n \lesssim \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2 \cdot d^{1/p-1/2}$, as $\ell_{p,d}$ is in John's position, and the claim follows. □

B.5.2. PROOF OF LEMMA B.3

Let $P = \frac{1}{\sqrt{d}}\mathbf{X}$, and note showing that $c\epsilon B_n \subset P\left(\mathcal{C}_d\right)$ is equivalent to show that

$$\left(P(\mathcal{C}_d)\right)^\circ \subset C\epsilon^{-1}B_n$$

By duality, as the radial function equals to $1/\|\cdot\|_{K^\circ}$, it is sufficient to show

$$\min_{u\in\mathbb{S}^{n-1}} \|\mathbf{X}^\top u\|_1 \geq \epsilon d.$$

Equivalently,

$$\min_{u\in\mathbb{S}^{n-1}} \sum_{j=1}^{d} |\sum_{i=1}^{n} X_{ij}\cdot u_i| \geq \epsilon d$$

with probability of $1-\exp(cd\log(\epsilon))$. The proof goes via the probabilistic method, for a fix $u\in\mathbb{S}^{n-1}$

$$\sum_{j=1}^{d} |\sum_{i=1}^{n} X_{ij}\cdot u_i| \geq c\epsilon d$$

with probability of $1-\exp(cd\log(\epsilon))$. To see this, note that are cd entries with magnitude $c_1\epsilon$, with a probability of $1-\exp(cd\log(\epsilon))$. We used the fact that

$$\Pr(|X_1\cdot u|\leq\epsilon)\leq C\epsilon.$$

which holds as the entries of X_1 are iid, and have a bounded density, see (Rudelson & Vershynin, 2015). Taking a net of $\epsilon/2$-over the sphere that has a cardinality of $(1+2/\epsilon)^n \leq \exp(-n\log(\epsilon))$ concludes the proof.