BENIGN OVERFITTING FOR GENERAL NORMS AND DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding why predictors can generalize despite interpolating noisy training data is a central puzzle in machine learning. Most work on such "benign overfitting" studies minimum-ℓ_2-norm linear regression, reflecting the inductive bias of gradient descent. However, modern optimizers such as Adam and Muon use non-Euclidean update geometries, favoring solutions associated with other norms. Analyzing regression for non-ℓ_2 norms is substantially more difficult, with known results essentially limited to Gaussian distributions. In this paper, we develop a method to analyze benign overfitting in linear regression for general norms *and* general (sub-Gaussian) distributions. As a special case, we prove that minimum-ℓ_p-norm interpolation with $p>1$ can benignly overfit even for non-Gaussian distributions, under suitable conditions. Perhaps surprisingly, for the ℓ_1 norm, benign overfitting *does not* hold in general for well-behaved (but non-Gaussian) distributions, showing that existing positive ℓ_1 results rely crucially on Gaussianity. Our proof analyzes the geometry of the dual optimization problem, using concentration and central limit tools to show it is approximately Euclidean in many high-dimensional cases.

1 INTRODUCTION

Classical intuition associates overfitting with poor generalization. However, high-dimensional models can often generalize well even when interpolating the training data (namely, achieving zero training error, which results in overfitting if the Bayes risk is positive). This phenomenon is known as benign overfitting (Bartlett et al., 2020; Muthukumar et al., 2020; Belkin et al., 2020; Hastie et al., 2022; Mallinar et al., 2022; Barzilai & Shamir, 2024), and understanding when it occurs and how it depends on the learning algorithm has become a fundamental question in machine learning theory.

Linear regression provides a canonical setting to study benign overfitting. Consider an input matrix $X \in \mathbb{R}^{n\times d}$ with n i.i.d. rows, a target vector $\mathbf{w}^\star \in \mathbb{R}^d$, and noisy labels $\mathbf{y} = X\mathbf{w}^\star + \boldsymbol{\xi}$, where $\boldsymbol{\xi} \in \mathbb{R}^n$ is a vector of independent noise terms. In the overparameterized regime where $d > n$, there are typically infinitely many predictors $\mathbf{w}$ that interpolate this training data, i.e. satisfy $X\mathbf{w} = \mathbf{y}$. Given a norm $\|\cdot\|$ on $\mathbb{R}^d$, a common choice is the minimum-norm interpolator, given by

$$\widehat{\mathbf{w}} \in \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w}\| \quad \text{subject to} \quad X\mathbf{w} = \mathbf{y}.$$

In this setting, benign overfitting corresponds to $\hat{\mathbf{w}}$ achieving approximately optimal performance (comparable to $\mathbf{w}^\star$), despite overfitting the noisy labels.

The large majority of works on linear regression and benign overfitting consider the Euclidean ℓ_2 norm. This choice is natural from an optimization perspective: In overparameterized linear regression, gradient descent initialized at zero is well-known to converge to the minimum-ℓ_2-norm interpolator. However, modern machine learning increasingly uses different optimizers such as Adam (Kingma & Ba, 2015) and Muon (Jordan et al., 2024), and it is well known that the choice of optimizer affects the solution reached at convergence (Gunasekar et al., 2018). For example, in various regimes, Adam and SignSGD are known to select the minimum-ℓ_∞-norm solution (Zhang et al., 2024; Bernstein & Newhouse, 2024; Gronich & Vardi, 2026). This underscores the importance of understanding benign overfitting beyond minimum-ℓ_2-norm interpolation.

Unfortunately, many techniques developed for the minimum-ℓ_2-norm interpolator crucially rely on it having a closed form, which is not typically true for other norms. For non-ℓ_2 norms, existing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

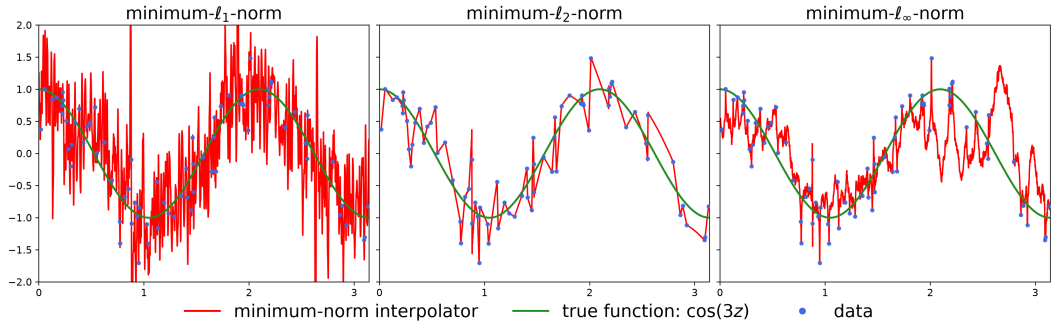

Figure 1: Minimum-ℓ_p interpolation for $p \in \{1, 2, \infty\}$, over i.i.d inputs $\mathbf{x} \in \mathbb{R}^{2000}$ with coordinates $x_j = \cos(jz)/j$ where $z \sim \mathrm{Unif}([0, \pi])$. The number of inputs is $n = 75$ and the labels are $y = \cos(3z) + \xi$, with $\xi \sim \mathcal{N}(0, 0.1)$. The horizontal axis shows z.

works focus on Gaussian inputs (and often also Gaussian noise), and use tools that crucially rely on this assumption, such as the Gaussian Minmax theorem (Koehler et al., 2021; Kur & Pathak, 2026). This is not just a technical limitation: As we will show later in this paper, the existence of benign overfitting can be very different for Gaussian and non-Gaussian distributions, even for well-studied settings such as minimum ℓ_1-norm interpolation. The previous discussions motivate the central question underpinning this work:

When and why do non-ℓ_2 minimum-norm interpolators exhibit benign overfitting in standard (possibly non-Gaussian) high-dimensional regression settings?

As in many previous works on benign overfitting (e.g. Bartlett et al. (2020); Koehler et al. (2021); Shamir (2023)), our analysis partitions the d-dimensional inputs into a small set S of "signal coordinates" used to fit the targets, and the remaining set N of "noise-fitting" coordinates used to fit the noise without incurring large excess risk. This split is used only in the analysis, and the algorithm does not rely on it. Prior work on the ℓ_2 interpolator shows that benign overfitting can occur when S is relatively low-dimensional and contains most of the target, while N has a large effective dimension (meaning that there are many directions with similar variance).

Our contributions and paper structure can be summarized as follows:

- We begin in Sec. 3 by considering the minimum-ℓ_p-norm interpolator for any fixed $p \in (1, \infty]$. We prove a quantitative risk bound that implies benign overfitting, under suitable assumptions broadly similar to those in previous works (Theorem 1 and Corollary 1). Specifically, suppose that the input and label-noise are independent and sub-Gaussian, the noise-fitting coordinates have common variance s_N^2, and the target vector $\mathbf{w}^\star$ is supported on S. Writing q for the dual exponent, defined by $1/p + 1/q = 1$, our results imply benign overfitting in the regime
$$s_N^2 |N|^{2/q} |S|^{\max\{2/p-1,0\}} \ll n \ll |N|^{\min\{2/q,1\}}.$$
For ℓ_2, this matches the necessary and sufficient criteria implied by Bartlett et al. (2020).
- In Sec. 4, we explain our method and main proof ideas for analyzing the minimum-norm interpolator. Roughly speaking, our analysis begins by leveraging the dual objective of the min-norm interpolation problem. We then demonstrate that under suitable conditions on the norm (which we prove hold for ℓ_p norms when $p \in (1, \infty]$), and for high-dimensional inputs where the coordinates are independent and sub-Gaussian, the dual problem can be well-approximated by a least-squares problem with ℓ_2 geometry, leading to a behavior similar to the ℓ_2 case. The techniques used in this approach, which include quantitative central limit and uniform concentration arguments, may be of independent interest.
- In Sec. 5, we show that benign overfitting *does not* generally hold for minimum-ℓ_1-norm interpolation. For scaled Rademacher inputs and suitable independent sub-Gaussian noise, the excess risk diverges as $\Omega(\sigma^2 \log n)$ with constant probability (Theorem 2). This is perhaps surprising in light of ℓ_1 benign overfitting results for *Gaussian* distributions (Koehler et al., 2021; Wang et al., 2022). Nevertheless, we demonstrate that the geometry of the minimum-ℓ_1-norm regression problem can be completely different between Gaussians and other well-behaved distributions. This is

also illustrated in Fig. 1, which shows a simple empirical example where the minimum-ℓ_1 interpolator clearly behaves much worse than the minimum-ℓ_2 and ℓ_∞-norm interpolators, in terms of generalization behavior.

- In Sec. 6, we generalize our theory to provide risk bounds for generic norms (beyond ℓ_p), and general sub-Gaussian distributions with independent coordinates (Theorem 3). The theorem identifies two conditions on N that suffice for benign overfitting, related to its expected dual norm on flat directions, and uniform concentration of the dual norm over the sphere. We prove these conditions indeed hold for ℓ_p norms as a special case.

All proofs are provided in the appendices. We defer discussion of related work to Appendix A.

2 Setting and Preliminaries

Consider an input distribution $\mathcal{D}_{\mathbf{x}}$ on $\mathbb{R}^d$. Let $\mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathcal{D}_{\mathbf{x}}$ be i.i.d. input samples, forming a matrix $X \in \mathbb{R}^{n \times d}$ with rows $\mathbf{x}_i^\top$. Let $\boldsymbol{\xi} := (\xi_1, \dots, \xi_n)^\top$ be a noise vector, $\mathbf{w}^\star \in \mathbb{R}^d$ be a target vector, and define the labels by $\mathbf{y} := X\mathbf{w}^\star + \boldsymbol{\xi}$. Throughout the paper, unless stated otherwise, $\|\cdot\|$ denotes an arbitrary norm on $\mathbb{R}^d$. For such a norm, the minimum-norm interpolator is any

$$\widehat{\mathbf{w}} \in \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^d} \{\|\mathbf{w}\| : X\mathbf{w} = \mathbf{y}\}.$$

For any $\mathbf{w} \in \mathbb{R}^d$, we define its excess risk as

$$\mathcal{R}(\mathbf{w}) := \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}} \left[\left(\mathbf{x}^\top (\mathbf{w} - \mathbf{w}^\star) \right)^2 \right].$$

Our goal is to identify conditions under which $\mathcal{R}(\widehat{\mathbf{w}}) \to 0$ (where $\hat{\mathbf{w}}$ is a minimum-norm interpolator), which is the standard notion of benign overfitting in the literature.

For our analysis, following previous works, we partition the input coordinates $\{1, \dots, d\}$ into two nonempty sets $S \sqcup N$, where S is the *signal block* (thought of as low dimensional, and ideally $\mathbf{w}^\star$ is mostly supported on S) and N is the *noise-fitting block* (thought of as high dimensional). After relabeling the coordinates, we may assume that $S = \{1, \dots, |S|\}$ and $N = \{|S|+1, \dots, d\}$. This is purely an indexing convention and does not affect the algorithm in any way.

Assumptions. For a random variable Z, write $\|Z\|_{\psi_2}$ for its sub-Gaussian norm[1], and call Z sub-Gaussian if $\|Z\|_{\psi_2} < \infty$ (see Vershynin (2026) for examples, e.g. bounded or Gaussian variables). We assume that the inputs and noise have independent, zero-mean sub-Gaussian coordinates:

Assumption 1 (Sub-Gaussian input distribution)**.** *The inputs $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,d}) \sim \mathcal{D}_{\mathbf{x}}$ have independent coordinates, and there are $b_1, \dots, b_d > 0$ and $R_X > 0$ such that for every $j \in [d]$,*

$$\mathbb{E}[x_j] = 0, \qquad \mathbb{E}[x_j^2] = b_j^2, \qquad \|x_j\|_{\psi_2} \le R_X b_j.$$

Assumption 2 (Sub-Gaussian noise)**.** *The noise vector $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^\top$ has independent coordinates that are also independent of X. There are $\sigma > 0$ and $R_\xi > 0$ such that for every $i \in [n]$,*

$$\mathbb{E}[\xi_i] = 0, \qquad \mathbb{E}[\xi_i^2] = \sigma^2, \qquad \|\xi_i\|_{\psi_2} \le R_\xi \sigma.$$

Assumptions 1 and 2 are rather standard in the literature. While we adopt them for simplicity, as Sec. 4 clarifies, our analysis can extend to milder assumptions, but this may impact the exact bounds.

Notation and additional definitions. For $m \in \mathbb{N}$ let $[m] := \{1, \dots, m\}$. For $T \subseteq [d]$, we write $\mathbf{v}_T$ for the restriction of a vector $\mathbf{v}$ to the coordinates in T, and X_T for the submatrix of X with columns indexed by T. For a partition $d = S \sqcup N$ of the input coordinates write

$$\mathbf{w} = (\mathbf{w}_S, \mathbf{w}_N), \qquad \mathbf{w}^\star = (\mathbf{w}_S^\star, \mathbf{w}_N^\star), \qquad X = (X_S, X_N).$$

Moreover, for any nonempty $T \subseteq [d]$, write:

$$b_T^- := \min_{j \in T} b_j, \qquad b_T^+ := \max_{j \in T} b_j, \qquad \Sigma_T := \mathbb{E}[\mathbf{x}_T \mathbf{x}_T^\top] = \operatorname{diag}\big((b_j^2)_{j \in T}\big),$$

and define the effective rank

$$r_T := \frac{\operatorname{tr}(\Sigma_T)}{\|\Sigma_T\|_{\mathrm{op}}} = \frac{\|\mathbf{b}_T\|_2^2}{(b_T^+)^2}.$$

[1] Defined by $\|Z\|_{\psi_2} := \inf\left\{t > 0 : \mathbb{E}\exp\left(Z^2/t^2\right) \le 2\right\}$

3 BENIGN OVERFITTING FOR ℓ_p NORMS ($p \in (1, \infty]$)

For any $p \in (1, \infty]$, let $q \in [1, \infty)$ be its dual exponent, given by $1/p + 1/q = 1$. To reduce the amount of notation in this section, we will make two simplifying assumptions, which will both be relaxed in Section 6. First, we assume that the coordinates in N share the same variance. While this is not strictly necessary, it simplifies the final bound while capturing the important underlying phenomenon. Moreover, it is well-known that for benign overfitting to occur, a necessary condition (in the ℓ_2 analyses and for general norms under Gaussian inputs) is essentially that there are many coordinates with similar variance (Bartlett et al., 2020; Koehler et al., 2021). Second, we assume that the target vector $\mathbf{w}^\star$ is supported on the S coordinates, meaning $\mathbf{w}^\star = (\mathbf{w}^\star_S, 0)$. For a more general target vector, the risk bounds would incur additional factors related to $\|\Sigma_N^{1/2}\mathbf{w}^\star_N\|_2$. Once again, such terms must be small anyway for benign overfitting to generally occur (Bartlett et al., 2020; Koehler et al., 2021). We are now ready to state our theorem:

Theorem 1 (Risk bound for ℓ_p norms)**.** *Let $[d] = S \sqcup N$ be any partition of coordinates, and suppose Assumptions 1 and 2 hold, that $\mathbf{w}^\star = (\mathbf{w}^\star_S, \mathbf{0})$ and that $(x_j)_{j\in N}$ have common variance s_N^2. For any $p \in (1, \infty]$, there is a positive $C = \mathrm{poly}\big(R_X, R_\xi, \sigma, \sigma^{-1}, (b_S^-)^{-1}, b_S^+, \|\mathbf{w}^\star_S\|_p\big)$ such that, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$ every minimum-ℓ_p-norm interpolator satisfies*

$$\mathcal{R}(\widehat{\mathbf{w}}) \le C\left[\frac{s_N|N|^{\frac{1}{q}}|S|^{\max\left\{\frac{1}{p}-\frac{1}{2},0\right\}}}{\sqrt{n}} + \left(\frac{r_S + \log(\frac{n}{\delta})}{n}\right)^{\frac{1}{4}} + \sqrt{\frac{n\log\left(\frac{n|N|}{\delta}\right)}{|N|}} + \frac{n\log\left(\frac{n|N|}{\delta}\right)}{|N|^{\min\left\{\frac{2}{q},1\right\}}}\right],$$

provided that the quantity in square brackets is at most C^{-1}. Moreover, C is non-increasing in p.

We now provide an asymptotic corollary that clarifies the conditions for benign overfitting. To formally avoid issues related to the order of limits, take $n \to \infty$ and view all other parameters (e.g. N, s_N, δ, S) as functions of n, yielding the following corollary:

Corollary 1 (Benign overfitting for ℓ_p norms)**.** *Consider a sequence of problem instances indexed by the sample size n, each satisfying the assumptions of Theorem 1. Suppose that $R_X, R_\xi, \sigma, \sigma^{-1}, (b_S^-)^{-1}, b_S^+$, and $\|\mathbf{w}^\star_S\|_p$ are uniformly bounded. If there is a sequence $\delta_n \to 0$ such that $\frac{r_S + \log(n/\delta_n)}{n} \to 0$ and if*

$$\frac{\big(n\log(n|N|/\delta_n)\big)^{\max\{1,q/2\}}}{|N|} \longrightarrow 0, \qquad \frac{s_N|N|^{1/q}|S|^{\max\{1/p-1/2,0\}}}{\sqrt{n}} \longrightarrow 0,$$

then, with probability tending to one, every minimum-ℓ_p-norm interpolator satisfies

$$\mathcal{R}(\widehat{\mathbf{w}}) \longrightarrow 0.$$

Analogously to related settings studied in prior work, the regime in which the results are meaningful is when relatively few coordinates can fit the targets, and many low-variance coordinates are used to interpolate the noise. Accordingly, the asymptotic conditions proved sufficient by Theorem 1 and Corollary 1 for benign overfitting are

$$r_S \ll n\,, \qquad \text{and} \qquad s_N^2|N|^{2/q}|S|^{\max\{2/p-1,0\}} \ll n \ll |N|^{\min\{2/q,1\}}. \tag{1}$$

Since $r_S \le |S|$, the first condition is always satisfied when $|S| \ll n$. The $n \ll |N|^{\min\{2/q,1\}}$ becomes increasingly demanding as $p \to 1$, but as we will later show in Sec. 5, this is not an artifact of our analysis: benign overfitting can fail for $p = 1$ even under the assumptions of Theorem 1.

In the ℓ_2 case and under the setting of Theorem 1 and Corollary 1, the asymptotic criteria for benign overfitting in our results match those of Bartlett et al. (2020), which were proved to be both necessary and sufficient. We remark that their results are stated in terms of various notions of effective rank, which can be computed explicitly in this setting to match (1). Likewise, the criteria for benign overfitting given in (1) match those of Koehler et al. (2021) for any $p \in [2, \infty]$ (although the exact convergence rates vary). However, unlike our results, that paper crucially requires the inputs to be Gaussian, e.g. by relying on the Convex Gaussian Minmax Theorem (Thrampoulidis et al., 2015).

The conditions of Theorem 1 and Corollary 1 are sufficient but not necessary. We will later prove in Sec. 6 a more general version (Theorem 3) that holds for general norms and drops both the common-variance and $\mathbf{w}^\star$-support assumptions.

4 PROOF IDEAS

We now outline the techniques underlying the results. We focus here on the ℓ_p norm case, but as we will further discuss in Sec. 6, the proof can readily be adapted to more general norms.

4.1 WARM-UP: GAUSSIAN INPUTS

For simplicity, we first consider the case where the inputs $\mathbf{x}$ are Gaussian, as it provides some useful intuition, and then detail how the framework extends to other distributions.

Fix $p \in (1, \infty)$, and let q be its dual exponent, defined by $1/p + 1/q = 1$ (our proof also handles $p = \infty$, but we omit it here for simplicity). Write $\mathbf{w} = (\mathbf{w}_S, \mathbf{w}_N)$. We decompose the minimum-norm problem by first minimizing over $\mathbf{w}_S$, and then minimizing $\mathbf{w}_N$ as a function of $\mathbf{w}_S$:

$$\begin{aligned}\min_{\mathbf{w}:X\mathbf{w}=\mathbf{y}} \|\mathbf{w}\|_p &= \min_{\mathbf{w}:X\mathbf{w}=\mathbf{y}} \left[\|\mathbf{w}_S\|_p^p + \|\mathbf{w}_N\|_p^p\right]^{1/p} \\ &= \min_{\mathbf{w}_S \in \mathbb{R}^S} \left[\|\mathbf{w}_S\|_p^p + \min_{\mathbf{w}_N : X_N \mathbf{w}_N = \mathbf{y} - X_S \mathbf{w}_S} \|\mathbf{w}_N\|_p^p\right]^{\frac{1}{p}}. \end{aligned} \tag{2}$$

We first focus on the inner minimization over $\mathbf{w}_N$, and later show that under the conditions of the theorem, this is the dominating term. We define the cost of fitting $\mathbf{w}_N$ and the residual as:

$$g(\mathbf{w}_S) := \min_{\mathbf{w}_N : X_N \mathbf{w}_N = \mathbf{z}(\mathbf{w}_S)} \|\mathbf{w}_N\|_p, \qquad \mathbf{z}(\mathbf{w}_S) := \mathbf{y} - X_S \mathbf{w}_S = X_S(\mathbf{w}_S^\star - \mathbf{w}_S) + \boldsymbol{\xi}.$$

A central part of our approach is to analyze $g(\mathbf{w}_S)$ through convex duality. Specifically, the minimization problem defining $g(\mathbf{w}_S)$ can equivalently be written in its dual form as:

$$g(\mathbf{w}_S) = \max_{\boldsymbol{\lambda} \in \mathbb{R}^n : \|X_N^\top \boldsymbol{\lambda}\|_q \le 1} \mathbf{z}(\mathbf{w}_S)^\top \boldsymbol{\lambda}. \tag{3}$$

Let $B_n(r) := \{\boldsymbol{\lambda} \in \mathbb{R}^n : \|\boldsymbol{\lambda}\|_2 \le r\}$ be the ℓ_2 ball of radius r. We will now show that the feasible set in (3) is well-approximated by $B_n(r)$ for a suitably chosen r. Specifically, when the rows of X_N are i.i.d. copies of a Gaussian vector $\mathbf{x}_N$, then, writing $\kappa_N := \mathbb{E}\|\mathbf{x}_N\|_q$, for every $\boldsymbol{\lambda} \in \mathbb{R}^n$,

$$X_N^\top \boldsymbol{\lambda} \overset{d}{=} \|\boldsymbol{\lambda}\|_2 \mathbf{x}_N \qquad \Longrightarrow \qquad \mathbb{E}\|X_N^\top \boldsymbol{\lambda}\|_q = \kappa_N \|\boldsymbol{\lambda}\|_2. \tag{4}$$

When $\mathbf{x}_N$ is high-dimensional, one can further show that, with high probability, $\|X_N^\top \boldsymbol{\lambda}\|_q$ is close to its expectation uniformly over $\boldsymbol{\lambda}$, meaning that

$$(1-\varepsilon)\kappa_N \|\boldsymbol{\lambda}\|_2 \le \|X_N^\top \boldsymbol{\lambda}\|_q \le (1+\varepsilon)\kappa_N \|\boldsymbol{\lambda}\|_2 \qquad \forall\, \boldsymbol{\lambda} \in \mathbb{R}^n, \tag{5}$$

for some small $\varepsilon > 0$. On this event,

$$B_n\left(\frac{1}{(1+\varepsilon)\kappa_N}\right) \subseteq \{\boldsymbol{\lambda} \in \mathbb{R}^n : \|X_N^\top \boldsymbol{\lambda}\|_q \le 1\} \subseteq B_n\left(\frac{1}{(1-\varepsilon)\kappa_N}\right).$$

Thus, the dual problem in (3) is maximizing a linear function over an approximate Euclidean ball, from which it readily follows that

$$\frac{\|\mathbf{z}(\mathbf{w}_S)\|_2}{(1+\varepsilon)\kappa_N} \le g(\mathbf{w}_S) \le \frac{\|\mathbf{z}(\mathbf{w}_S)\|_2}{(1-\varepsilon)\kappa_N}. \tag{6}$$

At the same time, recalling that $\mathbf{z}(\mathbf{w}_S) := X_S(\mathbf{w}_S^\star - \mathbf{w}_S) + \boldsymbol{\xi}$ and that the noise ξ_i has mean-zero and variance σ^2, it holds for sufficiently large n that with high probability, the residuals $\mathbf{z}(\mathbf{w}_S)$ can be well approximated as

$$\begin{aligned}\frac{1}{n}\|\mathbf{z}(\mathbf{w}_S)\|_2^2 &= (\mathbf{w}_S - \mathbf{w}_S^\star)^\top \left(\frac{1}{n} X_S^\top X_S\right)(\mathbf{w}_S - \mathbf{w}_S^\star) - \frac{2}{n}(\mathbf{w}_S - \mathbf{w}_S^\star)^\top X_S^\top \boldsymbol{\xi} + \frac{1}{n}\|\boldsymbol{\xi}\|_2^2 \\ &\approx \left\|\Sigma_S^{1/2}(\mathbf{w}_S - \mathbf{w}_S^\star)\right\|_2^2 + \sigma^2. \end{aligned} \tag{7}$$

Overall, we have shown that for any fixed $\mathbf{w}_S$,

$$\min_{\mathbf{w}_N : X_N \mathbf{w}_N = \mathbf{y} - X_S \mathbf{w}_S} \|\mathbf{w}_N\|_p \approx \frac{\|\mathbf{z}(\mathbf{w}_S)\|_2}{\kappa_N} \approx \frac{\sqrt{n}}{\kappa_N} \cdot \sqrt{\sigma^2 + \left\|\Sigma_S^{1/2}(\mathbf{w}_S - \mathbf{w}_S^\star)\right\|_2^2}. \tag{8}$$

Substituting this into (2), the original min-norm problem is roughly equivalent to:

$$\min_{\mathbf{w}:X\mathbf{w}=\mathbf{y}} \|\mathbf{w}\|_p \approx \min_{\mathbf{w}_S\in\mathbb{R}^S} \left[\|\mathbf{w}_S\|_p^p + \left(\frac{n}{\kappa_N^2}\right)^{p/2} \left(\sigma^2 + \left\|\Sigma_S^{1/2}(\mathbf{w}_S - \mathbf{w}_S^\star)\right\|_2^2\right)^{p/2}\right]^{\frac{1}{p}}.$$

When $n \gg \kappa_N^2$, the dominating term that we can minimize is $\|\Sigma_S^{1/2}(\mathbf{w}_S - \mathbf{w}_S^\star)\|_2^2$, so the min-norm problem is roughly equivalent to a least-squares problem on the signal block,

$$\min_{\mathbf{w}_S} \left\|\Sigma_S^{1/2}(\mathbf{w}_S - \mathbf{w}_S^\star)\right\|_2^2.$$

Therefore, recalling that $\kappa_N := \mathbb{E}[\|\mathbf{x}_N\|_q] \approx s_N|N|^{1/q}$, whenever $n \gg \kappa_N^2 \approx s_N^2|N|^{2/q}$ and Σ_S is well-conditioned, the minimum-ℓ_p-norm solution $\hat{\mathbf{w}}$ must approximate $\mathbf{w}^\star$ on the signal block:

$$\hat{\mathbf{w}}_S \approx \mathbf{w}_S^\star.$$

It remains to show that the effect of the noise-fitting block on the population error is negligible. To see this, when $\hat{\mathbf{w}}_S \approx \mathbf{w}_S^\star$, (8) implies that $\|\hat{\mathbf{w}}_N\|_p \approx \frac{\sqrt{n}}{\kappa_N}\sigma$. Using the standard inequality between the ℓ_2 norm and the ℓ_p norm, and assuming that $\mathbf{w}_N^\star = 0$,

$$\begin{aligned}\|\Sigma_N^{1/2}(\widehat{\mathbf{w}}_N - \mathbf{w}_N^\star)\|_2^2 &\le \|\Sigma_N\|_{\mathrm{op}} \|\widehat{\mathbf{w}}_N\|_2^2 \le |N|^{\max\{1-\frac{2}{p},0\}} \|\Sigma_N\|_{\mathrm{op}} \|\widehat{\mathbf{w}}_N\|_p^2 \\ &\approx |N|^{\max\{1-\frac{2}{p},0\}} \|\Sigma_N\|_{\mathrm{op}} \frac{n}{\kappa_N^2}\sigma^2.\end{aligned}$$

Finally, since $\kappa_N \asymp s_N|N|^{1/q}$ and $\|\Sigma_N\|_{\mathrm{op}} = s_N^2$, the last display is of order $\sigma^2 n/|N|^{\min\{2/q,1\}}$. Thus, ignoring logarithmic factors, it converges to 0 when $n \ll |N|^{\min\{2/q,1\}}$. In such a case, using $\hat{\mathbf{w}}_S \approx \mathbf{w}_S^\star$, the excess risk is small:

$$\mathcal{R}(\widehat{\mathbf{w}}) = \|\Sigma_S^{1/2}(\widehat{\mathbf{w}}_S - \mathbf{w}_S^\star)\|_2^2 + \|\Sigma_N^{1/2}(\widehat{\mathbf{w}}_N - \mathbf{w}_N^\star)\|_2^2 \approx 0.$$

4.2 Beyond Gaussians

For non-Gaussian inputs, the high-level proof approach is similar, but several arguments from the Gaussian proof require adaptation and additional ideas. The first is that for general input distributions, the rotational invariance used in (4) no longer holds. Nevertheless, we can get an approximate version using a central limit theorem (CLT) argument. To see this, first note that the vectors $X_N^\top\boldsymbol{\lambda}$ which determine the dual constraints in (3) have coordinates $(X_N^\top\boldsymbol{\lambda})_j = \sum_{i=1}^n \lambda_i x_{i,j}$. If one had $\lambda_i = 1/\sqrt{n}$ for every i, the CLT would let us approximate $(X_N^\top\lambda)_j$ with a suitable Gaussian. It is well known that in some cases, similar results hold when $\boldsymbol{\lambda}$ is "flat", in the sense that $\|\boldsymbol{\lambda}\|_\infty \ll \|\boldsymbol{\lambda}\|_2$. Lemma 13 in the appendix makes this idea quantitative, and leads to a statement analogous to (4) for flat $\boldsymbol{\lambda}$. Specifically, letting $\kappa_N := s_N\left(|N|\,\mathbb{E}_{G\sim\mathcal{N}(0,1)}[|G|^q]\right)^{1/q}$, the lemma states that, for every $\boldsymbol{\lambda} \in \mathbb{R}^n$,

$$\left|\frac{\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q}{\kappa_N\|\boldsymbol{\lambda}\|_2} - 1\right| \lesssim |N|^{-1/(2q)} + \frac{\|\boldsymbol{\lambda}\|_\infty}{\|\boldsymbol{\lambda}\|_2}.$$

Additionally, Proposition 2 in the appendix shows that $\|X_N^\top\boldsymbol{\lambda}\|_q$ is close to its expectation via uniform concentration, allowing us to obtain a result analogous to (5), namely that with high probability

$$\sup_{\boldsymbol{\lambda}\in\mathbb{S}^{n-1}} \left|\frac{\|X_N^\top\boldsymbol{\lambda}\|_q}{\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q} - 1\right| \lesssim \sqrt{\frac{n}{|N|}} + \frac{n^{q/2}}{|N|} + |N|^{-1/(2q)}$$

(ignoring log factors). In principle, combining the last two displayed equations implies that $\|X_N^\top\boldsymbol{\lambda}\|_q$ tends to be close to $\kappa_N\|\boldsymbol{\lambda}\|_2$, leading to the same Euclidean geometry as before and allowing our argument to go through. The remaining challenge is that the $\boldsymbol{\lambda}$ maximizing the dual problem in (3) does not necessarily need to be flat (in which case $\frac{\|\boldsymbol{\lambda}\|_\infty}{\|\boldsymbol{\lambda}\|_2}$ is not small, as required for the CLT argument). Nevertheless, as we will soon explain in more detail, Lemma 3 in the appendix shows that we can find a *close to optimal* $\boldsymbol{\lambda}$ that is flat (and thus the previous CLT argument applies) whenever $\mathbf{z}(\mathbf{w}_S)$ is flat. Moreover, we show that the existence of a near-optimal flat $\boldsymbol{\lambda}$ is enough for

the argument to go through. Specifically, we first show that $\mathbf{z}(\mathbf{w}_S)$ is flat: Due to the independent sub-Gaussian noise variables, we can show that up to some error terms,

$$\frac{\|\mathbf{z}(\mathbf{w}_S)\|_\infty}{\|\mathbf{z}(\mathbf{w}_S)\|_2} \lesssim \sqrt{\frac{\log(n)}{n}}.$$

Indeed, proving that the denominator scales as $\sqrt{n}$ is done as in (7), and the numerator scale of $\sqrt{\log n}$ arises from the maximum of n independent noise variables. Next, for every feasible $\boldsymbol{\lambda}$, one can zero out all coordinates bigger than some threshold $\tau > 0$, giving $\boldsymbol{\lambda}_\tau := \sum_{j=1}^n \mathbf{1}_{|\lambda_j| \le \tau} \lambda_j \mathbf{e}_j$. Then we prove that for the dual objective given in (3), after slight rescaling, $\boldsymbol{\lambda}_\tau$ is feasible and close to optimal when $\|\mathbf{z}(\mathbf{w}_S)\|_\infty \ll \|\mathbf{z}(\mathbf{w}_S)\|_2$:

$$\mathbf{z}(\mathbf{w}_S)^\top \boldsymbol{\lambda}_\tau \gtrsim \mathbf{z}(\mathbf{w}_S)^\top \boldsymbol{\lambda} - \frac{\|\mathbf{z}(\mathbf{w}_S)\|_\infty}{\tau} \gtrsim \mathbf{z}(\mathbf{w}_S)^\top \boldsymbol{\lambda} - \frac{\|\mathbf{z}(\mathbf{w}_S)\|_2}{\tau}\sqrt{\frac{\log(n)}{n}}. \tag{9}$$

There is now a tradeoff: a smaller choice of τ is better for the quantitative CLT argument, but causes the gap between $\mathbf{z}(\mathbf{w}_S)^\top \boldsymbol{\lambda}_\tau$ and $\mathbf{z}(\mathbf{w}_S)^\top \boldsymbol{\lambda}$ to increase. Taking $\tau \approx n^{-1/4}$ balances both, meaning that $\boldsymbol{\lambda}_\tau$ both satisfies the assumptions needed for the CLT argument, and is an approximate maximizer of the dual objective given in (3).

Combining the ingredients discussed in this subsection allows us to obtain a result analogous to (6), up to the error terms appearing in the theorem. The remaining details of the proof transfer from the Gaussian case more straightforwardly.

5 A Failure Example for Minimum-ℓ_1-Norm Interpolation

Our risk bounds in Thm. 1 give sufficient conditions for benign overfitting for any ℓ_p norm with $p \in (1, \infty]$, but not for ℓ_1. Indeed, the concentration and CLT arguments (described in Sec. 4), which showed that the dual optimization problem is approximately Euclidean, do not hold when $p = 1$. One may wonder if this gap is a proof artifact or reflects genuinely different behavior of the minimum-ℓ_1-norm predictor. As noted in the introduction, it is known that over *Gaussian* inputs (with suitable covariance), the minimum-ℓ_1 norm does exhibit benign overfitting (Ju et al., 2020; Koehler et al., 2021; Wang et al., 2022), with the excess risk scaling as $\sigma^2/\log(d/n) \to 0$ when $d/n \to \infty$. Perhaps surprisingly, we show that despite such results for Gaussians, minimum ℓ_1 interpolation does *not* generally enjoy benign overfitting. In particular, even for well-behaved mean-zero sub-Gaussian inputs and noise (satisfying Assumptions 1 and 2), the risk of the minimum-ℓ_1-norm predictor can diverge to infinity:

Theorem 2 (Failure of minimum-ℓ_1-norm interpolation)**.** *Fix $n \ge 2$, $d > 2n2^n$, $\sigma^2 > 0$, and inputs*

$$x_1 \sim \mathrm{Unif}\{\pm 1\}, \qquad \textit{and} \qquad x_j \sim \mathrm{Unif}\{\pm s_N\} \quad \forall\, j \in \{2, \dots, d\} \quad \textit{where} \;\; s_N \in (0, 1),$$

where coordinates are independent (and sub-Gaussian). There exists independent sub-Gaussian noise ξ with Var$(\xi) = \sigma^2$ such that for $\mathbf{w}^\star = \mathbf{0}$, with probability at least $1/5$, every minimum-ℓ_1-norm interpolator satisfies

$$\mathcal{R}(\widehat{\mathbf{w}}) \ge \sigma^2 \frac{\log n}{9}.$$

Moreover, all distributional and structural assumptions in Thm. 1 are satisfied: for $S = \{1\}, N = \{2, \dots, d\}$, Assumptions 1 and 2 hold with $b_j = s_N$ for all $j \in N$, and with R_X, R_ξ bounded by universal constants, uniformly in n, $|N|$, and s_N.

The construction in Thm. 2 uses dimension exponential in the number of samples. Over Gaussian inputs, the results of Koehler et al. (2021) yield benign overfitting for this same covariance and exponentially large dimension.

We now give intuition as to how the change from Gaussians to (scaled) Rademacher inputs completely changes the nature of the minimum-ℓ_1 interpolation problem. As in the proof of our positive results (cf. Sec. 4) we begin by decomposing the minimum-ℓ_1-norm interpolation problem as

$$\min_{\mathbf{w}:\, X\mathbf{w} = \mathbf{y}} \|\mathbf{w}\|_1 = \min_{w_S \in \mathbb{R}} \left(|w_S| + \min_{\mathbf{w}_N : X_N \mathbf{w}_N = \mathbf{y} - X_S w_S} \|\mathbf{w}_N\|_1 \right). \tag{10}$$

Now suppose that the inputs have N coordinates in $\{\pm s_N\}$ and with high probability every vector in $\{\pm s_N\}^n$ appears as a column of X_N. On this event, for any $\boldsymbol{\lambda} \in \mathbb{R}^n$,

$$\|X_N^\top \boldsymbol{\lambda}\|_\infty = s_N \max_{\mathbf{u} \in \{\pm 1\}^n} |\mathbf{u}^\top \boldsymbol{\lambda}| = s_N \|\boldsymbol{\lambda}\|_1. \tag{11}$$

Using ℓ_1 duality and letting $\mathbf{z}(w_S) := \mathbf{y} - X_S w_S$, for any w_S the inner minimization problem in (10) becomes

$$\min_{\mathbf{w}_N : X_N \mathbf{w}_N = \mathbf{z}(w_S)} \|\mathbf{w}_N\|_1 = \max_{\|X_N^\top \boldsymbol{\lambda}\|_\infty \le 1} \mathbf{z}(w_S)^\top \boldsymbol{\lambda} = \max_{s_N \|\boldsymbol{\lambda}\|_1 \le 1} \mathbf{z}(w_S)^\top \boldsymbol{\lambda} = \frac{1}{s_N} \|\mathbf{z}(w_S)\|_\infty.$$

Thus, any min-norm interpolator satisfies

$$\hat{w}_S \in \operatorname*{argmin}_{w_S \in \mathbb{R}} \left(|w_S| + \frac{1}{s_N} \|\mathbf{y} - X_S w_S\|_\infty \right). \tag{12}$$

By contrast, if the rows of X_N were i.i.d. copies of a Gaussian vector $\mathbf{x}_N$, then since $X_N^\top \boldsymbol{\lambda} \overset{d}{=} \|\boldsymbol{\lambda}\|_2 \mathbf{x}_N$, (11) would be replaced with $\|X_N^\top \boldsymbol{\lambda}\|_\infty \overset{d}{=} \|\boldsymbol{\lambda}\|_2 \|\mathbf{x}_N\|_\infty$. When $|N|$ is large, there is a deterministic value $\kappa_N > 0$ (which may depend on N) such that with high probability, $\|\boldsymbol{\lambda}\|_2 \|\mathbf{x}_N\|_\infty \approx \|\boldsymbol{\lambda}\|_2 \kappa_N$ for all $\boldsymbol{\lambda}$, and consequently one can prove that:

$$\hat{w}_S \approx \operatorname*{argmin}_{w_S \in \mathbb{R}} \left(|w_S| + \frac{1}{\kappa_N} \|\mathbf{y} - X_S w_S\|_2 \right). \tag{13}$$

The two problems in (12) and (13) have fundamentally different geometries: In the Rademacher case, the solution minimizes the ℓ_∞ (maximal) residual error, whereas in the Gaussian case, the solution (13) minimizes the ℓ_2 (average) error, analogously to our positive results for ℓ_p with $p \in (1, \infty]$ (Sec. 4). Our negative result reflects the intuition that minimizing the ℓ_∞ error is inherently brittle and generally inconsistent. Fig. 2 in the appendix visualizes this geometric contrast directly.

6 General Benign Overfitting Result

In Sec. 4 we presented our framework behind the benign overfitting results for ℓ_p norms. However, our proof technique does not crucially rely on the specific structure of these norms, and in this section, we generalize the results of Sec. 4 to allow for more general norms, input distributions, and target vectors. First, for any $\mathbf{v} \in \mathbb{R}^N$, define the dual of the induced N-block norm[2] by

$$\|\mathbf{v}\|_* := \sup_{\|(\mathbf{0}, \mathbf{u})\| \le 1} \mathbf{v}^\top \mathbf{u}.$$

In the case of ℓ_p norms, $\|\mathbf{v}\|_* = \|\mathbf{v}\|_q$ (where $1/p + 1/q = 1$). As in Sec. 4, we will split the minimum-norm interpolation problem into two: first, consider an arbitrary $\mathbf{w}_S$ and then minimize the norm of $\mathbf{w}_N$ under the constraint that $(\mathbf{w}_S, \mathbf{w}_N)$ interpolates the training data. The majority of our proof concerns the second minimization problem, whose dual form is given by

$$\min_{\mathbf{w}_N : X_N \mathbf{w}_N = \mathbf{y} - X_S \mathbf{w}_S} \|(\mathbf{0}, \mathbf{w}_N)\| = \max_{\boldsymbol{\lambda} \in \mathbb{R}^n : \|X_N^\top \boldsymbol{\lambda}\|_* \le 1} (\mathbf{y} - X_S \mathbf{w}_S)^\top \boldsymbol{\lambda}.$$

We now state two assumptions that capture the properties of $\|X_N^\top \boldsymbol{\lambda}\|_*$ that were needed for our proofs in Sec. 3 (and which are explicitly proven to hold under the conditions of Thm. 1). The first assumption is basically that an approximate CLT holds. Concretely, note that for X_N whose rows are i.i.d. copies of a Gaussian vector $\mathbf{x}_N$, then as explained in (4), writing $\kappa_N := \mathbb{E}\|\mathbf{x}_N\|_*$ gives $\mathbb{E}[\|X_N^\top \boldsymbol{\lambda}\|_*] = \kappa_N \|\boldsymbol{\lambda}\|_2$. For more general distributions, by CLT considerations, it is thus reasonable to assume that $\mathbb{E}[\|X_N^\top \boldsymbol{\lambda}\|_*] \approx \kappa_N \|\boldsymbol{\lambda}\|_2$ for a suitable κ_N and for all $\boldsymbol{\lambda}$ that are flat (in the sense that $\|\boldsymbol{\lambda}\|_\infty \ll \|\boldsymbol{\lambda}\|_2$).

Assumption 3. *There exist a scaling factor $\kappa_N > 0$, a constant $A > 0$, and a nondecreasing function $\varphi : [0, \infty) \to [0, \infty)$ such that for every $\boldsymbol{\lambda} \in \mathbb{R}^n$,*

$$K(\boldsymbol{\lambda}) := \frac{\mathbb{E}\big[\|X_N^\top \boldsymbol{\lambda}\|_*\big]}{\kappa_N} < \infty$$

and for every $\boldsymbol{\lambda} \in \mathbb{S}^{n-1}$:

$$K(\boldsymbol{\lambda}) \ge A \qquad \text{and} \qquad |K(\boldsymbol{\lambda}) - 1| \le \varphi(\|\boldsymbol{\lambda}\|_\infty).$$

[2] While this notation is natural as $\|\mathbf{v}\|_*$ is the dual norm of $\mathbf{v} \mapsto \|(\mathbf{0}, \mathbf{v})\|$, we note that it may in general not be the same as *first* taking the dual norm on $\mathbb{R}^d$ and *then* restricting to $\mathbb{R}^N$.

The lower bound with A is a regularity assumption, and the exact value of A is not crucial for the results. The next assumption we need is a uniform concentration assumption.

Assumption 4. *For every $\delta \in (0,1)$, there is a deterministic $\varepsilon_{\mathrm{unif}}(\delta) \geq 0$ such that, with probability at least $1-\delta$,*

$$\sup_{\boldsymbol{\lambda} \in \mathbb{S}^{n-1}} \left| \frac{\|X_N^\top \boldsymbol{\lambda}\|_*}{\mathbb{E}\|X_N^\top \boldsymbol{\lambda}\|_*} - 1 \right| \leq \varepsilon_{\mathrm{unif}}(\delta).$$

We note that even for Gaussian inputs, uniform concentration over *arbitrary* norms is a nontrivial question which has been extensively studied, e.g. via Dvoretzky's theorem and almost-Euclidean sections of convex bodies (Artstein-Avidan et al., 2015). For many "common" norms, however, including ℓ_p ($p \in (1,\infty]$), uniform convergence does hold, as we show in Proposition 2.

To relate the squared error to the norm $\|\cdot\|$, we define

$$\beta_S := \sup_{\mathbf{v} \in \mathbb{R}^S \setminus \{0\}} \frac{\|(\mathbf{v},\mathbf{0})\|}{\|\mathbf{v}\|_2}, \quad \beta_{N,\Sigma} := \sup_{\mathbf{v} \in \mathbb{R}^N \setminus \{0\}} \frac{(\mathbf{v}^\top \Sigma_N \mathbf{v})^{1/2}}{\|(\mathbf{0},\mathbf{v})\|}, \quad \nu_N := \left((\mathbf{w}_N^\star)^\top \Sigma_N \mathbf{w}_N^\star\right)^{1/2}.$$

We are now ready to state our general results:

Theorem 3 (Risk bound for general norms)**.** *Let $[d] = S \sqcup N$ be any partition of coordinates, and suppose Assumptions 1, 2, 3, and 4 hold. There exist $C, C' \geq 1$ and $c > 0$, depending (polynomially) only on A^{-1}, such that the following holds for every $\delta \in (0,1)$. Suppose that $n \geq C\left(r_S + \log(n/\delta)\right)$ and define $\gamma_N := \kappa_N/\sqrt{n}$ and*

$$\begin{aligned} \rho &:= \frac{\nu_N}{\sigma} + \frac{R_X \|\mathbf{b}_S\|_2 + (R_X b_S^+ + R_\xi \sigma)\sqrt{\log(n/\delta)}}{\sigma\sqrt{n}}, \\ \varepsilon_{\mathrm{flat}} &:= \sqrt{\rho} + \varepsilon_{\mathrm{unif}}(\delta/6) + \varphi(C'\sqrt{\rho}), \end{aligned} \tag{14}$$

and

$$\begin{aligned} r_\delta^2 := \frac{C}{(b_S^-)^2} \Bigg[&\sigma^2 \varepsilon_{\mathrm{flat}} + \gamma_N \|(\mathbf{w}_S^\star, \mathbf{0})\| \left(\sigma + \gamma_N \|(\mathbf{w}_S^\star, \mathbf{0})\|\right) + \sigma^2 \left(\frac{\gamma_N \beta_S}{b_S^-}\right)^2 \\ &+ \left(\frac{R_X R_\xi \sigma b_S^+}{b_S^-}\right)^2 \sqrt{\frac{r_S + \log(n/\delta)}{n}} \Bigg]. \end{aligned} \tag{15}$$

If $r_\delta \leq 1$ and if

$$\varepsilon_{\mathrm{flat}} + \frac{\gamma_N \beta_S}{b_S^-} + \left(\frac{R_X b_S^+ + R_\xi \sigma}{\min\{b_S^-, \sigma\}}\right)^2 \sqrt{\frac{r_S + \log(n/\delta)}{n}} \leq c, \tag{16}$$

then, with probability at least $1-\delta$, every minimum-norm interpolator satisfies the following:

$$\begin{aligned} &\|\widehat{\mathbf{w}}_S - \mathbf{w}_S^\star\|_2 < r_\delta, \\ &\mathcal{R}(\widehat{\mathbf{w}}) \leq (b_S^+)^2 r_\delta^2 + C\beta_{N,\Sigma}^2 \left(\frac{\sigma}{\gamma_N} + \|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S r_\delta\right)^2 + 2\nu_N^2. \end{aligned}$$

Once again, the conditions of the theorem that ensure benign overfitting are best highlighted when making the results asymptotic, as given by the following corollary:

Corollary 2 (Benign overfitting for general norms)**.** *Consider a sequence of problem instances indexed by the sample size n, each satisfying the assumptions of Thm. 3. Write $\gamma_N := \kappa_N/\sqrt{n}$, and suppose that $R_X, R_\xi, \sigma, \sigma^{-1}, (b_S^-)^{-1}, b_S^+, A^{-1}$ and $\|\mathbf{w}_S^\star\|_2$ are uniformly bounded. If there is a sequence $\delta_n \to 0$ such that, with $\varepsilon_{\mathrm{flat}}$ evaluated at $\delta = \delta_n$,*

$$\varepsilon_{\mathrm{flat}} \longrightarrow 0, \qquad \beta_S \gamma_N \longrightarrow 0, \qquad \frac{\beta_{N,\Sigma}}{\gamma_N} \longrightarrow 0,$$

then, with probability tending to one, every minimum-norm interpolator satisfies

$$\mathcal{R}(\widehat{\mathbf{w}}) \longrightarrow 0.$$

For reference, in the setting of Thm. 1, the conditions $\beta_S \gamma_N \to 0$ and $\frac{\beta_{N,\Sigma}}{\gamma_N} \to 0$ are equivalent to:

$$s_N^2 |N|^{2/q} |S|^{\max\{2/p-1,0\}} \ll n \ll |N|^{\min\{2/q,1\}}, \tag{17}$$

where $\varepsilon_{\mathrm{flat}}$ depends on the ε values from Assumptions 3 and 4. In the setting of Thm. 1, the second inequality in (17) and the condition $n \gg r_S$ together imply $\varepsilon_{\mathrm{flat}} \to 0$ and thus benign overfitting.

AI USE STATEMENT

The fundamental proof ideas in the paper were developed by the authors, without the use of AI. In particular, we did not use AI tools to develop the critical proof ideas and the mechanism underlying our results (described in Sec. 4), as well as early versions of the proofs. Nevertheless, we did use AI tools for finding references and known results that were needed for our proofs, sharpening proofs and weakening their assumptions, writing the code for the experiments, and general editing/writing. We have reviewed all AI-assisted work, including everything mentioned in the previous sentence. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shiri Artstein-Avidan, Apostolos Giannopoulos, and Vitali D Milman. *Asymptotic Geometric Analysis, Part I*, volume 202. American Mathematical Society, 2015.

Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024.

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

Daniel Barzilai and Ohad Shamir. Generalization in kernel regression under realistic assumptions. In *Forty-first International Conference on Machine Learning*, 2024.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Sergey G Bobkov. Berry–esseen bounds and edgeworth expansions in the central limit theorem for transport distances. *Probability Theory and Related Fields*, 170(1):229–262, 2018.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020.

Niladri S Chatterji and Philip M Long. Foolish crowds support benign overfitting. *Journal of Machine Learning Research*, 23(125):1–12, 2022.

Tin Sum Cheng, Aurelien Lucchi, Anastasis Kratsios, and David Belius. A comprehensive analysis on the learning curve in kernel ridge regression. *Advances in Neural Information Processing Systems*, 37:24659–24723, 2024.

Geoffrey Chinot, Matthias Löffler, and Sara van de Geer. On the robustness of minimum norm interpolators and regularized empirical risk minimizers. *The Annals of Statistics*, 50(4):2306–2333, 2022.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021.

Michal Derezinski, Feynman T Liang, and Michael Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. *Advances in neural information processing systems*, 33:5152–5164, 2020.

Konstantin Donhauser, Nicolo Ruggeri, Stefan Stojanovic, and Fanny Yang. Fast rates for noisy interpolation require rethinking the effect of inductive bias. In *International Conference on Machine Learning*, pp. 5397–5428. PMLR, 2022.

Eitan Gronich and Gal Vardi. The implicit bias of adam and muon on smooth homogeneous neural networks. In *Forty-third International Conference on Machine Learning*, 2026.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1832–1841. PMLR, 2018.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL `https://kellerjordan.github.io/posts/muon/`.

Peizhong Ju, Xiaojun Lin, and Jia Liu. Overfitting can be harmless for basis pursuit, but only to a degree. *Advances in Neural Information Processing Systems*, 33:7956–7967, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020.

Frederic Koehler, Lijia Zhou, Danica J. Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20657–20668, 2021.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference: A Journal of the IMA*, 11(4):1389–1456, 2022.

Gil Kur and Pierre Bizeul. Minimum norm interpolation via the local theory of banach spaces: The role of 2-uniform convexity. *arXiv preprint arXiv:2603.28956*, 2026.

Gil Kur and Reese Pathak. Minimum norm interpolation via the local theory of banach spaces: The role of gaussianity. *arXiv preprint arXiv:2607.07694*, 2026.

Gil Kur, Pedro Abdalla, Pierre Bizeul, and Fanny Yang. Minimum norm interpolation meets the local theory of banach spaces. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25726–25754. PMLR, 2024.

Tengyuan Liang and Alexander Rakhlin. Just interpolate. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Weihao Lu, Haobo Zhang, Yicheng Li, Manyun Xu, and Qian Lin. Optimal rate of kernel regression in large dimensions. *arXiv preprint arXiv:2309.04268*, 2023.

Neil Mallinar, James Simon, Amirhesam Abedsoltan, Parthe Pandit, Misha Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting. *Advances in neural information processing systems*, 35:1182–1195, 2022.

Neil Rohit Mallinar, Austin Zane, Spencer Frei, and Bin Yu. Minimum-norm interpolation under covariate shift. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 21–27 Jul 2024.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

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

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

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

Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, pp. 3889–3897. PMLR, 2021.

Ohad Shamir. The implicit bias of benign overfitting. *Journal of Machine Learning Research*, 24 (113):1–40, 2023.

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

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *J. Mach. Learn. Res.*, 24:123–1, 2023.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2 edition, 2026.

Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum ℓ_1-norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics*, pp. 10572–10602. PMLR, 2022.

Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue M Lu, and Jeffrey Pennington. Precise learning curves and higher-order scaling limits for dot-product kernel regression. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(11):114005, 2023.

Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data. *Advances in Neural Information Processing Systems*, 37:23988–24021, 2024.

Nikita Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*, 29:1–28, 2024.

CONTENTS

1 Introduction **1**

2 Setting and Preliminaries **3**

3 Benign Overfitting for ℓ_p Norms ($p \in (1, \infty]$) **4**

4 Proof Ideas **5**

4.1 Warm-Up: Gaussian Inputs 5

4.2 Beyond Gaussians 6

5 A Failure Example for Minimum-ℓ_1-Norm Interpolation **7**

6 General Benign Overfitting Result **8**

A Related Work **15**

B Auxiliary Lemmas **16**

C Proof of Theorem 3 and Related Lemmas **18**

C.1 Approximate Isometry on Flat Vectors 18

C.2 Uniform Approximation of the Dual Objective 20

C.3 Proof of Theorem 3 23

C.4 Proof of Corollary 2 25

D Proof of Theorem 1 and Related Results **26**

D.1 Setting 26

D.2 Moments and Expected Dual-Norm Geometry 27

D.3 Concentration of the Empirical Dual Norm 27

D.4 Gaussian Approximation via a Quantitative CLT 31

D.4.1 Expected ℓ_q geometry on flat directions 32

D.5 Proof of Theorem 1 33

E Proof of Theorem 2 – The ℓ_1 Failure Example **35**

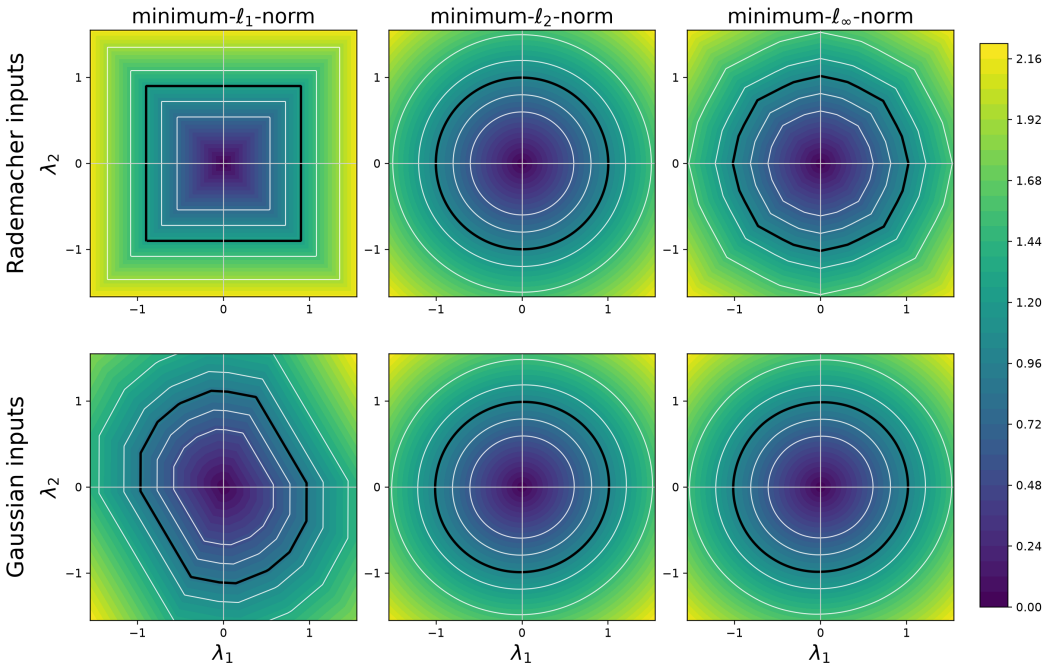

Figure 2: Dual geometry for Rademacher (top) and Gaussian (bottom) inputs with $n = 12$ and $|N| = 2^n$. Each panel shows normalized level sets of $\boldsymbol{\lambda} \mapsto \|X_N^\top \boldsymbol{\lambda}\|_q$ on a random flat two-dimensional subspace $\boldsymbol{\lambda} = a\mathbf{u} + b\mathbf{v}$, where $\mathbf{u}, \mathbf{v} \in \{\pm 1/\sqrt{n}\}^n$ are orthonormal and q is dual to the indicated minimum-ℓ_p norm. The black curve is the unit level set. The geometries differ substantially: For Gaussian inputs, the geometry is nearly Euclidean across all three norms. However, for Rademacher inputs, the geometry varies substantially between the norms.

A Related Work

Minimum-ℓ_2-norm interpolation. There is by now a very large body of work studying the generalization error of minimum-ℓ_2-norm interpolation in linear regression (Kobak et al., 2020; Bartlett et al., 2020; Muthukumar et al., 2020; Belkin et al., 2020; Derezinski et al., 2020; Negrea et al., 2020; Richards et al., 2021; Muthukumar et al., 2021; Hastie et al., 2022; Bach, 2024; Shamir, 2023; Tsigler & Bartlett, 2023; Mallinar et al., 2024), as well as random-feature/kernel regression (Liang & Rakhlin, 2020; Bordelon et al., 2020; Cui et al., 2021; Xiao et al., 2023; Mei et al., 2022; Lu et al., 2023; Cheng et al., 2024; Barzilai & Shamir, 2024). In many cases, sharp excess risk bounds are possible when analyzing the minimum-ℓ_2-norm interpolator under relatively mild assumptions. However, the overwhelming majority of these analyses *begin* with the known closed-form formula for the minimum ℓ_2-norm interpolator, namely $\hat{\mathbf{w}} = X^\top (XX^\top)^\dagger \mathbf{y}$, which is highly specific to the ℓ_2 case. As such, many of the techniques and tools developed in such works do not extend to the analysis of minimum-norm interpolation under more general norms.

General norms. Much less is known about minimum-norm interpolation for non-ℓ_2 norms, and existing works largely rely on Gaussian inputs (and often Gaussian noise as well) (Ju et al., 2020; Koehler et al., 2021; Chatterji & Long, 2022; Donhauser et al., 2022; Wang et al., 2022; Kur & Pathak, 2026). In most of these works, Gaussianity is crucially utilized by relying on the Convex Gaussian Minmax theorem (Thrampoulidis et al., 2015), which does not generalize to other distributions. Under such assumptions, perhaps the best studied norm is ℓ_1 (also known as basis pursuit), where for isotropic Gaussian inputs with variance σ^2, sharp rates scaling as $\sigma^2/\log(d/n)$ are known. While this converges to 0 when $d/n \to \infty$, we show that Gaussianity is critical for this bound, and provide an example of excess risk lower bounded as $\sigma^2 \log(n)$ for sub-Gaussian inputs and noise. Within these works, perhaps the closest to our setting is Koehler et al. (2021), who analyze general norms and split the inputs into a low-dimensional block for fitting the target and a high-dimensional

block for fitting the noise, as we do here. Nevertheless, the Gaussian Minmax theorem is central to their results, preventing them from analyzing more general distributions.

Kur et al. (2024); Kur & Bizeul (2026) provide sharp excess risk bounds for ℓ_p norms with $p \in (1, 2]$ that only assume sub-Gaussian inputs. Nevertheless, they assume Gaussian noise and isotropic inputs, and their risk bounds do not apply to more general norms such as ℓ_p with $p > 2$. Chinot et al. (2022) analyze more general input/noise distributions, allowing even for adversarial noise. In the case where the label noise is independent with variance σ^2, they establish risk bounds of the form $\mathcal{O}\left(\sigma^2\right)$, but these do not imply benign overfitting, meaning that the excess risk does not tend to 0 when $\sigma^2 > 0$.

B Auxiliary Lemmas

In this section we collect standard concentration results for later use. Recall that, for nonempty $T \subseteq [d]$,

$$\Sigma_T := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{x}_T\mathbf{x}_T^\top] = \operatorname{diag}\big((b_j^2)_{j\in T}\big).$$

The last identity follows from Assumption 1: the coordinates are independent and mean-zero, with variances b_j^2. For a nonempty coordinate set $T \subseteq [d]$ and $\delta \in (0, 1)$, define the quadratic concentration scale

$$\Delta_2(T, \delta) := \sqrt{\frac{r_T + \log(8/\delta)}{n}}. \tag{18}$$

For the empty set, define instead the scalar Bernstein scale

$$\Delta_2(\emptyset, \delta) := \left[\sqrt{\frac{\log(8/\delta)}{n}} + \frac{\log(8/\delta)}{n}\right].$$

For nonempty T, also define the maximum-coordinate scale

$$\Delta_\infty(T, \delta) := \frac{1}{\sqrt{n}}\left[R_X\|\mathbf{b}_T\|_2 + \left(R_X b_T^+ + R_\xi\sigma\right)\sqrt{\log\frac{8n}{\delta}}\right]. \tag{19}$$

Lemma 1. *Let $Z = (Z_1, \dots, Z_m)$ have independent, mean-zero, sub-Gaussian coordinates, and write $K_j := \|Z_j\|_{\psi_2}$. There exists an absolute constant $C > 0$, such that for every $t > 0$, with probability at least $1 - 2e^{-t}$,*

$$\left|\|Z\|_2^2 - \sum_{j=1}^m \mathbb{E}Z_j^2\right| \le C\left(\sqrt{t\sum_{j=1}^m K_j^4} + t\max_{j\in[m]} K_j^2\right), \tag{20}$$

and

$$\|Z\|_2 \le C\left[\left(\sum_{j=1}^m K_j^2\right)^{1/2} + \max_{j\in[m]} K_j\sqrt{t}\right]. \tag{21}$$

Proof. Throughout this proof, we use C to denote an absolute constant that may change from line to line. The variables $Z_j^2 - \mathbb{E}Z_j^2$ are independent, mean-zero, and, by (Vershynin, 2026, Lemmas 2.7.8, 2.8.5), have ψ_1-norms at most CK_j^2. Applying Bernstein's inequality (Vershynin, 2026, Theorem 2.9.1) gives (20). Moreover,

$$\sum_{j=1}^m \mathbb{E}Z_j^2 \le C\sum_{j=1}^m K_j^2, \qquad \left(\sum_{j=1}^m K_j^4\right)^{1/2} \le \max_{j\in[m]} K_j\left(\sum_{j=1}^m K_j^2\right)^{1/2}.$$

Thus, squaring the right-hand side of (21) shows it dominates the upper bound for $\|Z\|_2^2$ given by (20). □

Lemma 2. *Suppose Assumptions 1 and 2 hold. Then there exists an absolute constant $C > 0$ such that for every nonempty $T \subseteq [d]$, every $\delta \in (0,1)$, and every $n \geq C\,(r_T + \log(8/\delta))$, with probability at least $1-\delta$, the following bounds hold simultaneously:*

$$\left\| \frac{1}{n} X_T^\top X_T - \Sigma_T \right\|_{\mathrm{op}} \leq C R_X^2 (b_T^+)^2 \Delta_2(T,\delta),$$

$$\frac{1}{\sqrt{n}} \left(\max_{i\in[n]} \|\mathbf{x}_{i,T}\|_2 + \|\boldsymbol{\xi}\|_\infty \right) \leq C \Delta_\infty(T,\delta),$$

$$\left\| \frac{1}{n} X_T^\top \boldsymbol{\xi} \right\|_2 \leq C R_X R_\xi \sigma b_T^+ \Delta_2(T,\delta),$$

$$\left| \frac{1}{n} \|\boldsymbol{\xi}\|_2^2 - \sigma^2 \right| \leq C R_\xi^2 \sigma^2 \Delta_2(\emptyset,\delta).$$

Moreover,

$$\frac{1}{\sqrt{n}} \|X_N \mathbf{w}_N^\star\|_2 \leq \left(1 + C R_X^2 \Delta_2(\emptyset,\delta)\right)^{1/2} \nu_N.$$

Proof. Throughout this proof, we use C to denote an absolute constant that may change from line to line. For every $\mathbf{u} \in \mathbb{R}^T$, independence and centering of the coordinates, together with the sub-Gaussian linear-combination inequality (Vershynin, 2026, Proposition 2.7.1), give

$$\|\langle \mathbf{u}, \mathbf{x}_T \rangle\|_{\psi_2} \leq C \left(\sum_{j\in T} u_j^2 \|x_j\|_{\psi_2}^2 \right)^{1/2} \leq C R_X \left(\mathbf{u}^\top \Sigma_T \mathbf{u} \right)^{1/2}.$$

Thus condition (4) of Example 1 in Zhivotovskiy (2024) holds with $\kappa = C R_X$, and hence condition (2) of Theorem 1 holds for $M = \mathbf{x}_T \mathbf{x}_T^\top$. Applying that theorem with $t = \log(8/\delta)$, using

$$\frac{\mathrm{tr}(\Sigma_T)}{\|\Sigma_T\|_{\mathrm{op}}} = r_T, \qquad \|\Sigma_T\|_{\mathrm{op}} = (b_T^+)^2,$$

and the assumed sample-size condition, gives, with probability at least $1 - \delta/8$,

$$\left\| \frac{1}{n} X_T^\top X_T - \Sigma_T \right\|_{\mathrm{op}} \leq C R_X^2 (b_T^+)^2 \sqrt{\frac{r_T + \log(8/\delta)}{n}}.$$

For the row norm, apply Lemma 1 with $t = \log(16n/\delta)$ to each $\mathbf{x}_{i,T}$, using $\|x_{i,j}\|_{\psi_2} \leq R_X b_j$, and take a union bound over $i \in [n]$. This gives, with probability at least $1 - \delta/8$, after adjusting the absolute constant,

$$\frac{1}{\sqrt{n}} \max_{i\in[n]} \|\mathbf{x}_{i,T}\|_2 \leq \frac{C R_X}{\sqrt{n}} \left(\|\mathbf{b}_T\|_2 + b_T^+ \sqrt{\log(8n/\delta)} \right).$$

For the cross term, condition on $\boldsymbol{\xi}$, and write

$$Y_j := \frac{1}{n} \sum_{i=1}^n x_{i,j} \xi_i, \qquad j \in T.$$

The variables $(Y_j)_{j\in T}$ are then independent and mean-zero, and the sub-Gaussian linear-combination inequality gives

$$\|Y_j\|_{\psi_2} \leq \frac{C R_X b_j}{n} \|\boldsymbol{\xi}\|_2.$$

Conditionally applying Lemma 1 with $t = \log(16/\delta)$ therefore gives, with conditional probability at least $1 - \delta/8$,

$$\left\| \frac{1}{n} X_T^\top \boldsymbol{\xi} \right\|_2 \leq \frac{C R_X \|\boldsymbol{\xi}\|_2}{n} \left(\|\mathbf{b}_T\|_2 + b_T^+ \sqrt{\log(8/\delta)} \right)$$

after adjusting the absolute constant. The same lemma, with the same value of t, applied to $\boldsymbol{\xi}$ gives

$$\|\boldsymbol{\xi}\|_2 \le CR_\xi\sigma\left(\sqrt{n}+\sqrt{\log(8/\delta)}\right)$$

with probability at least $1-\delta/8$. Under the assumed sample-size condition, $\log(8/\delta)\le n$, and the last two displays, together with the exact identity

$$\|\mathbf{b}_T\|_2 = b_T^+\sqrt{r_T},$$

give

$$\left\|\frac{1}{n}X_T^\top\boldsymbol{\xi}\right\|_2 \le CR_XR_\xi\sigma b_T^+\sqrt{\frac{r_T+\log(8/\delta)}{n}}.$$

The ℓ_∞ noise bound follows from the standard sub-Gaussian tail bound, e.g. (Vershynin, 2026, Proposition 2.6.1), and a union bound over $i\in[n]$. Thus, with probability at least $1-\delta/8$,

$$\frac{1}{\sqrt{n}}\|\boldsymbol{\xi}\|_\infty \le CR_\xi\sigma\sqrt{\frac{\log(8n/\delta)}{n}}.$$

Adding this to the row-norm bound proves the stated $C\Delta_\infty(T,\delta)$ bound.

For the noise energy, apply (20) with $t=\log(16/\delta)$ to $\boldsymbol{\xi}$, using $\mathbb{E}\xi_i^2=\sigma^2$ and $\|\xi_i\|_{\psi_2}\le R_\xi\sigma$. With probability at least $1-\delta/8$, this gives

$$\left|\frac{1}{n}\|\boldsymbol{\xi}\|_2^2-\sigma^2\right| \le CR_\xi^2\sigma^2\Delta_2(\emptyset,\delta).$$

Finally, set $V_i := \mathbf{x}_{i,N}^\top\mathbf{w}_N^\star$. These variables are independent and mean-zero, with

$$\mathbb{E}V_i^2=\nu_N^2, \qquad \|V_i\|_{\psi_2}\le CR_X\nu_N.$$

Another application of (20), again with $t=\log(16/\delta)$, gives, with probability at least $1-\delta/8$,

$$\left|\frac{1}{n}\|X_N\mathbf{w}_N^\star\|_2^2-\nu_N^2\right| \le CR_X^2\Delta_2(\emptyset,\delta)\nu_N^2.$$

Taking the upper bound and then the square root proves the last conclusion. Taking a union bound over the above events completes the proof. □

C Proof of Theorem 3 and Related Lemmas

C.1 Approximate Isometry on Flat Vectors

Let

$$K(\boldsymbol{\lambda}) := \frac{\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_*}{\kappa_N}.$$

As a reminder, the norm $\|\cdot\|_*$ in the definition of $K(\boldsymbol{\lambda})$ is defined as the dual norm of the induced N-block norm, i.e. $\|\mathbf{v}\|_* = \sup_{\|(\mathbf{0},\mathbf{u})\|\le 1}\mathbf{v}^\top\mathbf{u}$.

Lemma 3 (Dual truncation)**.** *Suppose Assumptions 1 and 3 hold. For $\tau>0$, define the truncated vector*

$$\boldsymbol{\lambda}_\tau := \sum_{j=1}^n \mathbf{1}_{\{|\lambda_j|\le\tau\}}\lambda_j\mathbf{e}_j.$$

Then every $\boldsymbol{\lambda}$ with $K(\boldsymbol{\lambda})\le 1$ satisfies for every $\mathbf{z}\in\mathbb{R}^n$:

$$K(\boldsymbol{\lambda}_\tau)\le K(\boldsymbol{\lambda}), \qquad \mathbf{z}^\top\boldsymbol{\lambda}_\tau \ge \mathbf{z}^\top\boldsymbol{\lambda}-\frac{\|\mathbf{z}\|_\infty}{A^2\tau}.$$

Proof. If $\boldsymbol{\lambda} = 0$, the result is trivial. Otherwise, the assumed lower bound and homogeneity imply

$$K(\boldsymbol{\lambda}) = \|\boldsymbol{\lambda}\|_2 K\left(\frac{\boldsymbol{\lambda}}{\|\boldsymbol{\lambda}\|_2}\right) \geq A\|\boldsymbol{\lambda}\|_2,$$

so every feasible $\boldsymbol{\lambda}$ satisfies

$$\|\boldsymbol{\lambda}\|_2 \leq A^{-1}. \tag{22}$$

Fix a coordinate i, and let $\boldsymbol{\lambda}' = \boldsymbol{\lambda} - \lambda_i \mathbf{e}_i$. Since $\mathbf{x}_{i,N}$ is independent of the other rows and has mean zero, Jensen's inequality gives

$$\kappa_N K(\boldsymbol{\lambda}) = \mathbb{E}\left\|\sum_{j=1}^{n} \lambda_j \mathbf{x}_{j,N}\right\|_* \geq \mathbb{E}\left\|\sum_{j\neq i} \lambda_j \mathbf{x}_{j,N} + \lambda_i \mathbb{E}[\mathbf{x}_{i,N}]\right\|_* = \kappa_N K(\boldsymbol{\lambda}').$$

Sequentially applying this zeroing operation over all coordinates with $|\lambda_j| > \tau$ gives $K(\boldsymbol{\lambda}_\tau) \leq K(\boldsymbol{\lambda})$. Let $\mathcal{I}_\tau := \{j : |\lambda_j| > \tau\}$. Since

$$\boldsymbol{\lambda} - \boldsymbol{\lambda}_\tau = \sum_{j\in\mathcal{I}_\tau} \lambda_j \mathbf{e}_j,$$

equation (22) gives

$$\sum_{j\in\mathcal{I}_\tau} |\lambda_j| \leq \frac{1}{\tau}\sum_{j\in\mathcal{I}_\tau} \lambda_j^2 \leq \frac{\|\boldsymbol{\lambda}\|_2^2}{\tau} \leq \frac{1}{A^2\tau}.$$

Therefore,

$$\mathbf{z}^\top \boldsymbol{\lambda}_\tau = \mathbf{z}^\top \boldsymbol{\lambda} - \sum_{j\in\mathcal{I}_\tau} z_j \lambda_j \geq \mathbf{z}^\top \boldsymbol{\lambda} - \|\mathbf{z}\|_\infty \sum_{j\in\mathcal{I}_\tau} |\lambda_j| \geq \mathbf{z}^\top \boldsymbol{\lambda} - \frac{\|\mathbf{z}\|_\infty}{A^2\tau}.$$

□

Proposition 1 (Flat-vector dual approximation)**.** *Suppose Assumptions 1 and 3 hold. Let* $\mathbf{z} \in \mathbb{R}^n$ *be nonzero, and write* $\varepsilon_{\text{geom}} := \varphi\left(2\sqrt{\frac{\|\mathbf{z}\|_\infty}{\|\mathbf{z}\|_2}}\right)$*. If* $\varepsilon_{\text{geom}} \leq 1/2$*, then*

$$(1-\varepsilon_{\text{geom}})\|\mathbf{z}\|_2 \leq \max_{\boldsymbol{\lambda}: K(\boldsymbol{\lambda})\leq 1} \mathbf{z}^\top \boldsymbol{\lambda} \leq \left(1 + 2\varepsilon_{\text{geom}} + A^{-2}\sqrt{\frac{\|\mathbf{z}\|_\infty}{\|\mathbf{z}\|_2}}\right)\|\mathbf{z}\|_2.$$

Proof. First, note that the feasible set $\{\boldsymbol{\lambda} : K(\boldsymbol{\lambda}) \leq 1\}$ is compact: it is closed, and Assumption 3 gives $K(\boldsymbol{\lambda}) \geq A\|\boldsymbol{\lambda}\|_2$ implying it is bounded. Thus the maximum is attained. Let

$$\bar{\mathbf{z}} := \frac{\mathbf{z}}{\|\mathbf{z}\|_2}, \qquad \tau := \sqrt{\|\bar{\mathbf{z}}\|_\infty} = \sqrt{\frac{\|\mathbf{z}\|_\infty}{\|\mathbf{z}\|_2}}.$$

Since $\varphi(\cdot)$ is nondecreasing and $\|\bar{\mathbf{z}}\|_\infty \leq 2\tau$, Assumption 3 gives

$$K(\bar{\mathbf{z}}) \leq 1 + \varphi(\|\bar{\mathbf{z}}\|_\infty) \leq 1 + \varepsilon_{\text{geom}}.$$

Hence $\bar{\mathbf{z}}/(1+\varepsilon_{\text{geom}})$ is feasible, and therefore

$$\max_{\boldsymbol{\lambda}: K(\boldsymbol{\lambda})\leq 1} \mathbf{z}^\top \boldsymbol{\lambda} \geq \frac{\|\mathbf{z}\|_2}{1+\varepsilon_{\text{geom}}} \geq (1-\varepsilon_{\text{geom}})\|\mathbf{z}\|_2.$$

For the upper bound, let $\boldsymbol{\lambda}$ be any feasible point. By Lemma 3,

$$\mathbf{z}^\top \boldsymbol{\lambda} \leq \mathbf{z}^\top \boldsymbol{\lambda}_\tau + \frac{\|\mathbf{z}\|_\infty}{A^2\tau} \leq \|\mathbf{z}\|_2\|\boldsymbol{\lambda}_\tau\|_2 + \frac{\|\mathbf{z}\|_\infty}{A^2\tau}.$$

It remains to bound $\|\boldsymbol{\lambda}_\tau\|_2$. If $\|\boldsymbol{\lambda}_\tau\|_2 < 1/2$, then

$$\|\boldsymbol{\lambda}_\tau\|_2 \leq \frac{1}{1-\varepsilon_{\text{geom}}}.$$

Otherwise, $\|\boldsymbol{\lambda}_\tau\|_2 \geq 1/2$, and since every coordinate of $\boldsymbol{\lambda}_\tau$ has magnitude at most τ,

$$\frac{\|\boldsymbol{\lambda}_\tau\|_\infty}{\|\boldsymbol{\lambda}_\tau\|_2} \leq 2\tau.$$

Using $K(\boldsymbol{\lambda}_\tau) \leq 1$, homogeneity, and Assumption 3,

$$1 \geq K(\boldsymbol{\lambda}_\tau) = \|\boldsymbol{\lambda}_\tau\|_2 K\left(\frac{\boldsymbol{\lambda}_\tau}{\|\boldsymbol{\lambda}_\tau\|_2}\right) \geq \|\boldsymbol{\lambda}_\tau\|_2(1-\varepsilon_{\mathrm{geom}}).$$

Thus, in all cases,

$$\|\boldsymbol{\lambda}_\tau\|_2 \leq \frac{1}{1-\varepsilon_{\mathrm{geom}}}.$$

Consequently,

$$\mathbf{z}^\top \boldsymbol{\lambda} \leq \|\mathbf{z}\|_2 \left(\frac{1}{1-\varepsilon_{\mathrm{geom}}} + \frac{\|\mathbf{z}\|_\infty}{A^2\tau\|\mathbf{z}\|_2}\right).$$

Since $\tau = \sqrt{\|\mathbf{z}\|_\infty/\|\mathbf{z}\|_2}$ and $(1-\varepsilon_{\mathrm{geom}})^{-1} \leq 1 + 2\varepsilon_{\mathrm{geom}}$, this becomes

$$\mathbf{z}^\top \boldsymbol{\lambda} \leq \left(1 + 2\varepsilon_{\mathrm{geom}} + A^{-2}\sqrt{\frac{\|\mathbf{z}\|_\infty}{\|\mathbf{z}\|_2}}\right)\|\mathbf{z}\|_2.$$

Taking the maximum over feasible $\boldsymbol{\lambda}$ completes the proof. □

Corollary 3 (Random flat-vector dual approximation)**.** *Suppose Assumptions 1, 3, and 4 hold. Fix $\delta \in (0,1)$ such that $\varepsilon_{\mathrm{unif}} := \varepsilon_{\mathrm{unif}}(\delta) < 1/2$. Then, with probability at least $1-\delta$, the following holds for every nonzero $\mathbf{z} \in \mathbb{R}^n$. Define $\varepsilon_{\mathrm{geom}} := \varphi\left(2\sqrt{\frac{\|\mathbf{z}\|_\infty}{\|\mathbf{z}\|_2}}\right)$, and suppose that $\varepsilon_{\mathrm{geom}} \leq 1/2$. Then*

$$\begin{aligned}\frac{(1-\varepsilon_{\mathrm{unif}})(1-\varepsilon_{\mathrm{geom}})}{\kappa_N}\|\mathbf{z}\|_2 &\leq \max_{\boldsymbol{\lambda}:\|X_N^\top\boldsymbol{\lambda}\|_* \leq 1} \mathbf{z}^\top\boldsymbol{\lambda} \\ &\leq \frac{(1+2\varepsilon_{\mathrm{unif}})\left(1+2\varepsilon_{\mathrm{geom}} + A^{-2}\sqrt{\frac{\|\mathbf{z}\|_\infty}{\|\mathbf{z}\|_2}}\right)}{\kappa_N}\|\mathbf{z}\|_2.\end{aligned}$$

Proof. By Assumption 4, with probability at least $1-\delta$, the defining inequality in Assumption 4 holds. For $\boldsymbol{\lambda} = 0$ the next inequality is trivial, while for $\boldsymbol{\lambda} \neq 0$ it follows by applying Assumption 4 to $\boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|_2$. Thus, for every $\boldsymbol{\lambda} \in \mathbb{R}^n$,

$$(1-\varepsilon_{\mathrm{unif}})\kappa_N K(\boldsymbol{\lambda}) \leq \|X_N^\top\boldsymbol{\lambda}\|_* \leq (1+\varepsilon_{\mathrm{unif}})\kappa_N K(\boldsymbol{\lambda}).$$

Therefore,

$$\left\{\boldsymbol{\lambda} : K(\boldsymbol{\lambda}) \leq \frac{1}{(1+\varepsilon_{\mathrm{unif}})\kappa_N}\right\} \subseteq \left\{\boldsymbol{\lambda} : \|X_N^\top\boldsymbol{\lambda}\|_* \leq 1\right\} \subseteq \left\{\boldsymbol{\lambda} : K(\boldsymbol{\lambda}) \leq \frac{1}{(1-\varepsilon_{\mathrm{unif}})\kappa_N}\right\}.$$

By homogeneity of K,

$$\frac{1}{(1+\varepsilon_{\mathrm{unif}})\kappa_N}\max_{\boldsymbol{\lambda}:K(\boldsymbol{\lambda})\leq 1}\mathbf{z}^\top\boldsymbol{\lambda} \leq \max_{\boldsymbol{\lambda}:\|X_N^\top\boldsymbol{\lambda}\|_*\leq 1}\mathbf{z}^\top\boldsymbol{\lambda} \leq \frac{1}{(1-\varepsilon_{\mathrm{unif}})\kappa_N}\max_{\boldsymbol{\lambda}:K(\boldsymbol{\lambda})\leq 1}\mathbf{z}^\top\boldsymbol{\lambda}.$$

Apply Proposition 1. For the lower bound, use $(1+\varepsilon_{\mathrm{unif}})^{-1} \geq 1-\varepsilon_{\mathrm{unif}}$, and for the upper bound use $(1-\varepsilon_{\mathrm{unif}})^{-1} \leq 1+2\varepsilon_{\mathrm{unif}}$, valid since $\varepsilon_{\mathrm{unif}} < 1/2$. □

C.2 Uniform Approximation of the Dual Objective

For $\mathbf{w}_S \in \mathbb{R}^S$, define

$$\mathbf{z}(\mathbf{w}_S) := X_S(\mathbf{w}_S^\star - \mathbf{w}_S) + \boldsymbol{\xi},$$

and

$$\mathcal{W}_S := \left\{\mathbf{w}_S \in \mathbb{R}^S : \|\mathbf{w}_S - \mathbf{w}_S^\star\|_2 \leq 1\right\}.$$

Recall that

$$\nu_N^2 = \sum_{j \in N} b_j^2 (w_j^\star)^2.$$

Thus ν_N^2 is the population risk contribution of the true N-component.

On the event where the constraint is feasible, define

$$g(\mathbf{w}_S) := \min_{\mathbf{w}_N \colon X_N \mathbf{w}_N = \mathbf{z}(\mathbf{w}_S) + X_N \mathbf{w}_N^\star} \|(\mathbf{0}, \mathbf{w}_N)\|.$$

Lemma 4 (Dual representation)**.** *Fix a matrix X_N and $\mathbf{r} \in \mathbb{R}^n$, and suppose the constraint $X_N \mathbf{w}_N = \mathbf{r}$ is feasible. Then*

$$\min_{\mathbf{w}_N \colon X_N \mathbf{w}_N = \mathbf{r}} \|(\mathbf{0}, \mathbf{w}_N)\| = \max_{\boldsymbol{\lambda} \in \mathbb{R}^n} \left\{ \mathbf{r}^\top \boldsymbol{\lambda} : \|X_N^\top \boldsymbol{\lambda}\|_* \le 1 \right\},$$

where $\|\cdot\|_*$ is the dual norm associated with $\mathbf{w}_N \mapsto \|(\mathbf{0}, \mathbf{w}_N)\|$.

Proof. The Lagrangian is

$$L(\mathbf{w}_N, \boldsymbol{\lambda}) = \|(\mathbf{0}, \mathbf{w}_N)\| + \boldsymbol{\lambda}^\top (\mathbf{r} - X_N \mathbf{w}_N).$$

Hence the dual function is

$$\inf_{\mathbf{w}_N} L(\mathbf{w}_N, \boldsymbol{\lambda}) = \mathbf{r}^\top \boldsymbol{\lambda} + \inf_{\mathbf{w}_N} \left\{ \|(\mathbf{0}, \mathbf{w}_N)\| - (X_N^\top \boldsymbol{\lambda})^\top \mathbf{w}_N \right\}.$$

By the definition of the dual norm, the infimum in braces is 0 when $\|X_N^\top \boldsymbol{\lambda}\|_* \le 1$, attained at $\mathbf{w}_N = \mathbf{0}$, and is $-\infty$ otherwise. Thus the Lagrange dual is

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^n} \left\{ \mathbf{r}^\top \boldsymbol{\lambda} : \|X_N^\top \boldsymbol{\lambda}\|_* \le 1 \right\}.$$

Finally, the primal objective is finite and continuous everywhere, and the affine constraint is feasible by assumption. Therefore, Slater's condition holds, so strong duality holds and the dual optimum is attained. □

Fix an absolute constant $C_{\mathrm{conc}} > 0$ large enough to dominate the constant in both the sample-size condition and every conclusion of Lemma 2. For $\delta \in (0, 1)$, define

$$\begin{aligned}
\eta_N(\delta) &:= \left(1 + C_{\mathrm{conc}} R_X^2 \Delta_2(\emptyset, \delta/3)\right)^{1/2} \nu_N, \\
L_S(\delta) &:= \left(\sigma^2 - C_{\mathrm{conc}} R_\xi^2 \sigma^2 \Delta_2(\emptyset, \delta/3) - 2 C_{\mathrm{conc}} R_X R_\xi \sigma b_S^+ \Delta_2(S, \delta/3)\right)^{1/2}, \\
U_S(\delta) &:= C_{\mathrm{conc}} \Delta_\infty(S, \delta/3), \\
\rho_{S,N}(\delta) &:= \frac{U_S(\delta) + \eta_N(\delta)}{L_S(\delta) - \eta_N(\delta)}.
\end{aligned} \tag{23}$$

These quantities are used only when the right-hand sides are well-defined and $L_S(\delta) > \eta_N(\delta)$.

Lemma 5 (Uniform approximation of the N-block fitting norm)**.** *Suppose Assumptions 1, 2, 3, and 4 hold. Fix $\delta \in (0, 1)$, and let $\varepsilon_{\mathrm{unif}} := \varepsilon_{\mathrm{unif}}(\delta/3)$ be the quantity from Assumption 4. Let $\varphi(\cdot)$ and A be the quantities from Assumption 3, and define*

$$\varepsilon_{\mathrm{fit}}(\delta) := C \left(\rho_{S,N}(\delta) + \varepsilon_{\mathrm{unif}} + \varphi\left(2\sqrt{\rho_{S,N}(\delta)}\right) + \sqrt{\rho_{S,N}(\delta)} \right), \tag{24}$$

where $C = C(A) \ge 1$ is a sufficiently large fixed constant, chosen in particular so that the sample-size condition below implies the hypothesis of Lemma 2 with failure probability $\delta/3$. Suppose that

$$n \ge C\left(r_S + \log(8/\delta)\right), \qquad L_S(\delta) > \eta_N(\delta), \qquad \varepsilon_{\mathrm{unif}} < \frac{1}{2}, \qquad \varphi\left(2\sqrt{\rho_{S,N}(\delta)}\right) \le \frac{1}{2}.$$

Then, with probability at least $1 - \delta$, X_N has full row rank and, simultaneously for every $\mathbf{w}_S \in \mathcal{W}_S$,

$$(1 - \varepsilon_{\mathrm{fit}}(\delta)) \|\mathbf{z}(\mathbf{w}_S)\|_2 \le \kappa_N g(\mathbf{w}_S) \le (1 + \varepsilon_{\mathrm{fit}}(\delta)) \|\mathbf{z}(\mathbf{w}_S)\|_2.$$

Proof. First, by Assumption 4, with probability at least $1-\delta/3$,

$$(1-\varepsilon_{\mathrm{unif}})\kappa_N K(\boldsymbol{\lambda}) \le \|X_N^\top \boldsymbol{\lambda}\|_* \le (1+\varepsilon_{\mathrm{unif}})\kappa_N K(\boldsymbol{\lambda}) \qquad \text{for all } \boldsymbol{\lambda}\in\mathbb{R}^n. \tag{25}$$

Second, by Lemma 2, applied with $T=S$ and failure probability $\delta/3$, we have

$$\frac{1}{\sqrt{n}}\left(\max_{i\in[n]}\|\mathbf{x}_{i,S}\|_2 + \|\boldsymbol{\xi}\|_\infty\right) \le C_{\mathrm{conc}}\Delta_\infty(S,\delta/3) = U_S(\delta), \tag{26}$$

and

$$\left\|\frac{1}{n}X_S^\top\boldsymbol{\xi}\right\|_2 \le C_{\mathrm{conc}}R_X R_\xi \sigma b_S^+ \Delta_2(S,\delta/3), \qquad \left|\frac{1}{n}\|\boldsymbol{\xi}\|_2^2 - \sigma^2\right| \le C_{\mathrm{conc}}R_\xi^2\sigma^2\Delta_2(\emptyset,\delta/3). \tag{27}$$

The same application of Lemma 2 also gives

$$\frac{1}{\sqrt{n}}\|X_N\mathbf{w}_N^\star\|_2 \le \eta_N(\delta). \tag{28}$$

By a union bound, the uniform dual-concentration event and the basic concentration event hold simultaneously with probability at least $1-2\delta/3$, and hence at least $1-\delta$.

On the event (25), X_N has full row rank. Indeed, by (25) and Assumption 3,

$$\|X_N^\top\boldsymbol{\lambda}\|_* \ge (1-\varepsilon_{\mathrm{unif}})\kappa_N K(\boldsymbol{\lambda}) \ge (1-\varepsilon_{\mathrm{unif}})\kappa_N A\|\boldsymbol{\lambda}\|_2.$$

Since $\varepsilon_{\mathrm{unif}}<1$, no nonzero vector can lie in the kernel of $X_N^\top$. Fix $\mathbf{w}_S\in\mathcal{W}_S$, and write

$$\mathbf{u} := \mathbf{w}_S^\star - \mathbf{w}_S, \qquad \mathbf{r}(\mathbf{w}_S) := \mathbf{z}(\mathbf{w}_S) + X_N\mathbf{w}_N^\star.$$

Since X_N has full row rank, Lemma 4 gives

$$g(\mathbf{w}_S) = \max_{\boldsymbol{\lambda}\in\mathbb{R}^n}\left\{\mathbf{r}(\mathbf{w}_S)^\top\boldsymbol{\lambda} : \|X_N^\top\boldsymbol{\lambda}\|_* \le 1\right\}. \tag{29}$$

Using (26), Cauchy–Schwarz, and $\|\mathbf{u}\|_2\le 1$,

$$\frac{1}{\sqrt{n}}\|\mathbf{z}(\mathbf{w}_S)\|_\infty \le \frac{1}{\sqrt{n}}\left(\max_{i\in[n]}\|\mathbf{x}_{i,S}\|_2\,\|\mathbf{u}\|_2 + \|\boldsymbol{\xi}\|_\infty\right) \le U_S(\delta). \tag{30}$$

Using (27), and dropping the nonnegative quadratic term in $X_S^\top X_S$,

$$\frac{1}{\sqrt{n}}\|\mathbf{z}(\mathbf{w}_S)\|_2 \ge \left(\frac{1}{n}\|\boldsymbol{\xi}\|_2^2 - 2\|\mathbf{u}\|_2\left\|\frac{1}{n}X_S^\top\boldsymbol{\xi}\right\|_2\right)^{1/2} \ge L_S(\delta). \tag{31}$$

Together with (28), equations (30) and (31) imply

$$\frac{\|\mathbf{r}(\mathbf{w}_S)\|_\infty}{\|\mathbf{r}(\mathbf{w}_S)\|_2} \le \frac{\|\mathbf{z}(\mathbf{w}_S)\|_\infty + \|X_N\mathbf{w}_N^\star\|_2}{\|\mathbf{z}(\mathbf{w}_S)\|_2 - \|X_N\mathbf{w}_N^\star\|_2} \le \frac{U_S(\delta)+\eta_N(\delta)}{L_S(\delta)-\eta_N(\delta)} = \rho_{S,N}(\delta). \tag{32}$$

We next compare the Euclidean norms of $\mathbf{r}(\mathbf{w}_S)$ and $\mathbf{z}(\mathbf{w}_S)$. Their relative difference satisfies

$$\frac{\|\mathbf{r}(\mathbf{w}_S)-\mathbf{z}(\mathbf{w}_S)\|_2}{\|\mathbf{z}(\mathbf{w}_S)\|_2} = \frac{\|X_N\mathbf{w}_N^\star\|_2}{\|\mathbf{z}(\mathbf{w}_S)\|_2} \le \frac{\eta_N(\delta)}{L_S(\delta)} \le \rho_{S,N}(\delta). \tag{33}$$

The triangle and reverse triangle inequalities now give

$$(1-\rho_{S,N}(\delta))\|\mathbf{z}(\mathbf{w}_S)\|_2 \le \|\mathbf{r}(\mathbf{w}_S)\|_2 \le (1+\rho_{S,N}(\delta))\|\mathbf{z}(\mathbf{w}_S)\|_2. \tag{34}$$

By monotonicity of φ and the assumption on $\varphi(2\sqrt{\rho_{S,N}(\delta)})$, the conclusion of Corollary 3 applies on (25), uniformly over $\mathbf{w}_S\in\mathcal{W}_S$. Combining that conclusion with (29) gives

$$\begin{aligned}\kappa_N g(\mathbf{w}_S) &\ge (1-\varepsilon_{\mathrm{unif}})\left(1-\varphi(2\sqrt{\rho_{S,N}(\delta)})\right)\|\mathbf{r}(\mathbf{w}_S)\|_2 \\ &\ge (1-\varepsilon_{\mathrm{unif}})\left(1-\varphi(2\sqrt{\rho_{S,N}(\delta)})\right)(1-\rho_{S,N}(\delta))\|\mathbf{z}(\mathbf{w}_S)\|_2,\end{aligned}$$

and

$$\kappa_N g(\mathbf{w}_S) \le (1+2\varepsilon_{\mathrm{unif}})\left(1+2\varphi(2\sqrt{\rho_{S,N}(\delta)})+A^{-2}\min\left\{\sqrt{\rho_{S,N}(\delta)},1\right\}\right)\|\mathbf{r}(\mathbf{w}_S)\|_2$$
$$\le (1+2\varepsilon_{\mathrm{unif}})\left(1+2\varphi(2\sqrt{\rho_{S,N}(\delta)})+A^{-2}\min\left\{\sqrt{\rho_{S,N}(\delta)},1\right\}\right)$$
$$\times(1+\rho_{S,N}(\delta))\|\mathbf{z}(\mathbf{w}_S)\|_2.$$

Here the second inequality in each display uses (34). When $\rho_{S,N}(\delta)\le 1$, expanding the products and using $\varepsilon_{\mathrm{unif}}\le 1/2$ and $\varphi(2\sqrt{\rho_{S,N}(\delta)})\le 1/2$ gives the claimed bounds after choosing the constant in (24) sufficiently large depending only on A. When $\rho_{S,N}(\delta)>1$, the lower bound is trivial because $\varepsilon_{\mathrm{fit}}(\delta)\ge 1$, while the upper bound follows by absorbing $1+\rho_{S,N}(\delta)$ into (24). □

C.3 Proof of Theorem 3

Proof. Throughout this proof, C, C', and c denote constants that may depend on A. Set

$$\delta_0 := \frac{\delta}{2}, \qquad \varepsilon_{\mathrm{fit}} := \varepsilon_{\mathrm{fit}}(\delta_0), \qquad \Delta_S := \Delta_2(S,\delta_0), \qquad \Delta_0 := \Delta_2(\emptyset,\delta_0).$$

Since $\delta_0=\delta/2$, $r_S\ge 1$, and the theorem's sample-size condition holds, the definitions of Δ_2 give

$$\Delta_S+\Delta_0 \le 3\sqrt{\frac{r_S+\log(n/\delta)}{n}}.$$

We next verify the hypotheses of Lemma 5. Recall that η_N, L_S, U_S, and $\rho_{S,N}$ are defined in (23), while ρ and $\varepsilon_{\mathrm{flat}}$ are defined in (14). The preceding comparison and (16) give

$$L_S(\delta_0)>\eta_N(\delta_0), \qquad \rho_{S,N}(\delta_0)\le C\rho.$$

Together with the definition of $\varepsilon_{\mathrm{flat}}$, a sufficiently large C', and a sufficiently small c, this verifies all the hypotheses of Lemma 5. We also record various bounds that will be used below. Using (24), (16), and (15), and adjusting the constants in the theorem,

$$\begin{aligned} \varepsilon_{\mathrm{fit}} \le C\varepsilon_{\mathrm{flat}} \le \frac{1}{16}, \qquad C_{\mathrm{conc}}R_\xi^2\sigma^2\Delta_0 \le \frac{1}{4}\sigma^2, \\ C_{\mathrm{conc}}R_X^2(b_S^+)^2\Delta_S \le \frac{1}{2}(b_S^-)^2, \qquad C_{\mathrm{conc}}R_XR_\xi\sigma b_S^+\Delta_S \le \frac{1}{8}\min\left\{\sigma^2,(b_S^-)^2r_\delta\right\}. \end{aligned} \tag{35}$$

Next, we will prove the following inequality (after adjusting constants if necessary) which will be needed in the proof,

$$\frac{1}{8}(b_S^-)^2r_\delta^2 \ge 128\sigma^2\varepsilon_{\mathrm{fit}}+16\sigma\gamma_N\left(2\|(\mathbf{w}_S^\star,\mathbf{0})\|+\beta_Sr_\delta\right)+16\gamma_N^2\left(2\|(\mathbf{w}_S^\star,\mathbf{0})\|+\beta_Sr_\delta\right)^2. \tag{36}$$

To see this, after using $\varepsilon_{\mathrm{fit}}\le C\varepsilon_{\mathrm{flat}}$ and $(x+y)^2\le 2x^2+2y^2$, the only terms on the right-hand side not directly present in (15) are controlled, for any $\eta>0$, by

$$\sigma\gamma_N\beta_Sr_\delta \le \eta(b_S^-)^2r_\delta^2+\frac{\sigma^2}{4\eta}\left(\frac{\gamma_N\beta_S}{b_S^-}\right)^2, \qquad \gamma_N^2\beta_S^2r_\delta^2\le c^2(b_S^-)^2r_\delta^2.$$

Choose η and the constant c in (16) sufficiently small that, after accounting for the fixed numerical coefficients in (36), the two multiples of $(b_S^-)^2r_\delta^2$ sum to at most $(b_S^-)^2r_\delta^2/16$. The remaining terms are all present in (15); increasing its fixed constant makes their sum at most $(b_S^-)^2r_\delta^2/16$, proving (36).

Apply Lemma 5 and Lemma 2 with $T=S$, each with failure probability δ_0. By a union bound, there is an event $\mathcal{E}$ of probability at least $1-\delta$ on which X_N has full row rank and

$$(1-\varepsilon_{\mathrm{fit}})\|\mathbf{z}(\mathbf{w}_S)\|_2\le\kappa_Ng(\mathbf{w}_S)\le(1+\varepsilon_{\mathrm{fit}})\|\mathbf{z}(\mathbf{w}_S)\|_2 \qquad \text{for every } \mathbf{w}_S\in\mathcal{W}_S, \tag{37}$$

and

$$\begin{aligned} \left\|\frac{1}{n}X_S^\top X_S-\Sigma_S\right\|_{\mathrm{op}} &\le C_{\mathrm{conc}}R_X^2(b_S^+)^2\Delta_S, \\ \left\|\frac{1}{n}X_S^\top\boldsymbol{\xi}\right\|_2 &\le C_{\mathrm{conc}}R_XR_\xi\sigma b_S^+\Delta_S, \\ \left|\frac{1}{n}\|\boldsymbol{\xi}\|_2^2-\sigma^2\right| &\le C_{\mathrm{conc}}R_\xi^2\sigma^2\Delta_0. \end{aligned} \tag{38}$$

Work on $\mathcal{E}$. Define

$$F(\mathbf{w}_S) := \min_{\mathbf{w}_N:\, X_N\mathbf{w}_N = \mathbf{z}(\mathbf{w}_S) + X_N\mathbf{w}_N^\star} \|(\mathbf{w}_S, \mathbf{w}_N)\|,$$

where the defining constraint is feasible for every $\mathbf{w}_S$ because X_N has full row rank. The S-blocks of the minimum-norm interpolators are precisely the minimizers of F, with the corresponding N-blocks attaining the inner minimum in the definition of F. For every $\mathbf{w}_S \in \mathcal{W}_S$, the triangle inequality gives

$$g(\mathbf{w}_S) - \|(\mathbf{w}_S, \mathbf{0})\| \le F(\mathbf{w}_S) \le g(\mathbf{w}_S) + \|(\mathbf{w}_S, \mathbf{0})\|. \tag{39}$$

By (37), (39), and $\gamma_N = \kappa_N/\sqrt{n}$,

$$\gamma_N F(\mathbf{w}_S) \ge (1-\varepsilon_{\mathrm{fit}})\frac{\|\mathbf{z}(\mathbf{w}_S)\|_2}{\sqrt{n}} - \gamma_N\|(\mathbf{w}_S^\star, \mathbf{0})\| - \gamma_N\beta_S\|\mathbf{w}_S - \mathbf{w}_S^\star\|_2. \tag{40}$$

Evaluating the same bounds at $\mathbf{w}_S = \mathbf{w}_S^\star$, where $\mathbf{z}(\mathbf{w}_S^\star) = \boldsymbol{\xi}$, gives

$$\gamma_N F(\mathbf{w}_S^\star) \le (1+\varepsilon_{\mathrm{fit}})\frac{\|\boldsymbol{\xi}\|_2}{\sqrt{n}} + \gamma_N\|(\mathbf{w}_S^\star, \mathbf{0})\|. \tag{41}$$

Subtracting the upper bound (41) from the lower bound (40) yields, for every $\mathbf{w}_S \in \mathcal{W}_S$,

$$\begin{aligned}\gamma_N\left(F(\mathbf{w}_S) - F(\mathbf{w}_S^\star)\right) \ge (1-\varepsilon_{\mathrm{fit}})\frac{\|\mathbf{z}(\mathbf{w}_S)\|_2}{\sqrt{n}} - (1+\varepsilon_{\mathrm{fit}})\frac{\|\boldsymbol{\xi}\|_2}{\sqrt{n}} \\ - \gamma_N\left(2\|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S\|\mathbf{w}_S - \mathbf{w}_S^\star\|_2\right).\end{aligned} \tag{42}$$

Now fix $\mathbf{w}_S$ satisfying $\|\mathbf{w}_S - \mathbf{w}_S^\star\|_2 = r_\delta$, and write $\mathbf{u} := \mathbf{w}_S^\star - \mathbf{w}_S$. Since $r_\delta \le 1$, this point belongs to $\mathcal{W}_S$, so (42) applies. On $\mathcal{E}$,

$$\begin{aligned}\frac{1}{n}\|\mathbf{z}(\mathbf{w}_S)\|_2^2 &= \mathbf{u}^\top\left(\frac{1}{n}X_S^\top X_S\right)\mathbf{u} + \frac{2}{n}\mathbf{u}^\top X_S^\top\boldsymbol{\xi} + \frac{1}{n}\|\boldsymbol{\xi}\|_2^2 \\ &\ge \frac{1}{n}\|\boldsymbol{\xi}\|_2^2 + \left((b_S^-)^2 - C_{\mathrm{conc}}R_X^2(b_S^+)^2\Delta_S\right)r_\delta^2 - 2C_{\mathrm{conc}}R_X R_\xi\sigma b_S^+\Delta_S r_\delta.\end{aligned}$$

The bounds in (35) give

$$(b_S^-)^2 - C_{\mathrm{conc}}R_X^2(b_S^+)^2\Delta_S \ge \frac{1}{2}(b_S^-)^2, \qquad 2C_{\mathrm{conc}}R_X R_\xi\sigma b_S^+\Delta_S r_\delta \le \frac{1}{4}(b_S^-)^2 r_\delta^2.$$

Therefore,

$$\frac{1}{\sqrt{n}}\|\mathbf{z}(\mathbf{w}_S)\|_2 \ge \sqrt{\frac{1}{n}\|\boldsymbol{\xi}\|_2^2 + \frac{1}{4}(b_S^-)^2 r_\delta^2}. \tag{43}$$

Since $1-\varepsilon_{\mathrm{fit}} > 0$, substituting (43) into (42) yields

$$\begin{aligned}\gamma_N\left(F(\mathbf{w}_S) - F(\mathbf{w}_S^\star)\right) &\ge (1-\varepsilon_{\mathrm{fit}})\sqrt{\frac{1}{n}\|\boldsymbol{\xi}\|_2^2 + \frac{1}{4}(b_S^-)^2 r_\delta^2} - (1+\varepsilon_{\mathrm{fit}})\frac{\|\boldsymbol{\xi}\|_2}{\sqrt{n}} \\ &\quad - \gamma_N\left(2\|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S r_\delta\right) \\ &\overset{(*)}{\ge} \frac{(b_S^-)^2 r_\delta^2}{8\sqrt{\frac{1}{n}\|\boldsymbol{\xi}\|_2^2 + \frac{1}{4}(b_S^-)^2 r_\delta^2}} - 2\varepsilon_{\mathrm{fit}}\sqrt{\frac{1}{n}\|\boldsymbol{\xi}\|_2^2 + \frac{1}{4}(b_S^-)^2 r_\delta^2} \\ &\quad - \gamma_N\left(2\|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S r_\delta\right).\end{aligned} \tag{44}$$

Here $(*)$ uses, for $a, b \ge 0$ that $\sqrt{a+b} - \sqrt{a} \ge \frac{b}{2\sqrt{a+b}}$, and $\sqrt{a+b} + \sqrt{a} \le 2\sqrt{a+b}$. Combining (38) and (35) yields

$$\frac{\|\boldsymbol{\xi}\|_2^2}{n} \le \sigma^2 + C_{\mathrm{conc}}R_\xi^2\sigma^2\Delta_0 \le \frac{5}{4}\sigma^2,$$

and it follows that

$$\sqrt{\frac{1}{n}\|\boldsymbol{\xi}\|_2^2 + \frac{1}{4}(b_S^-)^2 r_\delta^2} \le 2\sigma + \frac{1}{2}b_S^- r_\delta.$$

Using this upper bound in both square-root terms in (44) gives

$$\gamma_N\left(F(\mathbf{w}_S) - F(\mathbf{w}_S^\star)\right) \ge \frac{(b_S^-)^2 r_\delta^2}{8\left(2\sigma + \frac{1}{2}b_S^- r_\delta\right)} - 2\varepsilon_{\mathrm{fit}}\left(2\sigma + \frac{1}{2}b_S^- r_\delta\right) - \gamma_N\left(2\|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S r_\delta\right).$$

Therefore, to ensure that $F(\mathbf{w}_S) > F(\mathbf{w}_S^\star)$, it suffices to show that

$$(b_S^-)^2 r_\delta^2 > 16\varepsilon_{\mathrm{fit}} \left(2\sigma + \frac{1}{2} b_S^- r_\delta\right)^2 + 8\gamma_N \left(2\|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S r_\delta\right) \left(2\sigma + \frac{1}{2} b_S^- r_\delta\right). \tag{45}$$

Since $(x+y)^2 \le 2x^2 + 2y^2$ and $\varepsilon_{\mathrm{fit}} \le 1/16$, the first term on the right-hand side is at most $128\sigma^2\varepsilon_{\mathrm{fit}} + \frac{1}{2}(b_S^-)^2 r_\delta^2$. Next, let $x := b_S^- r_\delta$ and $y := \gamma_N(2\|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S r_\delta)$ so that the second term on the right-hand side is $16\sigma y + 4xy$. Young's inequality gives $4xy \le x^2/4 + 16y^2$. Hence, the second term is at most

$$16\sigma\gamma_N \left(2\|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S r_\delta\right) + \frac{1}{4}(b_S^-)^2 r_\delta^2 + 16\gamma_N^2 \left(2\|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S r_\delta\right)^2.$$

Consequently, to prove (45), it suffices to show that

$$\frac{1}{4}(b_S^-)^2 r_\delta^2 > 128\sigma^2\varepsilon_{\mathrm{fit}} + 16\sigma\gamma_N \left(2\|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S r_\delta\right) + 16\gamma_N^2 \left(2\|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S r_\delta\right)^2.$$

By (36), the right-hand side is at most $(b_S^-)^2 r_\delta^2/8$. Thus (45) holds, proving

$$F(\mathbf{w}_S) > F(\mathbf{w}_S^\star) \qquad \text{whenever } \|\mathbf{w}_S - \mathbf{w}_S^\star\|_2 = r_\delta. \tag{46}$$

The function F is convex, hence (46) implies that every minimizer satisfies

$$\|\widehat{\mathbf{w}}_S - \mathbf{w}_S^\star\|_2 < r_\delta.$$

Indeed, otherwise the line segment between $\mathbf{w}_S^\star$ and a minimizer outside the radius-r_δ ball would intersect the sphere $\|\mathbf{w}_S - \mathbf{w}_S^\star\|_2 = r_\delta$ at a point with objective value at most $F(\mathbf{w}_S^\star)$, a contradiction.

It remains to bound the risk. Since $F(\widehat{\mathbf{w}}_S) \le F(\mathbf{w}_S^\star)$, the triangle inequality and the signal bound just proved give

$$\|(0, \widehat{\mathbf{w}}_N)\| \le F(\widehat{\mathbf{w}}_S) + \|(\widehat{\mathbf{w}}_S, \mathbf{0})\| \le F(\mathbf{w}_S^\star) + \|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S r_\delta.$$

Applying (41) and the $\|\boldsymbol{\xi}\|_2^2$ bound in (38) therefore gives

$$\|(0, \widehat{\mathbf{w}}_N)\| \le \frac{1+\varepsilon_{\mathrm{fit}}}{\gamma_N} \sqrt{\sigma^2 + C_{\mathrm{conc}} R_\xi^2 \sigma^2 \Delta_0} + 2\|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S r_\delta. \tag{47}$$

Finally, Assumption 1 gives

$$\mathcal{R}(\widehat{\mathbf{w}}) = \sum_{j \in S} b_j^2 (\widehat{w}_j - w_j^\star)^2 + \sum_{j \in N} b_j^2 (\widehat{w}_j - w_j^\star)^2.$$

The first term on the right-hand side is at most $(b_S^+)^2 r_\delta^2$. For the second term,

$$\sum_{j \in N} b_j^2 (\widehat{w}_j - w_j^\star)^2 \le 2 \sum_{j \in N} b_j^2 \widehat{w}_j^2 + 2 \sum_{j \in N} b_j^2 (w_j^\star)^2 \le 2\beta_{N,\Sigma}^2 \|(0, \widehat{\mathbf{w}}_N)\|^2 + 2\nu_N^2.$$

Equations (35) and (47) imply

$$\|(0, \widehat{\mathbf{w}}_N)\| \le C \left(\frac{\sigma}{\gamma_N} + \|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S r_\delta\right).$$

Substituting this bound proves the desired risk bound. □

C.4 Proof of Corollary 2

Proof. Since $\varepsilon_{\mathrm{flat}} \ge \sqrt{\rho}$ and $\rho \ge \nu_N/\sigma$, the first limit in the corollary statement implies $\rho \to 0$ and $\nu_N \to 0$. The definition of ρ and the uniform bounds in the statement also imply

$$\frac{r_S + \log(n/\delta_n)}{n} \longrightarrow 0.$$

Moreover, $\|(\mathbf{w}_S^\star, \mathbf{0})\| \le \beta_S \|\mathbf{w}_S^\star\|_2$. Thus, the assumptions of the corollary and the definitions in (14)–(15) imply that the regularity condition in (16) holds for all sufficiently large n and that $r_{\delta_n} \to 0$.

It remains to apply the risk bound in Theorem 3. Its first and last terms tend to zero, while

$$\beta_{N,\Sigma} \left(\frac{\sigma}{\gamma_N} + \|(\mathbf{w}_S^\star, \mathbf{0})\| + \beta_S r_{\delta_n}\right) = \frac{\beta_{N,\Sigma}}{\gamma_N} \left(\sigma + \gamma_N \|(\mathbf{w}_S^\star, \mathbf{0})\| + \gamma_N \beta_S r_{\delta_n}\right) \longrightarrow 0.$$

Applying Theorem 3 with $\delta = \delta_n$ proves the claim. □

D Proof of Theorem 1 and Related Results

D.1 Setting

Fix $p \in (1, \infty]$, and let $q \in [1, \infty)$ be its dual exponent, defined by $1/p + 1/q = 1$.

Definition 1. *For $\alpha > 0$, define the ψ_α Orlicz functional of a real-valued random variable Y by*

$$\|Y\|_{\psi_\alpha} := \inf\left\{ s > 0 : \mathbb{E}\exp\left(\left|\frac{Y}{s}\right|^\alpha\right) \le 2 \right\}.$$

When $\alpha \ge 1$ this is an Orlicz norm; for $\alpha < 1$ we use it as the standard ψ_α quasi-norm.

We locally index the columns of X_N by $[|N|]$. Throughout this section, $s_N > 0$, and the noise-fitting coordinates have the product form

$$x_{i,k} = s_N \zeta_{i,k}, \qquad i \in [n],\ k \in [|N|],$$

where the variables $\{\zeta_{i,k}\}_{i,k}$ are mutually independent and, for every $i \in [n]$ and $k \in [|N|]$,

$$\mathbb{E}\zeta_{i,k} = 0, \qquad \mathbb{E}\zeta_{i,k}^2 = 1, \qquad \|\zeta_{i,k}\|_{\psi_2} \le R_X.$$

It is well known that any sub-Gaussian random variable satisfies $\|\zeta_{i,k}\|_{\psi_2} \ge \sqrt{\mathbb{E}\zeta_{i,k}^2} = 1$, so $R_X \ge 1$. For $\boldsymbol{\lambda} \in \mathbb{R}^n$ and $k \in [|N|]$, write

$$b_{\boldsymbol{\lambda},k} := \sum_{i=1}^n \lambda_i \zeta_{i,k}.$$

For fixed $\boldsymbol{\lambda}$, the variables $\{b_{\boldsymbol{\lambda},k}\}_{k=1}^{|N|}$ are independent. Moreover,

$$X_N^\top \boldsymbol{\lambda} = s_N (b_{\boldsymbol{\lambda},1}, \dots, b_{\boldsymbol{\lambda},|N|}).$$

Lemma 6 (Centering a ψ_α random variable)**.** *There is an absolute constant $C_{\mathrm{ctr}} > 0$ such that, for every $\alpha \in (0, 2]$ and every real-valued random variable Y with $\|Y\|_{\psi_\alpha} < \infty$,*

$$\|Y - \mathbb{E}Y\|_{\psi_\alpha} \le \left(\frac{C_{\mathrm{ctr}}}{\alpha}\right)^{1/\alpha} \|Y\|_{\psi_\alpha}. \tag{48}$$

Proof. Fix $K > \|Y\|_{\psi_\alpha}$. Markov's inequality and Def. 1 give for any $t > 0$,

$$\Pr(|Y| \ge Kt) = \Pr\left(\exp\left(\left|\frac{Y}{K}\right|^\alpha\right) \ge e^{t^\alpha}\right) \le e^{-t^\alpha}\mathbb{E}\exp\left(\left|\frac{Y}{K}\right|^\alpha\right) \le 2e^{-t^\alpha},$$

and hence, by tail integration,

$$|\mathbb{E}Y| \le \mathbb{E}|Y| \le 2K\int_0^\infty e^{-t^\alpha}\,dt = 2\Gamma(1 + 1/\alpha)K,$$

where Γ is the Euler gamma function. By Stirling's inequality, $\Gamma(1 + 1/\alpha) \le (C/\alpha)^{1/\alpha}$ for $0 < \alpha \le 2$. As such, there is an absolute constant $C_1 > 0$ such that

$$|\mathbb{E}Y|^\alpha \le \frac{C_1}{\alpha}K^\alpha, \qquad \text{where} \quad 0 < \alpha \le 2.$$

Also, $\alpha \le 2$ implies $|x - y|^\alpha \le 2(|x|^\alpha + |y|^\alpha)$. Therefore, for $D_\alpha := (C_{\mathrm{ctr}}/\alpha)^{1/\alpha}$, where C_{ctr} will be chosen below,

$$\left|\frac{Y - \mathbb{E}Y}{D_\alpha K}\right|^\alpha \le \frac{2\alpha}{C_{\mathrm{ctr}}}\left|\frac{Y}{K}\right|^\alpha + \frac{2C_1}{C_{\mathrm{ctr}}}.$$

Choose the absolute constant C_{ctr} large enough that

$$\frac{2\alpha}{C_{\mathrm{ctr}}} \le \frac{1}{2} \quad \text{for all } \alpha \le 2, \qquad \frac{2C_1}{C_{\mathrm{ctr}}} \le \frac{\log 2}{2}.$$

Exponentiating the preceding bound and applying Jensen's inequality now gives

$$\mathbb{E}\exp\left(\left|\frac{Y - \mathbb{E}Y}{D_\alpha K}\right|^\alpha\right) \le e^{(\log 2)/2}\left[\mathbb{E}\exp\left(\left|\frac{Y}{K}\right|^\alpha\right)\right]^{1/2} \le 2.$$

Letting $K \downarrow \|Y\|_{\psi_\alpha}$ proves (48). □

D.2 Moments and Expected Dual-Norm Geometry

Lemma 7 (Moments of sub-Gaussian linear forms)**.** *Under Setting D.1, there exist absolute constants $c, C > 0$ such that, for every $r \geq 1$, every $k \in [|N|]$, and every $\boldsymbol{\lambda} \in \mathbb{S}^{n-1}$,*

$$\frac{1}{\sqrt{2}} \left(\frac{c}{R_X^4} \right)^{1/r} \leq \left(\mathbb{E} |b_{\boldsymbol{\lambda},k}|^r \right)^{1/r} \leq C R_X \sqrt{r}.$$

Proof. Choose an absolute constant $C_0 \geq 1$ large enough that the standard sub-Gaussian sum and moment bounds (Vershynin, 2026, Propositions 2.7.1 and 2.6.1) give $\|b_{\boldsymbol{\lambda},k}\|_{\psi_2} \leq C_0 R_X$ and

$$\left(\mathbb{E} |b_{\boldsymbol{\lambda},k}|^r \right)^{1/r} \leq C_0 R_X \sqrt{r}.$$

For the lower bound, $\mathbb{E} b_{\boldsymbol{\lambda},k}^2 = 1$, and the upper bound with $r = 4$ gives $\mathbb{E} b_{\boldsymbol{\lambda},k}^4 \leq (2C_0 R_X)^4$. Applying Paley–Zygmund to $b_{\boldsymbol{\lambda},k}^2$ yields

$$\Pr\left(|b_{\boldsymbol{\lambda},k}| \geq \frac{1}{\sqrt{2}} \right) \geq \frac{1}{4(2C_0 R_X)^4}.$$

Set $c_0 := 1/[4(2C_0)^4]$. The elementary inequality $\mathbb{E} Z \geq t \Pr(Z \geq t)$, applied to $Z = |b_{\boldsymbol{\lambda},k}|^r$, gives

$$\mathbb{E} |b_{\boldsymbol{\lambda},k}|^r \geq 2^{-r/2} \frac{c_0}{R_X^4}.$$

Taking the r-th root proves the result with $c = c_0$ and $C = C_0$. □

Lemma 8 (Expected ℓ_q norm)**.** *Under Setting D.1, there exist absolute constants $c, C > 0$ such that, for every $\boldsymbol{\lambda} \in \mathbb{S}^{n-1}$,*

$$\frac{c}{R_X^4} s_N |N|^{1/q} \leq \mathbb{E} \|X_N^\top \boldsymbol{\lambda}\|_q \leq C R_X \sqrt{q}\, s_N |N|^{1/q}.$$

Proof. Take C to be the upper constant in Lemma 7, and take c to be $1/\sqrt{2}$ times that lemma's lower constant. By Jensen's inequality, using the concavity of $u \mapsto u^{1/q}$,

$$\mathbb{E} \|X_N^\top \boldsymbol{\lambda}\|_q = s_N \mathbb{E} \left(\sum_{k=1}^{|N|} |b_{\boldsymbol{\lambda},k}|^q \right)^{1/q} \leq s_N \left(\sum_{k=1}^{|N|} \mathbb{E} |b_{\boldsymbol{\lambda},k}|^q \right)^{1/q} \leq C R_X \sqrt{q}\, s_N |N|^{1/q}.$$

For the lower bound, use $\|\mathbf{v}\|_q \geq |N|^{1/q-1} \|\mathbf{v}\|_1$ for $\mathbf{v} \in \mathbb{R}^N$. Then

$$\mathbb{E} \|X_N^\top \boldsymbol{\lambda}\|_q \geq s_N |N|^{1/q-1} \sum_{k=1}^{|N|} \mathbb{E} |b_{\boldsymbol{\lambda},k}| \geq \frac{c}{R_X^4} s_N |N|^{1/q}.$$

□

D.3 Concentration of the Empirical Dual Norm

Lemma 9 (Sub-Weibull mean concentration)**.** *Let $m \geq 1$, and let $Y_1, \ldots, Y_m$ be independent mean-zero random variables with $\|Y_i\|_{\psi_\alpha} \leq B$ for some $\alpha, B > 0$. Then, for every $u > 0$,*

$$\Pr\left(\left| \frac{1}{m} \sum_{i=1}^m Y_i \right| \geq u \right) \leq 2 \exp\left(-c_\alpha \min\left\{ \frac{m u^2}{B^2}, \left(\frac{m^{\min(1/\alpha,1)} u}{B} \right)^\alpha \right\} \right),$$

where $c_\alpha > 0$ depends only on α and is non-decreasing for $\alpha \in (0, 2]$.

Proof. By (Kuchibhotla & Chakrabortty, 2022, Theorem 3.1), there is a constant $C_\alpha > 0$, depending only on α, such that, for every $t \geq 0$,

$$\Pr\left(\left| \frac{1}{m} \sum_{i=1}^m Y_i \right| \geq C_\alpha B \left(\sqrt{\frac{t}{m}} + \frac{t^{1/\alpha}}{m^{\min(1/\alpha,1)}} \right) \right) \leq 2e^{-t}.$$

Apply this bound with

$$t = c_\alpha \min\left\{ \frac{mu^2}{B^2}, \left(\frac{m^{\min(1/\alpha,1)} u}{B} \right)^\alpha \right\},$$

where $c_\alpha > 0$ is chosen small enough that $C_\alpha(c_\alpha^{1/2} + c_\alpha^{1/\alpha}) \le 1$. Since

$$B\sqrt{\frac{t}{m}} \le \sqrt{c_\alpha} u, \qquad B\frac{t^{1/\alpha}}{m^{\min(1/\alpha,1)}} \le c_\alpha^{1/\alpha} u,$$

the threshold in the cited bound is at most u, proving the stated tail inequality.

Finally, the explicit formula for C_α in the cited theorem is bounded above on every interval $[\alpha_0, 2]$ by a constant depending only on α_0. Hence the admissible choices of c_α can be made uniformly positive on every such interval. For $\alpha \in (0, 2]$, replace each choice by the infimum of the chosen constants over $[\alpha, 2]$. The resulting constants remain positive and admissible, and are non-decreasing in α, as claimed. □

Lemma 10 (Bias of the empirical q-root)**.** *Let $Y_1, \dots, Y_m$ be independent nonnegative random variables with $\mathbb{E}Y_i^2 < \infty$. Then, for every $q \ge 1$,*

$$\left| \mathbb{E}\left(\frac{1}{m}\sum_{i=1}^m Y_i \right)^{1/q} - \left(\frac{1}{m}\sum_{i=1}^m \mathbb{E}Y_i \right)^{1/q} \right| \le \left(\frac{1}{m^2}\sum_{i=1}^m \mathit{Var}(Y_i) \right)^{1/(2q)}.$$

Proof. For $u, v \ge 0$, the map $u \mapsto u^{1/q}$ is $1/q$-Hölder, so

$$|u^{1/q} - v^{1/q}| \le |u - v|^{1/q}.$$

Using this and Jensen's inequality,

$$\begin{aligned}
\left| \mathbb{E}\left(\frac{1}{m}\sum_{i=1}^m Y_i \right)^{1/q} - \left(\frac{1}{m}\sum_{i=1}^m \mathbb{E}Y_i \right)^{1/q} \right| &\le \mathbb{E}\left| \frac{1}{m}\sum_{i=1}^m (Y_i - \mathbb{E}Y_i) \right|^{1/q} \\
&\le \left[\mathbb{E}\left| \frac{1}{m}\sum_{i=1}^m (Y_i - \mathbb{E}Y_i) \right|^2 \right]^{1/(2q)} \\
&= \left(\frac{1}{m^2}\sum_{i=1}^m \mathrm{Var}(Y_i) \right)^{1/(2q)}.
\end{aligned}$$

The equality uses the independence and centering of the variables $Y_i - \mathbb{E}Y_i$. □

Lemma 11 (Fixed-vector concentration)**.** *Under Setting D.1, there is an absolute constant $C > 0$ and, for every $q \ge 1$, a constant $c(q) > 0$ that may depend on q such that for every $\boldsymbol{\lambda} \in \mathbb{S}^{n-1}$ and every $\tau \in (0, 1)$,*

$$\begin{aligned}
&\Pr\left(\left| \frac{\|X_N^\top \boldsymbol{\lambda}\|_q}{\mathbb{E}\|X_N^\top \boldsymbol{\lambda}\|_q} - 1 \right| > C R_X^5 \sqrt{q} \left(\tau + |N|^{-1/(2q)} \right) \right) \\
&\quad \le 2\exp\left(-\frac{c(q)}{R_X^{2q+8}} \min\left\{ |N|\tau^2, (|N|\tau)^{2/q} \right\} \right).
\end{aligned}$$

Moreover, for every $q_0 < \infty$, the constant $c(q)$ is bounded below by a positive constant depending only on q_0, uniformly over $q \in [1, q_0]$.

Proof. Fix $\boldsymbol{\lambda} \in \mathbb{S}^{n-1}$, and write

$$Y_k := |b_{\boldsymbol{\lambda},k}|^q, \qquad \bar{\mu} := \frac{1}{|N|}\sum_{k=1}^{|N|} \mathbb{E}Y_k.$$

By Lemma 7, there are absolute constants $c, C > 0$ such that

$$2^{-q/2}\frac{c}{R_X^4} \le \mathbb{E}Y_k \le (CR_X\sqrt{q})^q \quad \text{for every } k \in [|N|], \qquad 2^{-q/2}\frac{c}{R_X^4} \le \bar{\mu} \le (CR_X\sqrt{q})^q. \tag{49}$$

The standard sub-Gaussian sum bound gives $\|b_{\boldsymbol{\lambda},k}\|_{\psi_2} \le CR_X$, and hence

$$\|Y_k\|_{\psi_{2/q}} = \|b_{\boldsymbol{\lambda},k}\|_{\psi_2}^q \le (CR_X)^q.$$

Applying Lemma 6 with $\alpha = 2/q$ yields

$$\|Y_k - \mathbb{E}Y_k\|_{\psi_{2/q}} \le (CR_X\sqrt{q})^q,$$

after adjusting the absolute constant C.

Set $B := (CR_X\sqrt{q})^q$. The bounds in (49) show that there is a constant $a(q) > 0$, depending only on q, such that

$$\frac{\bar{\mu}}{B} \ge \frac{a(q)}{R_X^{q+4}}.$$

We may assume that $a(q) \le 1$. Moreover, for every $q_0 < \infty$, it can be chosen uniformly bounded below over $q \in [1, q_0]$. Applying Lemma 9 with $\alpha = 2/q$ and $u = \tau\bar{\mu}$ gives

$$\Pr\left(\left|\frac{1}{|N|}\sum_{k=1}^{|N|} Y_k - \bar{\mu}\right| > \tau\bar{\mu}\right) \le 2\exp\left(-c_{2/q}\min\left\{\frac{|N|\tau^2\bar{\mu}^2}{B^2}, \left(\frac{|N|^{\min(q/2,1)}\tau\bar{\mu}}{B}\right)^{2/q}\right\}\right).$$

If $q \ge 2$, the second term in the minimum is $(|N|\tau\bar{\mu}/B)^{2/q}$. If $1 \le q \le 2$, it is $|N|\tau^{2/q}(\bar{\mu}/B)^{2/q}$, which is at least $|N|\tau^2(\bar{\mu}/B)^{2/q}$ because $\tau \in (0,1)$. In this case, $(|N|\tau)^{2/q} \ge |N|\tau^2$ as well. Since $R_X \ge 1$ and $a(q) \le 1$, the preceding lower bound for $\bar{\mu}/B$ therefore gives, in both cases,

$$\min\left\{\frac{|N|\tau^2\bar{\mu}^2}{B^2}, \left(\frac{|N|^{\min(q/2,1)}\tau\bar{\mu}}{B}\right)^{2/q}\right\} \ge \frac{a(q)^2}{R_X^{2q+8}}\min\left\{|N|\tau^2, (|N|\tau)^{2/q}\right\}.$$

Set $c(q) := c_{2/q}a(q)^2$. The preceding bound becomes

$$\Pr\left(\left|\frac{1}{|N|}\sum_{k=1}^{|N|} Y_k - \bar{\mu}\right| > \tau\bar{\mu}\right) \le 2\exp\left(-\frac{c(q)}{R_X^{2q+8}}\min\left\{|N|\tau^2, (|N|\tau)^{2/q}\right\}\right). \tag{50}$$

Suppose that

$$\left|\frac{1}{|N|}\sum_{k=1}^{|N|} Y_k - \bar{\mu}\right| \le \tau\bar{\mu}.$$

Using $|s^{1/q} - 1| \le |s - 1|$, Lemma 10, and the triangle inequality then gives

$$\begin{aligned}
\left|\|X_N^\top\boldsymbol{\lambda}\|_q - \mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q\right| &\le s_N|N|^{1/q}\bar{\mu}^{1/q}\left|\left(\frac{|N|^{-1}\sum_{k=1}^{|N|} Y_k}{\bar{\mu}}\right)^{1/q} - 1\right| \\
&\quad + s_N|N|^{1/q}\left(\frac{1}{|N|^2}\sum_{k=1}^{|N|}\mathrm{Var}(Y_k)\right)^{1/(2q)} \\
&\le \tau s_N|N|^{1/q}\bar{\mu}^{1/q} + s_N|N|^{1/q}\left(\frac{1}{|N|^2}\sum_{k=1}^{|N|}\mathrm{Var}(Y_k)\right)^{1/(2q)} \\
&\le CR_X\sqrt{q}\, s_N|N|^{1/q}\left(\tau + |N|^{-1/(2q)}\right),
\end{aligned}$$

where the last inequality uses (49) and the upper moment bound in Lemma 7, applied with $2q$, after adjusting the absolute constant C. Moreover, Lemma 8 gives $\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q \ge (c/R_X^4)s_N|N|^{1/q}$. Hence, after another adjustment of C,

$$\left|\frac{\|X_N^\top\boldsymbol{\lambda}\|_q}{\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q} - 1\right| \le CR_X^5\sqrt{q}\left(\tau + |N|^{-1/(2q)}\right).$$

Thus, the deviation in the statement can occur only if the sample-mean deviation exceeds $\tau\bar{\mu}$. Its probability is bounded by (50), proving the stated deviation bound.

Finally, if $q \le q_0$, the monotonicity of the constants in Lemma 9 gives $c_{2/q} \ge c_{2/q_0} > 0$. Moreover, $a(q)$ is bounded below by a positive constant depending only on q_0 over the same interval. Hence so is $c(q) = c_{2/q}a(q)^2$, as claimed. □

Proposition 2 (Uniform concentration)**.** *Under Setting D.1, there is an absolute constant $C > 0$ and, for every $q \ge 1$, a constant $c(q) > 0$ that may depend on q such that for every $\tau, \delta \in (0,1)$, with $L = \sqrt{n} + \sqrt{\log(2|N|/\delta)}$,*

$$\Pr\left(\sup_{\boldsymbol{\lambda}\in\mathbb{S}^{n-1}}\left|\frac{\|X_N^\top\boldsymbol{\lambda}\|_q}{\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q} - 1\right| > CR_X^5\sqrt{q}\left(\tau + |N|^{-1/(2q)}\right)\right)$$
$$\le \delta + 2\exp\left(n\log\left(\frac{CR_X^{10}\sqrt{q}\,L}{\tau}\right) - \frac{c(q)}{R_X^{2q+8}}\min\left\{|N|\tau^2, (|N|\tau)^{2/q}\right\}\right).$$

Moreover, for every $q_0 < \infty$, $c(q)$ is bounded below by a positive constant depending only on q_0, uniformly over $q \in [1, q_0]$.

Proof. Choose absolute constants $0 < c_0 \le 1 \le C_0$ so that the expectation bounds in Lemma 8 and the sub-Gaussian tail bound below hold. Let $\mathbf{z}_N \in \mathbb{R}^{n\times N}$ be the matrix with entries $\zeta_{i,k}$, so that $X_N = s_N\mathbf{z}_N$. For each column, the sub-Gaussian linear-form bound (Vershynin, 2026, Proposition 2.7.1) and a Euclidean $1/2$-net of $\mathbb{S}^{n-1}$ give

$$\Pr\left(\|(\mathbf{z}_N)_{:,k}\|_2 > \frac{C_0}{2}R_X(\sqrt{n} + \sqrt{t})\right) \le 2e^{-t}, \qquad t \ge 0.$$

Taking $t = \log(2|N|/\delta)$ and applying a union bound over $k \in [|N|]$ show that, with probability at least $1 - \delta$,

$$\max_{k\in[|N|]}\|(\mathbf{z}_N)_{:,k}\|_2 \le C_0R_XL. \tag{51}$$

We work on this event. For every $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \mathbb{S}^{n-1}$,

$$\|X_N^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}')\|_q \le C_0R_Xs_NL|N|^{1/q}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2. \tag{52}$$

Lemma 8 and homogeneity also give

$$\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q \ge \frac{c_0}{R_X^4}s_N|N|^{1/q}, \qquad \mathbb{E}\|X_N^\top\mathbf{v}\|_q \le C_0R_X\sqrt{q}\,s_N|N|^{1/q}\|\mathbf{v}\|_2 \tag{53}$$

for every $\boldsymbol{\lambda} \in \mathbb{S}^{n-1}$ and $\mathbf{v} \in \mathbb{R}^n$. Moreover, (51) implies

$$\|X_N^\top\boldsymbol{\lambda}\|_q \le C_0R_Xs_NL|N|^{1/q} \qquad \text{for every } \boldsymbol{\lambda} \in \mathbb{S}^{n-1}. \tag{54}$$

Define $\Phi(\boldsymbol{\lambda}) := \|X_N^\top\boldsymbol{\lambda}\|_q/\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q$, and set $C_1 := C_0/c_0 + C_0^2/c_0^2$. For $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \mathbb{S}^{n-1}$, equations (52)–(54) give

$$|\Phi(\boldsymbol{\lambda}) - \Phi(\boldsymbol{\lambda}')| \le \frac{\|X_N^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}')\|_q}{\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q} + \|X_N^\top\boldsymbol{\lambda}'\|_q\frac{\mathbb{E}\|X_N^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}')\|_q}{\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q\mathbb{E}\|X_N^\top\boldsymbol{\lambda}'\|_q}$$
$$\le C_1R_X^{10}\sqrt{q}\,L\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2. \tag{55}$$

Let $\mathcal{N}_r$ be a Euclidean r-net of $\mathbb{S}^{n-1}$, where

$$r := \frac{\tau + |N|^{-1/(2q)}}{4C_1R_X^{10}\sqrt{q}\,L}.$$

Since $r < 1$ and $\tau + |N|^{-1/(2q)} \ge \tau$, the standard volumetric bound allows us to choose the net so that

$$|\mathcal{N}_r| \le \left(\frac{12C_1R_X^{10}\sqrt{q}\,L}{\tau}\right)^n. \tag{56}$$

Let $C_2 > 0$ be the absolute constant in Lemma 11, and choose the absolute constant C in the proposition so that

$$C \geq \max\{12C_1, C_2 + 1/4\}.$$

If

$$\sup_{\boldsymbol{\lambda} \in \mathbb{S}^{n-1}} |\Phi(\boldsymbol{\lambda}) - 1| > C R_X^5 \sqrt{q} \left(\tau + |N|^{-1/(2q)}\right),$$

then (55) and the definition of r imply that there is some $\boldsymbol{\lambda}_0 \in \mathcal{N}_r$ with

$$|\Phi(\boldsymbol{\lambda}_0) - 1| > C_2 R_X^5 \sqrt{q} \left(\tau + |N|^{-1/(2q)}\right). \tag{57}$$

The proposition now follows from a union bound, using Lemma 11, (56), and the probability δ that (51) fails. We take $c(q)$ to be the constant in Lemma 11, which also gives the asserted uniform lower bound on $c(q)$ over bounded intervals. □

D.4 Gaussian Approximation via a Quantitative CLT

For $r \geq 1$, let W_r denote the r-Wasserstein distance on probability measures on $\mathbb{R}$:

$$W_r(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathbb{R}\times\mathbb{R}} |x - y|^r \, d\pi(x, y) \right)^{1/r},$$

where $\Pi(\mu, \nu)$ is the set of all couplings of μ and ν.

Theorem 4 (Bobkov (2018), Theorem 1.1)**.** *Fix $r \geq 1$, and let $U_1, \ldots, U_n$ be independent real-valued random variables such that $\mathbb{E}U_i = 0$ and $\mathbb{E}|U_i|^{r+2} < \infty$. Set $V = \sum_{i=1}^n U_i$, and assume $\mathbb{E}V^2 = 1$. There exists a constant $A(r) > 0$, which may be chosen continuously in r, such that*

$$W_r\left(\mathcal{L}(V), \mathcal{N}(0,1)\right) \leq A(r) \left(\sum_{i=1}^n \mathbb{E}|U_i|^{r+2} \right)^{1/r}.$$

We will use this theorem only through the following moment comparison, stated in a slightly more general non-identically distributed form.

Lemma 12 (Gaussian moment comparison for sub-Gaussian linear forms)**.** *Let $\zeta_1, \ldots, \zeta_n$ be independent random variables satisfying*

$$\mathbb{E}\zeta_i = 0, \qquad \mathbb{E}\zeta_i^2 = 1, \qquad \|\zeta_i\|_{\psi_2} \leq R_X.$$

There is a locally bounded function $M : [1, \infty) \to (0, \infty)$ such that for every $q \geq 1$, every $\boldsymbol{\lambda} \in \mathbb{S}^{n-1}$, and $G \sim \mathcal{N}(0, 1)$,

$$\left| \mathbb{E} \left| \sum_{i=1}^n \lambda_i \zeta_i \right|^q - \mathbb{E}|G|^q \right| \leq M(q) R_X^{q+2/q} \|\boldsymbol{\lambda}\|_\infty.$$

Proof. Fix an absolute constant $C_0 \geq 1$ large enough for the standard sub-Gaussian moment and linear-form bounds used below, and set $W := \sum_{i=1}^n \lambda_i \zeta_i$. Then $\mathbb{E}W = 0$ and $\mathbb{E}W^2 = 1$. Moreover,

$$\sum_{i=1}^n \mathbb{E}|\lambda_i \zeta_i|^{q+2} \leq \left(C_0 R_X \sqrt{q+2}\right)^{q+2} \sum_{i=1}^n |\lambda_i|^{q+2} \leq \left(C_0 R_X \sqrt{q+2}\right)^{q+2} \|\boldsymbol{\lambda}\|_\infty^q.$$

Theorem 4, applied with $U_i = \lambda_i \zeta_i$, therefore gives

$$W_q\left(\mathcal{L}(W), \mathcal{N}(0,1)\right) \leq A(q) \left(C_0 R_X \sqrt{q+2}\right)^{1+2/q} \|\boldsymbol{\lambda}\|_\infty. \tag{58}$$

For any coupling $(\widetilde{W}, \widetilde{G})$ with $\widetilde{W} \sim \mathcal{L}(W)$ and $\widetilde{G} \sim \mathcal{N}(0, 1)$, the reverse triangle inequality in L_q gives

$$\left| (\mathbb{E}|W|^q)^{1/q} - (\mathbb{E}|G|^q)^{1/q} \right| \leq \|\widetilde{W} - \widetilde{G}\|_{L_q}.$$

Taking the infimum over all couplings and using (58),

$$\left|(\mathbb{E}|W|^q)^{1/q} - (\mathbb{E}|G|^q)^{1/q}\right| \le A(q)\left(C_0 R_X\sqrt{q+2}\right)^{1+2/q}\|\boldsymbol{\lambda}\|_\infty.$$

The linear form W is C_0R_X-sub-Gaussian. As noted after Setting D.1, $R_X \ge 1$, so

$$(\mathbb{E}|W|^q)^{1/q} \le C_0 R_X\sqrt{q}, \qquad (\mathbb{E}|G|^q)^{1/q} \le C_0 R_X\sqrt{q}.$$

Using $|x^q - y^q| \le q(x+y)^{q-1}|x-y|$ for $x, y \ge 0$, we obtain

$$|\mathbb{E}|W|^q - \mathbb{E}|G|^q| \le M(q) R_X^{q+2/q}\|\boldsymbol{\lambda}\|_\infty,$$

where

$$M(q) := q(2C_0\sqrt{q})^{q-1}A(q)\left(C_0\sqrt{q+2}\right)^{1+2/q}.$$

Since $A(q)$ is continuous, $M(q)$ is locally bounded. □

D.4.1 Expected ℓ_q geometry on flat directions

Lemma 13 (Expected ℓ_q geometry on flat directions)**.** *Under Setting D.1, let $G \sim \mathcal{N}(0,1)$, $m_q := \mathbb{E}|G|^q$, and $\kappa_N = s_N(|N|m_q)^{1/q}$. There are absolute constants $c, C > 0$ and a locally bounded function $D : [1,\infty) \to (0,\infty)$ such that every nonzero $\boldsymbol{\lambda} \in \mathbb{R}^n$ satisfies*

$$\frac{\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q}{\kappa_N\|\boldsymbol{\lambda}\|_2} \ge \frac{c}{R_X^4\sqrt{q}}$$

and

$$\left|\frac{\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q}{\kappa_N\|\boldsymbol{\lambda}\|_2} - 1\right| \le CR_X|N|^{-1/(2q)} + D(q)R_X^{q+4-2/q}\frac{\|\boldsymbol{\lambda}\|_\infty}{\|\boldsymbol{\lambda}\|_2}.$$

Proof. By homogeneity, it suffices to consider $\boldsymbol{\lambda} \in \mathbb{S}^{n-1}$. Lemma 8 and the standard bound $m_q^{1/q} \le C\sqrt{q}$ give, for absolute constants $c, C > 0$,

$$\frac{\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q}{\kappa_N} \ge \frac{c}{R_X^4\sqrt{q}}.$$

For the second bound, set

$$Z := \frac{1}{|N|}\sum_{k=1}^{|N|}|b_{\boldsymbol{\lambda},k}|^q, \qquad \mu := \mathbb{E}Z.$$

Since $\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q/\kappa_N = m_q^{-1/q}\mathbb{E}Z^{1/q}$, the triangle inequality gives

$$\left|\frac{\mathbb{E}\|X_N^\top\boldsymbol{\lambda}\|_q}{\kappa_N} - 1\right| \le m_q^{-1/q}\left|\mathbb{E}Z^{1/q} - \mu^{1/q}\right| + m_q^{-1/q}\left|\mu^{1/q} - m_q^{1/q}\right|.$$

Lemma 10, applied to the variables $|b_{\boldsymbol{\lambda},k}|^q$, together with Lemma 7 applied with $r = 2q$ and the Gaussian lower bound $m_q^{1/q} \ge c\sqrt{q}$, gives

$$m_q^{-1/q}\left|\mathbb{E}Z^{1/q} - \mu^{1/q}\right| \le CR_X|N|^{-1/(2q)}.$$

To bound the second term, we apply Lemma 12 separately to each $b_{\boldsymbol{\lambda},k}$, obtaining

$$|\mu - m_q| \le M(q)R_X^{q+2/q}\|\boldsymbol{\lambda}\|_\infty.$$

Set $a(q) := \min\{2^{-q/2}c, m_q\}$. The lower moment bound in Lemma 7 and $R_X \ge 1$ imply that both μ and m_q are at least $a(q)R_X^{-4}$. The mean value theorem therefore gives

$$m_q^{-1/q}\left|\mu^{1/q} - m_q^{1/q}\right| \le \frac{a(q)^{1/q-1}}{q\,m_q^{1/q}}R_X^{4-4/q}|\mu - m_q| \le D(q)R_X^{q+4-2/q}\|\boldsymbol{\lambda}\|_\infty,$$

where we may take $D(q) = a(q)^{1/q-1}M(q)/(q\,m_q^{1/q})$. This function is locally bounded because $a(q)$ and m_q are positive and continuous and $M(q)$ is locally bounded. Combining the two estimates in the triangle-inequality display proves the result for unit vectors, and homogeneity gives the stated result. □

D.5 Proof of Theorem 1

We prove Theorem 1 by specializing Theorem 3 to minimum-ℓ_p-norm interpolation. Throughout this section, q denotes the dual exponent from Theorem 1, and $\|\cdot\| = \|\cdot\|_p$.

Proof. The assumptions of Setting D.1 hold with $\zeta_{i,j} = x_{i,j}/s_N$: these variables are independent, mean-zero, variance one, and satisfy $\|\zeta_{i,j}\|_{\psi_2} \le R_X$.

To establish the claimed monotonicity, we fix $p_0 > 1$ and prove the result uniformly for all $p \ge p_0$. Then the dual exponent satisfies $1 \le q \le p_0/(p_0-1)$. Throughout the proof, $C \ge 1$ and $c > 0$ may change from line to line, with C and c^{-1} having the polynomial dependence stated in the theorem.

Denote the quantity in square brackets in the theorem by

$$\eta := \frac{s_N |N|^{1/q} |S|^{\max\{1/p-1/2,0\}}}{\sqrt{n}} + \left(\frac{r_S + \log(n/\delta)}{n}\right)^{1/4} + \sqrt{\frac{n\log(n|N|/\delta)}{|N|}} + \frac{n\log(n|N|/\delta)}{|N|^{\min\{2/q,1\}}}.$$

After adjusting the constants in the statement, it suffices to prove $\mathcal{R}(\widehat{\mathbf{w}}) \le C\eta$ under the assumption $\eta \le c$.

We first verify the sample-size condition of Theorem 3. The second term in η gives

$$n \ge C\big(r_S + \log(n/\delta)\big).$$

The last term similarly gives

$$|N| \ge C\big(n\log(n|N|/\delta)\big)^{\max\{1,q/2\}}.$$

This lower bound also implies

$$\frac{\big(n\log(n|N|/\delta)\big)^{q/2}}{|N|} \le \frac{n\log(n|N|/\delta)}{|N|^{\min\{2/q,1\}}}, \qquad |N|^{-1/(2q)} \le \left(\frac{r_S + \log(n/\delta)}{n}\right)^{1/4}.$$

For the first inequality, distinguish $q \le 2$ and $q \ge 2$ and use the preceding lower bound on $|N|$; the second follows from the same bound and $r_S \ge 1$.

We next verify Assumption 3. For the ℓ_p norm, the dual norm associated with $\mathbf{w}_N \mapsto \|(\mathbf{0}, \mathbf{w}_N)\|_p$ is $\|\cdot\|_q$. For $G \sim \mathcal{N}(0,1)$, take

$$\kappa_N := s_N\big(|N|\mathbb{E}|G|^q\big)^{1/q}.$$

Lemma 13 verifies the assumption with

$$A = \frac{a}{R_X^4\sqrt{q}}, \qquad \varphi(t) = CR_X|N|^{-1/(2q)} + D(q)R_X^{q+4-2/q}t,$$

where $D(q)$ depends only on q and is bounded uniformly for $1 \le q \le p_0/(p_0-1)$, and $a > 0$ is an absolute constant. In particular, $A^{-1} \le CR_X^4$. The proof of Theorem 3 uses A only through A^{-2}, so its constants have the polynomial dependence asserted in the theorem.

We now verify Assumption 4. Choose

$$\tau := K\left(\sqrt{\frac{n\log(n|N|/\delta)}{|N|}} + \frac{\big(n\log(n|N|/\delta)\big)^{q/2}}{|N|}\right),$$

where K is a sufficiently large polynomial in R_X, with degree and coefficients depending only on p_0. The preceding comparisons give $\tau \le K\eta$, so decreasing c ensures $\tau \in (0,1)$. In Proposition 2, with failure parameter $\delta/12$, we have $L = \sqrt{n} + \sqrt{\log(24|N|/\delta)} \le C\sqrt{n\log(n|N|/\delta)}$. Moreover, the constant $c(q)$ in that proposition is bounded below uniformly for $1 \le q \le p_0/(p_0-1)$. The logarithmic term in its probability bound is at most $Cn\log(n|N|/\delta)$, while

$$\min\Big\{|N|\tau^2, (|N|\tau)^{2/q}\Big\} \ge K^{2/q} n\log(n|N|/\delta).$$

Since $R_X \geq 1$ and $q \leq p_0/(p_0-1)$, we may choose K so that the second term in the probability bound of Proposition 2 is at most $\delta/12$. Thus Assumption 4 holds at failure level $\delta/6$, with

$$\varepsilon_{\mathrm{unif}}(\delta/6) \leq C\left[\sqrt{\frac{n\log(n|N|/\delta)}{|N|}} + \frac{\left(n\log(n|N|/\delta)\right)^{q/2}}{|N|} + |N|^{-1/(2q)}\right] \leq C\eta.$$

It remains to verify the regularity and radius conditions of Theorem 3. Since $\mathbf{w}_N^\star = 0$, we have $\nu_N = 0$, and $\|\mathbf{b}_S\|_2 = b_S^+\sqrt{r_S}$. The definition in (14) therefore gives

$$\rho \leq C\sqrt{\frac{r_S + \log(n/\delta)}{n}}.$$

Using the displayed formula for φ, the uniform bound on $D(q)$, and $R_X \geq 1$, we obtain

$$\varphi(C'\sqrt{\rho}) \leq C\left[|N|^{-1/(2q)} + \left(\frac{r_S + \log(n/\delta)}{n}\right)^{1/4}\right] \leq C\eta.$$

The definition of $\varepsilon_{\mathrm{flat}}$ now yields

$$\varepsilon_{\mathrm{flat}} \leq C\eta.$$

This controls the first term in (16). To control the second, note that $(\mathbb{E}|G|^q)^{1/q}$ is bounded above and below by positive constants uniformly for $1 \leq q \leq p_0/(p_0-1)$. Therefore

$$c\frac{s_N|N|^{1/q}}{\sqrt{n}} \leq \gamma_N \leq C\frac{s_N|N|^{1/q}}{\sqrt{n}}, \qquad \beta_S = |S|^{\max\{1/p-1/2,0\}},$$

and hence $\gamma_N\beta_S/b_S^- \leq C\eta$. Finally, the third term in (16) satisfies

$$\left(\frac{R_X b_S^+ + R_\xi\sigma}{\min\{b_S^-,\sigma\}}\right)^2\sqrt{\frac{r_S+\log(n/\delta)}{n}} \leq C\eta.$$

This verifies (16) when $\eta \leq c$.

It remains only to check the radius condition. Since $\beta_S \geq 1$, the bound $\gamma_N\beta_S \leq C\eta$ also controls γ_N. Combining this with $\eta \leq c$, all the terms involving γ_N in (15) are at most $C\eta$. The bound $\varepsilon_{\mathrm{flat}} \leq C\eta$ controls the first term. Finally, since $\eta \leq c \leq 1$,

$$\sqrt{\frac{r_S+\log(n/\delta)}{n}} \leq \left(\frac{r_S+\log(n/\delta)}{n}\right)^{1/4} \leq \eta$$

controls the remaining term. Thus

$$r_\delta^2 \leq C\eta.$$

Decreasing c once more ensures that $r_\delta \leq 1$. We have now verified all the hypotheses of Theorem 3.

We may therefore apply Theorem 3. For the ℓ_p norm,

$$\beta_{N,\Sigma} = s_N \sup_{\|\mathbf{v}\|_p=1}\|\mathbf{v}\|_2 = s_N|N|^{\max\{1/2-1/p,0\}} = s_N|N|^{\max\{1/q-1/2,0\}}.$$

The regularity bound and $r_\delta \leq 1$ imply $\beta_S r_\delta \leq C/\gamma_N$. Consequently, the risk bound in Theorem 3 gives

$$\mathcal{R}(\widehat{\mathbf{w}}) \leq Cr_\delta^2 + Cs_N^2|N|^{2\max\{1/q-1/2,0\}}\left(\frac{1}{\gamma_N}+1\right)^2.$$

Using $(a+1)^2 \leq 2a^2+2$ and the definition of γ_N, the $1/\gamma_N^2$ contribution equals

$$s_N^2|N|^{2\max\{1/q-1/2,0\}}\frac{1}{\gamma_N^2} = (\mathbb{E}|G|^q)^{-2/q}\frac{n}{|N|^{\min\{2/q,1\}}}.$$

The Gaussian factor is bounded uniformly for $1 \leq q \leq p_0/(p_0-1)$. The contribution without $1/\gamma_N^2$ is no larger than a constant multiple of the same rate. Indeed,

$$2\max\{1/q-1/2,0\} + \min\{2/q,1\} = 2/q$$

and $\beta_S \geq 1$ give

$$\frac{s_N^2 |N|^{2\max\{1/q-1/2,0\}}}{n/|N|^{\min\{2/q,1\}}} \leq \frac{s_N^2 |N|^{2/q} \beta_S^2}{n} \leq C\eta^2.$$

Therefore

$$\mathcal{R}(\widehat{\mathbf{w}}) \leq C \left[\frac{n}{|N|^{\min\{2/q,1\}}} + r_\delta^2 \right].$$

Finally, $r_\delta^2 \leq C\eta$, and

$$\frac{n}{|N|^{\min\{2/q,1\}}} \leq \frac{n \log(n|N|/\delta)}{|N|^{\min\{2/q,1\}}} \leq \eta.$$

Thus $\mathcal{R}(\widehat{\mathbf{w}}) \leq C\eta$, as claimed.

Finally, choosing the constants uniformly over dual exponents in $[1, q]$, a range that shrinks as p increases, makes C non-increasing and the small tolerance c non-decreasing in p. Increasing C if necessary so that $C \geq c^{-1}$ preserves the claimed monotonicity and ensures that $\eta \leq C^{-1}$ implies $\eta \leq c$. □

E Proof of Theorem 2 – The ℓ_1 Failure Example

Proof. Let $S = \{1\}, N = \{2, \ldots, d\}$, and recall that the inputs are given by

$$x_S \sim \mathrm{Unif}\{\pm 1\}, \qquad \text{and} \qquad x_j \sim \mathrm{Unif}\{\pm s_N\} \quad \forall\, j \in N.$$

Let the label noise be independent copies of $\xi := \sigma(Z_n/\rho_n)\varepsilon$, where $\varepsilon \sim \mathrm{Unif}\{\pm 1\}$, Z_n, ρ_n are defined as

$$Z_n = \begin{cases} 1 + \sqrt{\log(n)} & \text{w.p. } \frac{1}{n} \\ 1 & \text{w.p. } 1 - \frac{1}{n} \end{cases}, \qquad \rho_n := \mathbb{E}[Z_n^2]^{1/2},$$

and Z_n and ε are independent.

We first verify Assumptions 1 and 2. By symmetry, $\mathbb{E}\xi = 0$, and by the definition of ρ_n, $\mathbb{E}\xi^2 = \sigma^2$. Moreover,

$$\rho_n^2 = \mathbb{E}Z_n^2 = \frac{\log n + 2\sqrt{\log(n)} + 1}{n} + 1 - \frac{1}{n} = 1 + \frac{\log n + 2\sqrt{\log(n)}}{n}.$$

It is easy to check that $\rho_n \in [1, 3/2]$. Moreover, since $(1 + \sqrt{a})^2 \leq 2(1 + a)$ for every $a \geq 0$, we have

$$\mathbb{E} \exp\left(\frac{\xi^2}{16\sigma^2}\right) \leq \left(1 - \frac{1}{n}\right) e^{1/16} + \frac{1}{n} e^{(1+\log n)/8} \leq e^{1/16} + e^{1/8} 2^{-7/8} \leq 2$$

for all $n \geq 2$. Hence $\|\xi\|_{\psi_2} \leq 4\sigma$. Moreover, all coordinates of $\mathbf{x}$ are scaled Rademacher random variables, so Assumption 1 also holds with a universal constant. Write

$$X_N = s_N R, \qquad R \in \{\pm 1\}^{n \times |N|},$$

and let $\mathcal{E}_{\mathrm{cube}}$ be the event that every vector in $\{\pm 1\}^n$ occurs as a column of R. A union bound gives

$$\Pr(\mathcal{E}_{\mathrm{cube}}^c) \leq 2^n \left(1 - 2^{-n}\right)^{|N|} \leq \exp\left(n \log 2 - \frac{|N|}{2^n}\right).$$

Thus, for $|N| \geq 2n2^n$,

$$\Pr(\mathcal{E}_{\mathrm{cube}}^c) \leq e^{-(2-\log 2)n} \leq e^{-n}.$$

Work on $\mathcal{E}_{\mathrm{cube}}$. Since the columns of R contain all sign vectors, X_N has full row rank. Hence, for every $\mathbf{z} \in \mathbb{R}^n$, ℓ_1-duality (see Lemma 4) gives

$$\min_{\mathbf{w}_N :\, X_N \mathbf{w}_N = \mathbf{z}} \|\mathbf{w}_N\|_1 = \max_{\boldsymbol{\lambda} \in \mathbb{R}^n} \left\{ \mathbf{z}^\top \boldsymbol{\lambda} : \|X_N^\top \boldsymbol{\lambda}\|_\infty \leq 1 \right\}. \tag{59}$$

Moreover, for every $\boldsymbol{\lambda} \in \mathbb{R}^n$, since R contains all sign vectors,

$$\|X_N^\top \boldsymbol{\lambda}\|_\infty = s_N \max_{\mathbf{u} \in \{\pm 1\}^n} |\mathbf{u}^\top \boldsymbol{\lambda}| = s_N \|\boldsymbol{\lambda}\|_1. \tag{60}$$

Combining (59) and (60), and using the duality between ℓ_1 and ℓ_∞, yields

$$\min_{\mathbf{w}_N:\, X_N\mathbf{w}_N=\mathbf{z}} \|\mathbf{w}_N\|_1 = \max_{\|\boldsymbol{\lambda}\|_1 \le s_N^{-1}} \mathbf{z}^\top \boldsymbol{\lambda} = \frac{1}{s_N}\|\mathbf{z}\|_\infty. \tag{61}$$

Since $\mathbf{w}^\star = 0$, the min-norm interpolation problem is

$$\min_{\mathbf{w}:\, X_S w_S + X_N \mathbf{w}_N = \boldsymbol{\xi}} \|\mathbf{w}\|_1 = \min_{w_S \in \mathbb{R}} \min_{\mathbf{w}_N : X_N \mathbf{w}_N = \boldsymbol{\xi} - X_S w_S} \left(|w_S| + \|\mathbf{w}_N\|_1\right),$$

so that by (61), the min-norm interpolator must satisfy

$$\widehat{w}_S \in \operatorname*{argmin}_{w_S \in \mathbb{R}} \left(|w_S| + \frac{1}{s_N}\|\boldsymbol{\xi} - X_S w_S\|_\infty \right).$$

Let

$$H(w) := |w| + \frac{1}{s_N}\|\boldsymbol{\xi} - X_S w\|_\infty, \tag{62}$$

and define

$$\eta_i := x_{i,S}\xi_i, \qquad \eta_- := \min_{i\in[n]} \eta_i, \qquad \eta_+ := \max_{i\in[n]} \eta_i, \qquad M_\eta := \frac{\eta_+ + \eta_-}{2}.$$

Since $x_{i,S} \in \{\pm 1\}$, $|\xi_i - x_{i,S} w| = |x_{i,S}\xi_i - w|$, and therefore,

$$\begin{aligned}
\|\boldsymbol{\xi} - X_S w\|_\infty &= \max_{i\in[n]} |\eta_i - w| =_{(*)} \max\{|\eta_+ - w|, |\eta_- - w|\} \\
&= \max\left\{ \left| \frac{\eta_+ - \eta_-}{2} + M_\eta - w \right|, \left| -\frac{\eta_+ - \eta_-}{2} + M_\eta - w \right| \right\} \\
&=_{(**)} \frac{\eta_+ - \eta_-}{2} + |M_\eta - w|,
\end{aligned} \tag{63}$$

where (*) uses that the maximization problem is one-dimensional and (**) uses the inequality $\max\{|a+b|, |-a+b|\} = a + |b|$ for $a \ge 0$ and $b \in \mathbb{R}$. By (62) and (63),

$$\begin{aligned}
H(w) - H(M_\eta) &= |w| - |M_\eta| + \frac{1}{s_N}|w - M_\eta| \\
&\ge \left(\frac{1}{s_N} - 1\right) |w - M_\eta|.
\end{aligned}$$

Because $s_N < 1$, the right-hand side is strictly positive unless $w = M_\eta$. Thus M_η is the unique minimizer of H, and every minimum-ℓ_1-norm interpolator satisfies

$$\widehat{w}_S = M_\eta = \frac{\max_{i\in[n]}(x_{i,S}\xi_i) + \min_{i\in[n]}(x_{i,S}\xi_i)}{2}. \tag{64}$$

Let $\mathcal{E}_{\text{out}}$ be the event that exactly one of $\xi_1, \ldots, \xi_n$ has magnitude $\sigma(1 + \sqrt{\log n})/\rho_n$. Since each label-noise coordinate has this magnitude with probability $1/n$, and there are n choices for which coordinate has this magnitude,

$$\Pr(\mathcal{E}_{\text{out}}) = n \cdot \frac{1}{n}\left(1 - \frac{1}{n}\right)^{n-1} = \left(1 - \frac{1}{n}\right)^{n-1} \ge \frac{1}{e}.$$

On $\mathcal{E}_{\text{out}}$, exactly one η_i has magnitude $\sigma(1+\sqrt{\log n})/\rho_n$, while all the others have magnitude σ/ρ_n. If the exceptional coordinate is positive, then $\eta_+ = \sigma(1 + \sqrt{\log n})/\rho_n$ and $\eta_- \in \{-\sigma/\rho_n, \sigma/\rho_n\}$; the analogous statement holds with the roles reversed if it is negative. Hence (64) and $\rho_n < 3/2$ give

$$|\widehat{w}_S| \ge \frac{\sigma\sqrt{\log n}}{2\rho_n} \ge \frac{\sigma\sqrt{\log(n)}}{3}. \tag{65}$$

Finally, since $\mathbf{w}^\star = 0$, the excess risk is

$$\mathcal{R}(\widehat{\mathbf{w}}) = \widehat{w}_S^2 + s_N^2 \|\widehat{\mathbf{w}}_N\|_2^2 \ge \widehat{w}_S^2 \ge \sigma^2 \frac{\log n}{9}.$$

Combining the preceding bounds on the probabilities of $\mathcal{E}_{\text{cube}}$ and $\mathcal{E}_{\text{out}}$,

$$\Pr\left(\mathcal{E}_{\text{cube}} \cap \mathcal{E}_{\text{out}}\right) \ge \frac{1}{e} - e^{-n} \ge \frac{1}{5}.$$

Together with the preceding risk bound, this proves the theorem. □