UNIVERSALITY LAWS FOR RANDOMIZED DIMENSION REDUCTION, WITH APPLICATIONS

SAMET OYMAK AND JOEL A. TROPP

ABSTRACT. Dimension reduction is the process of embedding high-dimensional data into a lower dimensional space to facilitate its analysis. In the Euclidean setting, one fundamental technique for dimension reduction is to apply a random linear map to the data. This dimension reduction procedure succeeds when it preserves certain geometric features of the set. The question is how large the embedding dimension must be to ensure that randomized dimension reduction succeeds with high probability.

This paper studies a natural family of randomized dimension reduction maps and a large class of data sets. It proves that there is a phase transition in the success probability of the dimension reduction map as the embedding dimension increases. For a given data set, the location of the phase transition is the same for all maps in this family. Furthermore, each map has the same stability properties, as quantified through the restricted minimum singular value. These results can be viewed as new universality laws in high-dimensional stochastic geometry.

Universality laws for randomized dimension reduction have many applications in applied mathematics, signal processing, and statistics. They yield design principles for numerical linear algebra algorithms, for compressed sensing measurement ensembles, and for random linear codes. Furthermore, these results have implications for the performance of statistical estimation methods under a large class of random experimental designs.

Date: 30 November 2015. Revised 7 August 2017 and 5 September 2017.

2010 *Mathematics Subject Classification.* Primary: 60D05, 60F17. Secondary: 60B20.

Key words and phrases. Conic geometry, convex geometry, dimension reduction, invariance principle, limit theorem, random code, random matrix, randomized numerical linear algebra, signal reconstruction, statistical estimation, stochastic geometry, universality.

CONTENTS

List of Figures

1. Overview of the Universality Phenomenon

This paper concerns a fundamental question in high-dimensional stochastic geometry:

(Q1) Is it likely that a random subspace of fixed dimension does *not* intersect a given set?

This problem has its roots in the earliest research on spherical integral geometry [San52, San76], and it also arises in asymptotic convex geometry [Gor88]. In recent years, this question has attracted fresh attention [Don06c, RV08, Sto09, DT09b, CRPW12, ALMT14, Sto13, MT13, OH13, OTH13, TOH15, TAH15] because it is central to the analysis of randomized dimension reduction.

This paper establishes that a striking universality phenomenon takes place in the stochastic geometry problem (Q1). For a given set, the answer to this question is essentially the same for every distribution on random subspaces that is induced by a natural model for random linear maps. Universality also manifests itself in metric variants of (Q1), where we ask how far the random subspace lies from the set. We discuss the implications of these results in high-dimensional geometry, random matrix theory, numerical analysis, optimization, statistics, signal processing, and beyond.

1.1. Randomized Dimension Reduction.

Dimension reduction is the operation of mapping a set from a large space into a smaller space. Ideally, this action distills the "information" in the set, and it allows us to develop more efficient algorithms for processing that information. In the setting of Euclidean spaces, a fundamental method for dimension reduction is to apply a random linear map to each point in the set. It is important that the random linear map preserve geometric features of the set. In particular, we do *not* want the linear map to map a point in the set to the origin. Equivalently, the null space of the random linear map should *not* intersect the set. We see that (Q1) emerges naturally in the context of randomized dimension reduction.

1.2. Technical Setting.

Let us introduce a framework in which to study this problem. It is natural to treat (Q1) as a question in spherical geometry because it is scale invariant. Fix the *ambient dimension* D, and consider a closed subset Ω of the Euclidean unit sphere in $\mathbb{R}^D$. For the moment, we also assume that Ω is *spherically convex*; that is, Ω is the intersection of a convex cone[1] with the unit sphere. Construct a random linear map $\mathbf{\Pi} : \mathbb{R}^D \to \mathbb{R}^d$, where the *embedding dimension* d does not exceed the ambient dimension D. As we vary the distribution of the random linear map $\mathbf{\Pi}$, the map $\mathbf{\Pi} \mapsto \operatorname{null}(\mathbf{\Pi})$ induces different distributions on the subspaces in $\mathbb{R}^D$ with codimension at most d. We may now reformulate (Q1) in this language:

(Q2) For a given embedding dimension d, what is the probability that $\Omega \cap \operatorname{null}(\mathbf{\Pi}) = \emptyset$? Equivalently, what is the probability that $\mathbf{0} \notin \mathbf{\Pi}(\Omega)$?

We say that the random projection *succeeds* when $\mathbf{0} \notin \mathbf{\Pi}(\Omega)$. Conversely, when $\mathbf{0} \in \mathbf{\Pi}(\Omega)$, we say that the random projection *fails*. See Figure 1.1 for an illustration.

We have the intuition that, for a fixed choice of Ω, the projection is more likely to succeed as the embedding dimension d increases. Furthermore, a random linear map $\mathbf{\Pi}$ with fixed embedding dimension d is less likely to succeed as the size of the set increases. We will justify these heuristics in complete detail.

1.3. A Phase Transition for Uniformly Random Partial Isometries.

We begin with a case where the literature already contains a comprehensive answer to the question (Q2).

The most natural type of random embedding is a *uniformly random partial isometry*. That is, $\mathbf{\Pi} : \mathbb{R}^D \to \mathbb{R}^d$ is a partial isometry[2] whose null space, $\operatorname{null}(\mathbf{\Pi})$, is drawn uniformly at random from the Haar measure on the Grassmann manifold of subspaces in $\mathbb{R}^D$ with codimension d. The invariance properties of the distribution of $\mathbf{\Pi}$ allow for a complete analysis of its action on Ω, the spherically convex set [SW08, Chap. 6.5]. Recent research [ALMT14, MT14a, GNP14] has shown how to convert the complicated exact formulas into interpretable results.

[1] A *convex cone* is a convex set K that satisfies $\alpha K = K$ for all $\alpha > 0$.

[2] A *partial isometry* $\mathbf{\Pi}$ satisfies the condition $\mathbf{\Pi}\mathbf{\Pi}^* = \mathbf{\Pi}$, where * is the transpose operation and $\mathbf{I}$ is the identity map.

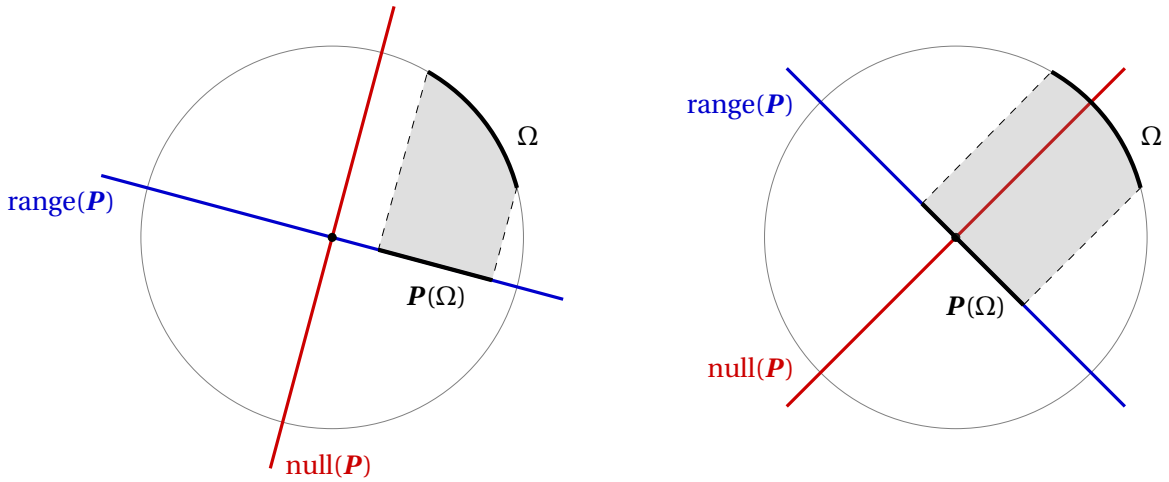

FIGURE 1.1: *Geometry of a Random Linear Map.* We can identify a linear map $\mathbf{\Pi} : \mathbb{R}^D \to \mathbb{R}^d$ with an orthogonal projector $\boldsymbol{P} : \mathbb{R}^D \to \mathbb{R}^D$ whose range is the orthogonal complement of null($\mathbf{\Pi}$). In this diagram, $\boldsymbol{P} : \mathbb{R}^2 \to \mathbb{R}^2$ is a random orthogonal projector applied to a closed, spherically convex set Ω. The likelihood of a given configuration depends on the statistical dimension $\delta(\Omega)$ of the set. LEFT: *Success Regime.* The null space, null($\boldsymbol{P}$), does not intersect Ω, and the image $\boldsymbol{P}(\Omega)$ does not contain the origin. RIGHT: *Failure Regime.* The null space, null($\boldsymbol{P}$), intersects the set Ω, and the image $\boldsymbol{P}(\Omega)$ contains the origin.

The modern theory is expressed in terms of a geometric functional $\delta(\Omega)$, called the *statistical dimension*:

$$\delta(\Omega) := \mathbb{E}\left[\left(\max_{\boldsymbol{t}\in\Omega} \boldsymbol{g}\cdot\boldsymbol{t}\right)_+^2\right], \quad \text{where } \boldsymbol{g} \text{ is NORMAL}(\mathbf{0}, \mathbf{I}).$$

The statistical dimension is increasing with respect to set inclusion, and its values range from zero (for the empty set) up to D (for the whole sphere). Furthermore, the functional can be computed accurately in many cases of interest. See Section 3.3 for more details.

The statistical dimension demarcates a phase transition in the behavior of a uniformly random partial isometry $\mathbf{\Pi}$ as the embedding dimension d varies. For a closed, spherically convex set Ω, the results [ALMT14, Thm. I and Prop. 10.2] demonstrate that

$$\begin{aligned} d \le \delta(\Omega) - \mathrm{C}\sqrt{\delta(\Omega)} \quad &\text{implies} \quad \mathbf{0} \in \mathbf{\Pi}(\Omega) \quad \text{with high probability;} \\ d \ge \delta(\Omega) + \mathrm{C}\sqrt{\delta(\Omega)} \quad &\text{implies} \quad \mathbf{0} \notin \mathbf{\Pi}(\Omega) \quad \text{with high probability.} \end{aligned} \tag{1.1}$$

The number C is a positive universal constant. In other terms, a uniformly random projection $\mathbf{\Pi}(\Omega)$ of a spherically convex set Ω is likely to succeed precisely when the embedding dimension d is larger than the statistical dimension $\delta(\Omega)$. See Figure 1.2 for a plot of the exact probability that a uniformly random partial isometry annihilates a point in a specific set Ω.

Remark 1.1 (Related Work)**.** The results [ALMT14, Thm. 7.1] and [MT13, Thm. A] contain good bounds for the probabilities in (1.1). The probabilities can be approximated more precisely by introducing a second geometric functional [GNP14]. These estimates depend on the spherical Crofton formula [SW08, Eqns. (6.62), (6.63)], which gives the *exact* probabilities in a less interpretable form. Related phase transition results can also be obtained via the Gaussian Minimax Theorem; see [Gor88, Cor. 3.4], [ALMT14, Rem. 2.9], [Sto13], or [TOH15, Thm. II.1]. See [TH15] for other results on uniformly random partial isometries.

1.4. **Other Types of Random Linear Maps?** The research outlined in Section 1.3 delivers a complete account of how a uniformly random partial isometry behaves in the presence of some convexity. In

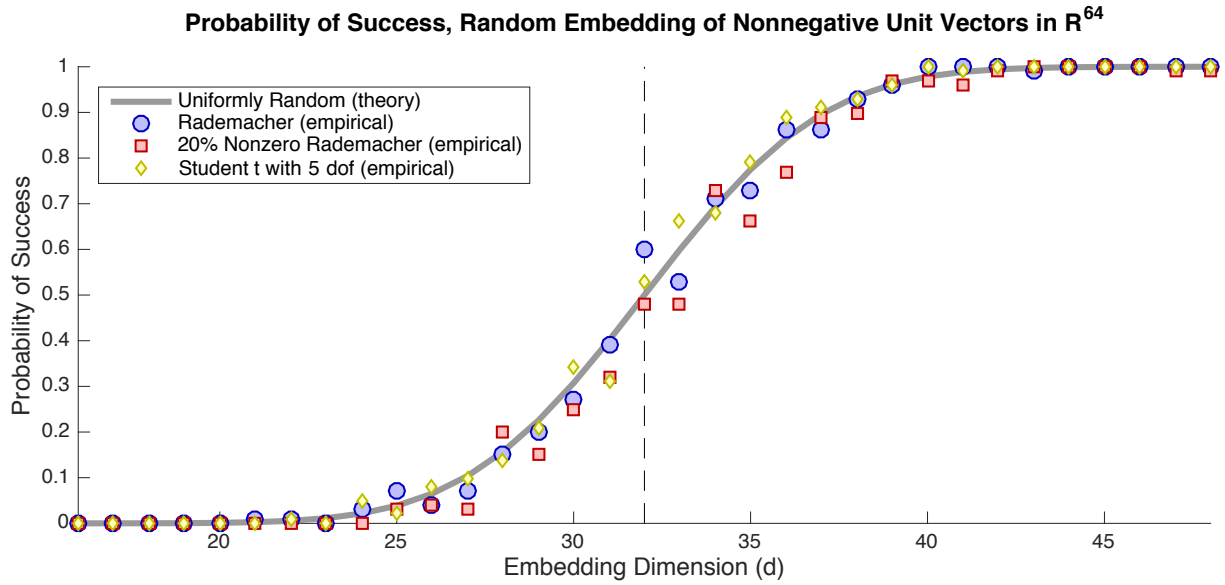

FIGURE 1.2: *Universality of the Embedding Dimension.* This plot describes the behavior of four types of random linear maps applied to the set Ω of unit vectors with nonnegative coordinates: $\Omega := \{\boldsymbol{t} \in \mathbb{R}^{64} : \|\boldsymbol{t}\| = 1, t_i \geq 0\}$. The **dashed line** marks the statistical dimension of the set: $\delta(\Omega) = 32$. The **gray curve** interpolates the exact probability that a uniformly random partial isometry $\boldsymbol{\Pi} : \mathbb{R}^{64} \to \mathbb{R}^d$ succeeds (i.e., $\mathbf{0} \notin \boldsymbol{\Pi}(\Omega)$) as a function of the embedding dimension d. The **markers** indicate the empirical probability (over 100 trials) that dimension reduction succeeds for a random linear map with the specified distribution. See Sections 1.4 and 1.9 for further details.

contrast, the literature contains almost no precise information about the behavior of other random linear maps.

Nevertheless, in applications, we may prefer—or be forced—to work with other types of random linear maps. Here is a motivating example. Many algorithms for numerical linear algebra now depend on randomized dimension reduction. In this context, uniformly random partial isometries are expensive to construct, to store, and to perform arithmetic with. It is more appealing to implement a random linear map that is discrete, or sparse, or structured. The lack of detailed theoretical information about how these linear maps behave makes it difficult to design numerical methods with guaranteed performance.

We can, however, use computation to investigate the behavior of other types of random linear maps. Figure 1.2 presents the results of the following experiment. Consider the set Ω of unit vectors in $\mathbb{R}^{64}$ with nonnegative coordinates:

$$\Omega := \big\{\boldsymbol{t} \in \mathbb{R}^{64} : \|\boldsymbol{t}\| = 1 \text{ and } t_i \geq 0 \text{ for } i = 1, \dots, 64\big\}.$$

According to (3.6), below, the statistical dimension $\delta(\Omega) = 32$, so the formula (1.1) tells us to expect a phase transition in the behavior of a uniformly random partial isometry when the embedding dimension $d = 32$. Using [ALMT14, Ex. 5.3 and Eqn. (5.10)], we can compute the exact probability that a uniformly random partial isometry $\boldsymbol{\Pi} : \mathbb{R}^{64} \to \mathbb{R}^d$ succeeds as a function of d. Against this baseline, we compare the empirical probability (over 100 trials) that a random linear map with independent Rademacher[3] entries succeeds. We also display experiments for a 20% nonzero Rademacher linear map and for a linear map with Student t_5 entries. See Section 1.9 for more details.

From this experiment, we discover that all three linear maps act almost exactly the same way as a uniformly random partial isometry! This universality phenomenon is remarkable because the four

[3] A *Rademacher* random variable takes the two values ± 1 with equal probability.

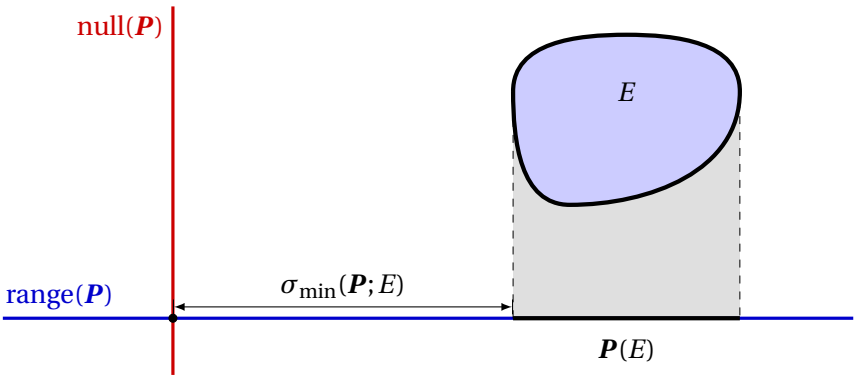

FIGURE 1.3: *Geometry of the Restricted Minimum Singular Value.* We can identify a partial unitary linear map $\mathbf{\Pi} : \mathbb{R}^D \to \mathbb{R}^d$ with an orthogonal projector $\boldsymbol{P} : \mathbb{R}^D \to \mathbb{R}^D$ whose range is the orthogonal complement of null($\mathbf{\Pi}$). In this diagram, $\boldsymbol{P} : \mathbb{R}^2 \to \mathbb{R}^2$ is an orthogonal projector applied to a compact convex set E. The restricted minimum singular value $\sigma_{\min}(\boldsymbol{P}; E)$ is the distance from the origin to the image $\boldsymbol{P}(E)$.

linear maps have rather different distributions. At present, the literature contains no information about when—or why—this phenomenon occurs.

1.5. **A Universality Law for the Embedding Dimension.** The central goal of this paper is to show that there is a substantial class of random linear maps for which the phase transition in the embedding dimension is universal. Here is a rough statement of the main result.

Let Ω be a closed, spherically convex set in $\mathbb{R}^D$. Suppose that the entries of the matrix of the random linear map $\mathbf{\Pi} : \mathbb{R}^D \to \mathbb{R}^d$ are independent, standardized,[4] and symmetric,[5] with a modest amount of regularity.[6] In particular, we may consider random linear maps that have an arbitrarily small, but constant, proportion of nonzero entries. For this class of random linear maps, we will demonstrate that

$$\begin{aligned} d \le \delta(\Omega) - o(D) \quad &\text{implies} \quad \mathbf{0} \in \mathbf{\Pi}(\Omega) \quad \text{with high probability;} \\ d \ge \delta(\Omega) + o(D) \quad &\text{implies} \quad \mathbf{0} \notin \mathbf{\Pi}(\Omega) \quad \text{with high probability.} \end{aligned} \tag{1.2}$$

The little-o notation suppresses constants that depend only on the regularity of the random variables that populate $\mathbf{\Pi}$. See Theorem I in Section 3.4 for a more complete statement.

The result (1.2) states that a random linear map $\mathbf{\Pi}$ is likely to succeed for a spherically convex set Ω precisely when the embedding dimension d exceeds the statistical dimension $\delta(\Omega)$ of the set. We learn that the phase transition in the embedding dimension is universal over our class of random linear maps, provided that Ω is not too small as compared with the ambient dimension D. Note that a random linear map $\mathbf{\Gamma} : \mathbb{R}^D \to \mathbb{R}^d$ with standard normal entries has the same behavior as a $d \times D$ uniformly random partial isometry because, almost surely, the null space of $\mathbf{\Gamma}$ is a uniformly random subspace of $\mathbb{R}^D$ with codimension d. This analysis explains the dominant features of the experiment in Figure 1.2!

1.6. **A Universality Law for the Restricted Minimum Singular Value.** It is also a matter of significant interest to understand the *stability* of randomized dimension reduction. We quantify the stability of the random linear map $\mathbf{\Pi} : \mathbb{R}^D \to \mathbb{R}^d$ on a compact, convex set E in $\mathbb{R}^D$ using the *restricted minimum singular value*:

$$\sigma_{\min}(\mathbf{\Pi}; E) := \min_{\boldsymbol{t} \in E} \|\mathbf{\Pi}\boldsymbol{t}\| .$$

[4] A *standardized* random variable has mean zero and variance one.

[5] A *symmetric* random variable X has the same distribution as its negation $-X$.

[6] For concreteness, we may assume that the entries of $\mathbf{\Pi}$ have five uniformly bounded moments.

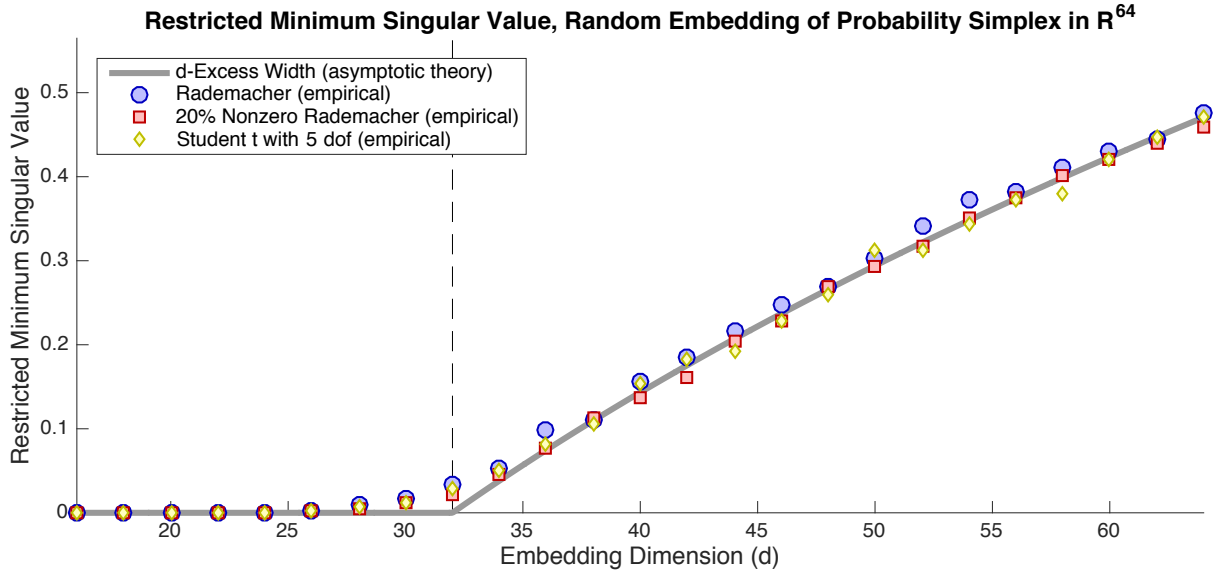

FIGURE 1.4: *Universality of the Restricted Minimum Singular Value.* This plot describes the behavior of three types of random linear maps applied to the probability simplex $\Delta_{64} := \{\boldsymbol{t} \in \mathbb{R}^{64} : \sum_i t_i = 1, t_i \geq 0\}$. The **dashed line** marks the minimum dimension $d = 32$ where uniformly random embedding of the set Δ_{64} is likely to succeed. The **gray curve** interpolates the value of the positive part $\big(\mathscr{E}_d(\Delta_{64})\big)_+$ of the d-excess width of Δ_{64}, obtained from the asymptotic calculation (4.4). The **markers** give an empirical estimate (over 100 trials) for the restricted minimum singular value $\sigma_{\min}(\boldsymbol{\Pi}; \Delta_{64})$ of a random linear map $\boldsymbol{\Pi} : \mathbb{R}^{64} \to \mathbb{R}^d$ drawn from the specified distribution. See Section 1.9 for more information.

When the restricted minimum singular value $\sigma_{\min}(\boldsymbol{\Pi}; E)$ is large, the random image $\boldsymbol{\Pi}(E)$ is far from the origin, so the embedding is very stable. That is, we can deform either the linear map $\boldsymbol{\Pi}$ or the set E and still be sure that the embedding succeeds. When the restricted minimum singular value is small, the random dimension reduction map is unstable. When it is zero, the random dimension reduction map fails. See Figure 1.3 for a diagram.

Our second major theorem is a universality law for the restricted minimum singular values of a random linear map. This result is expressed using a geometric functional $\mathscr{E}_d(E)$, called the *d-excess width* of E:

$$\mathscr{E}_d(E) := \mathbb{E} \min_{\boldsymbol{t} \in E} \big(\sqrt{d}\, \|\boldsymbol{t}\| + \boldsymbol{g} \cdot \boldsymbol{t} \big), \quad \text{where } \boldsymbol{g} \text{ is } \textsc{normal}(\mathbf{0}, \mathbf{I}).$$

The d-excess width increases with the parameter d, and it decreases with respect to set inclusion. The typical scale of $\mathscr{E}_d(E)$ is $O(\sqrt{d})$. In addition, the excess width can be evaluated precisely in many situations of interest. See Section 4.2 for more details.

Now, suppose that the entries of the matrix of $\boldsymbol{\Pi} : \mathbb{R}^D \to \mathbb{R}^d$ are independent, standardized, and symmetric, with a modest amount of regularity. For a compact, convex subset E of the unit ball in $\mathbb{R}^D$, we will establish that

$$\sigma_{\min}(\boldsymbol{\Pi}; E) = \big(\mathscr{E}_d(E)\big)_+ + o(\sqrt{d}) \quad \text{with high probability.} \tag{1.3}$$

The little-o notation suppresses constants that depend only on the regularity of the random variables. Theorem II in Section 4.3 contains a more detailed statement of (1.3).

In summary, provided that the set E is not too small, the restricted minimum singular value $\sigma_{\min}(\boldsymbol{\Pi}; E)$ depends primarily on the geometry of the set E and the embedding dimension d, rather than on the distribution of the random linear map $\boldsymbol{\Pi}$. See Figure 1.4 for a numerical illustration of this fact.

1.7. **Applications of Universality.** Randomized dimension reduction and, more generally, random matrices have become ubiquitous in the information sciences. As a consequence, the universality laws that we outlined in Section 1.5 and 1.6 have a wide range of implications.

Signal Processing: The main idea in the field of compressed sensing is that we can acquire information about a structured signal by taking random linear measurements. The literature contains extensive empirical evidence that many types of random measurements behave in an indistinguishable fashion. Our work gives the first explanation of this phenomenon. (Section 5)

Stochastic Geometry: Our results also indicate that the facial structure of the convex hull of independent random vectors, drawn from an appropriate class, does not depend heavily on the distribution. (Section 5.3.1)

Coding Theory: Random linear codes provide an efficient way to protect transmissions against error. We demonstrate that a class of random codes is resilient against sparse corruptions. The number of errors that can be corrected does not depend on the specific choice of codebook. (Section 6)

Numerical Analysis: Our research provides an engineering design principle for numerical algorithms based on randomized dimension reduction. We can select the random linear map that is most favorable for implementation and still be confident about the detailed behavior of the algorithm. This approach allows us to develop efficient numerical methods that also have rigorous performance guarantees. (Section 7)

Random Matrix Theory: Our work leads to a new proof of the Bai–Yin law for the minimum singular value of a random matrix with independent entries. (Section 7.1)

High-Dimensional Statistics: The LASSO is a widely used method for performing regression and variable selection. We demonstrate that the prediction error associated with a LASSO estimator is universal across a large class of random designs and statistical error models. We also show that least-absolute-deviation (LAD) regression can correct a small number of arbitrary statistical errors for a wide class of random designs. (Section 8 and Remark 6.2)

Neuroscience: Our universality laws may even have broader scientific significance. It has been conjectured, with some experimental evidence, that the brain may use dimension reduction to compress information [GG15]. Our universality laws suggest that many types of uncoordinated (i.e., random) activity lead to dimension reduction methods with the same behavior. This result indicates that the hypothesis of neural dimensionality reduction may be biologically plausible.

1.8. **The Scope of the Universality Phenomenon.** The universality phenomenon developed in this paper extends beyond the results that we establish, but there are some (apparently) related problems where universality does not hold. Let us say a few words about these examples and non-examples.

First, it does not appear important that the random linear map $\mathbf{\Pi}$ has independent entries. There is extensive evidence that structured random linear maps also have some universality properties; for example, see [DT09a].

Second, the restricted minimum singular value is not the only type of functional where universality is visible. For instance, suppose that f is a convex, Lipschitz function. Consider the quantity

$$\min_{\boldsymbol{t}\in E} \big(\|\mathbf{\Pi}\boldsymbol{t}\|^2 + f(\boldsymbol{t}) \big).$$

Optimization problems of this form play a central role in contemporary statistics and machine learning. It is likely that the value of this optimization problem is universal over a wide class of random linear maps. Furthermore, we believe that our techniques can be adapted to address this question. On the other hand, geometric functionals involving non-Euclidean norms need not exhibit universality. Consider the ℓ_1 *restricted minimum singular value*

$$\min_{\boldsymbol{t}\in E} \|\mathbf{\Pi}\boldsymbol{t}\|_{\ell_1}. \tag{1.4}$$

There are nontrivial sets E where the value of the optimization problem (1.4) varies a lot with the choice of the random linear map $\mathbf{\Pi}$. For instance, let $\mathbf{e}_1 \in \mathbb{R}^D$ be the first standard basis vector, and define the shifted Euclidean ball

$$E_\alpha := \big\{ \boldsymbol{t} \in \mathbb{R}^D : \|\boldsymbol{t} - \mathbf{e}_1\| \le \alpha \big\} \quad \text{for } 0 \le \alpha \le 1.$$

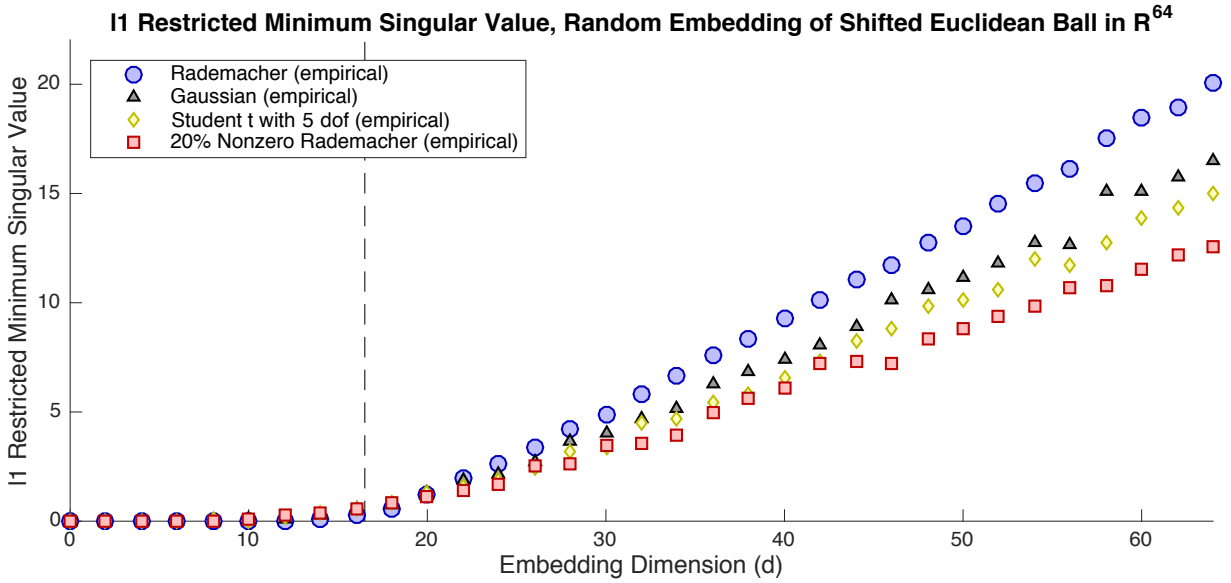

FIGURE 1.5: *Non-Universality of the ℓ_1 Restricted Minimum Singular Value.* This plot describes the behavior of four types of random linear maps applied to the set $E_{1/2} := \{\boldsymbol{t} \in \mathbb{R}^{64} : \|\boldsymbol{t} - \mathbf{e}_1\| \le 1/2\}$, where $\mathbf{e}_1$ is the first standard basis vector. The **dashed line** stands at the phase transition $d \approx 16$ for the embedding dimension of the set $E_{1/2}$. The **markers** give an empirical estimate (over 100 trials) for the ℓ_1 restricted minimum singular value $\min_{\boldsymbol{t} \in E_{1/2}} \|\boldsymbol{\Pi}\boldsymbol{t}\|_{\ell_1}$ of a random linear map $\boldsymbol{\Pi} : \mathbb{R}^{64} \to \mathbb{R}^d$ with the specified distribution as a function of the embedding dimension d. See Section 1.9 for more details.

Using Theorem I and the calculation [ALMT14, Sec. 3.4] of the statistical dimension of a circular cone, we can verify that there is a universal phase transition for successful embedding of the set E_α when the embedding dimension $d = \alpha^2 D + O(1)$. The result (1.3) implies that the minimum restricted singular value of E_α also takes a universal value. At the same time, Figure 1.5 illustrates that the functional (1.4) is not universal for the set E_α.

Finally, functionals involving maximization do not necessarily exhibit universality. The *restricted maximum singular value* is defined as

$$\sigma_{\max}(\boldsymbol{\Pi}; E) := \max_{\boldsymbol{t} \in E} \|\boldsymbol{\Pi}\boldsymbol{t}\|.$$

It is not hard to produce examples where the restricted maximum singular value depends on the choice of the random matrix $\boldsymbol{\Pi}$. For instance, Figure 1.6 demonstrates that the random linear map $\boldsymbol{\Pi}$ has a substantial impact on the maximum singular value $\sigma_{\max}(\boldsymbol{\Pi}; \Delta_D)$ restricted to the probability simplex Δ_D in $\mathbb{R}^D$. This observation may surprise researchers in random matrix theory because the ordinary maximum singular value is universal over the class of random matrices with independent entries [BS10, Thm. 3.10].

1.9. **Reproducible Research.** This paper is accompanied by `MATLAB` code [Tro15a] that reproduces each figure from stored data. This software can also repeat the numerical experiments to obtain new instances of each figure. By modifying the parameters in the code, the reader may explore how changes affect the universality phenomenon. We omit meticulous descriptions of the numerical experiments from the text because these recitations are tiresome for the reader and the code offers superior documentation.

1.10. **Roadmap.** This paper is divided into five parts. Part I offers a complete presentation of our universality laws, some comments about the proofs, and some prospects for further research. Part II outlines the applications of universality in several disciplines, and it contains more empirical confirmation of our analysis. Part III presents the proof that the restricted minimum singular value exhibits universal

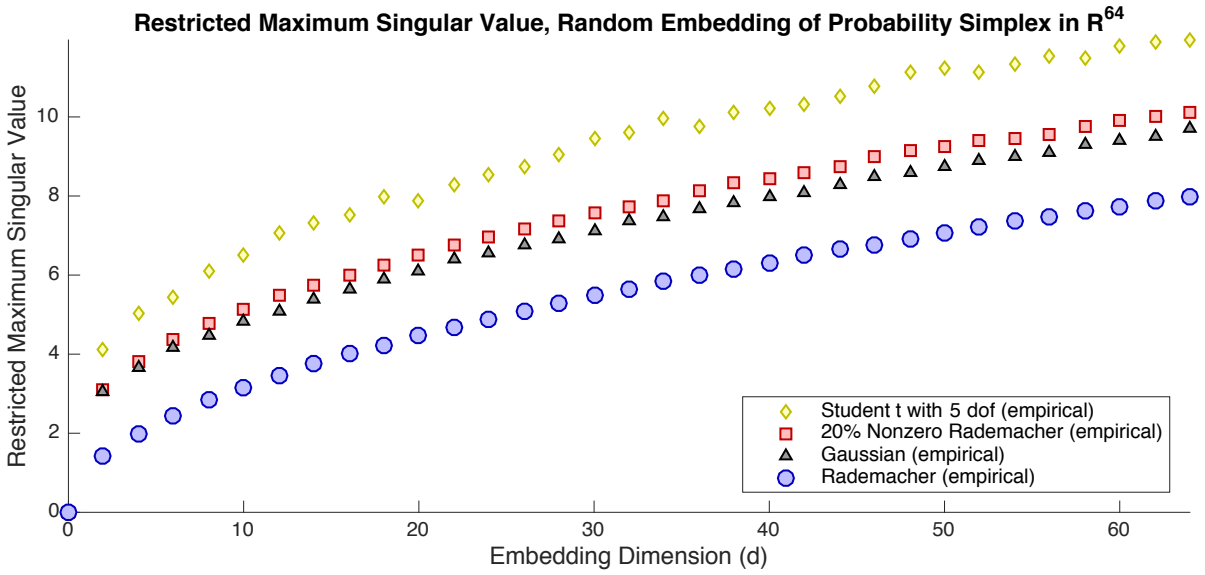

FIGURE 1.6: *Non-Universality of the Restricted Maximum Singular Value.* This plot describes the behavior of four types of random linear maps applied to the probability simplex $\Delta_{64} := \{\boldsymbol{t} \in \mathbb{R}^{64} : \sum_i t_i = 1, t_i \geq 0\}$. The **markers** give an empirical estimate (over 100 trials) for the restricted maximum singular value $\sigma_{\max}(\boldsymbol{\Pi}; \Delta_{64})$ of a random linear map $\boldsymbol{\Pi} : \mathbb{R}^{64} \to \mathbb{R}^d$ with the specified distribution. See Section 1.9 for more details.

behavior; this argument also yields the condition in which randomized embedding is likely to succeed. Part IV contains the proof of the condition in which randomized embedding is likely to fail. Finally, Part V includes background results, the acknowledgments, and the list of works cited.

1.11. **Notation.** Let us summarize our notation. We use italic lowercase letters (for example, a) for scalars, boldface lowercase letters ($\boldsymbol{a}$) for vectors, and boldface uppercase letters ($\boldsymbol{A}$) for matrices. Uppercase italic letters (A) may denote scalars, sets, or random variables, depending on the context. Roman letters (c, C) denote universal constants that may change from appearance to appearance. We sometimes delineate specific constant values with subscripts ($\mathrm{C}_{\mathrm{rad}}$).

Given a vector $\boldsymbol{a}$ and a set J of indices, we write $\boldsymbol{a}_J$ for the vector restricted to those indices. In particular, a_j is the jth coordinate of the vector. Given a matrix $\boldsymbol{A}$ and sets I and J of row and column indices, we write $\boldsymbol{A}_{IJ}$ for the submatrix indexed by I and J. In particular, a_{ij} is the component in the (i, j) position of $\boldsymbol{A}$. If there is a single index $\boldsymbol{A}_J$, it always refers to the *column* submatrix indexed by J.

We always work in a real Euclidean space. The symbol B^n is the unit ball in $\mathbb{R}^n$, and S^{n-1} is the unit sphere in $\mathbb{R}^n$. The unadorned norm $\|\cdot\|$ refers to the ℓ_2 norm of a vector or the ℓ_2 operator norm of a matrix. We use the notation $\boldsymbol{s} \cdot \boldsymbol{t}$ for the standard inner product of vectors $\boldsymbol{s}$ and $\boldsymbol{t}$ with the same length. We write * for the transpose of a vector or a matrix.

For a real number a, we define the positive-part and negative-part functions:

$$(a)_+ := \max\{a, 0\} \quad \text{and} \quad (a)_- := \max\{-a, 0\}.$$

These functions bind before powers, so $(a)_+^2$ is the square of the positive part of a.

The symbols $\mathbb{E}$ and Var refer to the expectation and variance of a random variable, and $\mathbb{P}\{\cdot\}$ returns the probability of an event. We use the convention that powers bind before the expectation, so $\mathbb{E}\, X^2$ returns the expectation of the square. We write $\mathbb{1}A$ for the 0–1 indicator random variable of the event A.

A *standardized* random variable has mean zero and variance one. A *symmetric* random variable X has the same distribution as its negation $-X$. We reserve the letter γ for a standard normal random variable; the boldface letters $\boldsymbol{\gamma}, \boldsymbol{g}, \boldsymbol{h}$ are always standard normal vectors; and $\boldsymbol{\Gamma}$ is a standard normal matrix. The dimensions are determined by context.

Part I. Main Results

This part of the paper introduces two new universality laws, one for the phase transition in the embedding dimension and a second one for the restricted minimum singular values of a random linear map. We also include some high-level remarks about the proofs, but we postpone the details to Parts III and IV.

In Section 2, we introduce two models for random linear maps that we use throughout the paper. Section 3 presents the universality result for the embedding dimension, and Section 4 presents the result for restricted singular values.

2. Random Matrix Models

To begin, we present two models for random linear maps that arise in our study of universality. One model includes bounded random matrices with independent entries, while the second allows random matrices with heavy-tailed entries.

2.1. Bounded Random Matrix Model.

Our first model contains matrices whose entries are uniformly bounded. This model is useful for some applications, and it plays a central role in the proofs of our universality results.

Model 2.1 (Bounded Matrix Model)**.** Fix a parameter $B \geq 1$. A random matrix in this model has the following properties:

- **Independence.** The entries are stochastically independent random variables.
- **Standardization.** Each entry has mean zero and variance one.
- **Symmetry.** Each entry has a symmetric distribution.
- **Boundedness.** Each entry X of the matrix is uniformly bounded: $|X| \leq B$.

Identical distribution of entries is not required. In some cases, which we will note, the symmetry requirement can be dropped.

This model includes several types of random matrices that appear frequently in practice.

Example 2.2 (Rademacher Matrices)**.** Consider a random matrix whose entries are independent, Rademacher random variables. This type of random matrix meets the requirements of Model 2.1 with $B = 1$. Rademacher matrices provide the simplest example of a random linear map. They are appealing for many applications because they are discrete.

Example 2.3 (Sparse Rademacher Matrices)**.** Let $\alpha \in (0, 1]$ be a thinning parameter. Consider a random variable X with distribution

$$X = \begin{cases} +\alpha^{-1/2}, & \text{with probability } \alpha/2; \\ -\alpha^{-1/2}, & \text{with probability } \alpha/2; \\ 0, & \text{otherwise.} \end{cases}$$

A random matrix whose entries are independent copies of X satisfies Model 2.1 with $B = \alpha^{-1/2}$. These random matrices are useful because we can control the sparsity.

2.2. Heavy-Tailed Random Matrix Model.

Next, we introduce a more general class of random matrices that includes heavy-tailed examples. Our main results concern random linear maps from this model.

Model 2.4 (p-Moment Model)**.** Fix parameters $p > 4$ and $\nu \geq 1$. A random matrix in this model has the following properties:

- **Independence.** The entries are stochastically independent random variables.
- **Standardization.** Each entry has mean zero and variance one.
- **Symmetry.** Each entry has a symmetric distribution.
- **Bounded Moments.** Each entry X has a uniformly bounded pth moment: $\mathbb{E}|X|^p \leq \nu^p$.

Identical distribution of entries is not required.

Model 2.4 subsumes Model 2.1, but it also encompasses many other types of random matrices.

Example 2.5 (Gaussian Matrices)**.** Consider an $m \times n$ random matrix $\mathbf{\Gamma}$ whose entries are independent, standard normal random variables. The matrix $\mathbf{\Gamma}$ satisfies the requirements of Model 2.4 for each $p > 4$ with $\nu \le \sqrt{p}$.

In some contexts, we can use a Gaussian random matrix to study the behavior of a uniformly random partial isometry. Indeed, the null space, $\operatorname{null}(\mathbf{\Gamma})$, of the standard normal matrix is a uniformly random subspace of $\mathbb{R}^n$ with codimension $\min\{m, n\}$, almost surely.

Model 2.4 contains several well-studied classes of random matrices.

Example 2.6 (Subgaussian Matrices)**.** Suppose that the entries of the random matrix are independent, and each entry X is symmetric, standardized, and uniformly subgaussian. That is, there is a parameter $\alpha > 0$ where

$$\left(\mathbb{E}|X|^p\right)^{1/p} \le \alpha\sqrt{p} \quad \text{for all } p \ge 1.$$

These matrices are included in Model 2.4 for each $p > 4$ with $\nu \le \alpha\sqrt{p}$. Rademacher, sparse Rademacher, and Gaussian matrices fall in this category.

Example 2.7 (Log-Concave Entries)**.** Suppose that the entries of the random matrix are independent, and each entry X is a symmetric, standardized, log-concave random variable. Recall that a real log-concave random variable X has a density f of the form

$$f(x) := \frac{1}{Z}\mathrm{e}^{-h(x)}$$

where $h : \mathbb{R} \to \mathbb{R}$ is convex and Z is a normalizing constant. It can be shown [BGVV14, Thm. 2.4.6] that these matrices are included in Model 2.4 for any $p > 4$.

In contrast with most research on randomized dimension reduction, we allow the random linear map to have entries with heavy tails. Here is one such example.

Example 2.8 (Student t Matrices)**.** Suppose that each entry of the random matrix is an independent Student t random variable with α degrees of freedom, for $\alpha > 4$. This matrix also follows Model 2.4 for each $p < \alpha$.

Finally, we present a general construction that takes a matrix from Model 2.4 and produces a sparse matrix that also satisfies the model, albeit with a larger value of the parameter ν.

Example 2.9 (Sparsified Random Matrix)**.** Let $\mathbf{\Phi}$ be a random matrix that satisfies Model 2.4 for some value $p > 4$ and $\nu \ge 1$. Let $\alpha \in (0, 1]$ be a thinning parameter, and construct a new random matrix $\mathbf{\Phi}^{(\alpha)}$ whose entries $\varphi_{ij}^{(\alpha)}$ are independent random variables with the distribution

$$\varphi_{ij}^{(\alpha)} = \begin{cases} \alpha^{-1/2}\varphi_{ij}, & \text{with probability } \alpha \\ 0, & \text{with probability } 1 - \alpha. \end{cases}$$

Then the sparsified random matrix $\mathbf{\Phi}^{(\alpha)}$ still follows Model 2.4 with the same value of p and with a modified value ν' of the other parameter: $\nu' = \alpha^{-1/p}\nu$.

3. A Universality Law for the Embedding Dimension

In this section, we present detailed results which show that, for a large class of sets, the embedding dimension is universal over a large class of linear maps.

3.1. **Embedding Dimension: Problem Formulation.** Let us begin with a more rigorous statement of the problem.

- Fix the ambient dimension D.
- Let E be a nonempty, compact subset of $\mathbb{R}^D$ that does not contain the origin.
- Let $\mathbf{\Pi} : \mathbb{R}^D \to \mathbb{R}^d$ be a random linear map with embedding dimension d.

We are interested in understanding the probability that the random projection $\mathbf{\Pi}(E)$ does not contain the origin. That is, we want to study the following predicate:

$$\mathbf{0} \notin \mathbf{\Pi}(E) \quad \text{or, equivalently,} \quad E \cap \operatorname{null}(\mathbf{\Pi}) = \emptyset. \tag{3.1}$$

We say that the random projection *succeeds* when the property (3.1) holds; otherwise, the random projection *fails*.

Our goal is to argue that there is a large class of sets and a large class of random linear maps where the probability that (3.1) holds depends primarily on the choice of the embedding dimension d and an appropriate measure of the size of the set E. In particular, the probability does *not* depend significantly on the distribution of the random linear map $\mathbf{\Pi}$.

3.2. **Some Concepts from Spherical Geometry.** The property (3.1) does not reflect the scale of the points in the set E. As a consequence, it is appropriate to translate the problem into a question about spherical geometry. We begin with a definition.

Definition 3.1 (Spherical Retraction)**.** The *spherical retraction* map $\boldsymbol{\theta} : \mathbb{R}^D \to \mathbb{R}^D$ is defined as

$$\boldsymbol{\theta}(\boldsymbol{t}) := \begin{cases} \boldsymbol{t} / \|\boldsymbol{t}\|, & \boldsymbol{t} \neq \mathbf{0} \\ \mathbf{0}, & \boldsymbol{t} = \mathbf{0}. \end{cases}$$

For every set E in $\mathbb{R}^D$, we have the equivalence

$$\mathbf{0} \notin \mathbf{\Pi}(E) \quad \text{if and only if} \quad \mathbf{0} \notin \mathbf{\Pi}(\boldsymbol{\theta}(E)). \tag{3.2}$$

Therefore, we may pass to the spherical retraction $\boldsymbol{\theta}(E)$ of the set E without loss of generality.

To obtain a complete analysis of when random projection succeeds or fails, we must restrict our attention to sets that have a convexity property.

Definition 3.2 (Spherical Convexity)**.** A nonempty subset Ω of the unit sphere in $\mathbb{R}^D$ is *spherically convex* when the pre-image $\boldsymbol{\theta}^{-1}(\Omega) \cup \{\mathbf{0}\}$ is a convex cone. By convention, the empty set is also spherically convex.

In particular, suppose that T is a compact, convex set that does not contain the origin. Then the retraction $\boldsymbol{\theta}(T)$ is compact and spherically convex.

Next, we introduce a polarity operation for spherical sets that supports some crucial duality arguments.

Definition 3.3 (Spherical Polarity)**.** Let Ω be a subset of the unit sphere S^{D-1} in $\mathbb{R}^D$. The *polar* of Ω is the set

$$\Omega^\circ := \left\{ \boldsymbol{x} \in \mathsf{S}^{D-1} : \boldsymbol{x} \cdot \boldsymbol{t} \leq 0 \text{ for all } \boldsymbol{t} \in \Omega \right\}.$$

By convention, the polar of the empty set is the whole sphere: $\emptyset^\circ := \mathsf{S}^{D-1}$.

This definition is simply the spherical analog of polarity for cones. It can be verified that Ω° is always closed and spherically convex. Furthermore, the polarity operation is an involution on the class of closed, spherically convex sets in S^{D-1}.

3.3. **The Statistical Dimension Functional.** We have the intuition that, for a given compact subset E of $\mathbb{R}^D$, the probability that a random linear map $\mathbf{\Pi} : \mathbb{R}^D \to \mathbb{R}^d$ succeeds decreases with the "content" of the set E. In the present context, the correct notion of content involves a geometric functional called the statistical dimension.

Definition 3.4 (Statistical Dimension)**.** Let Ω be a nonempty subset of the unit sphere in $\mathbb{R}^D$. The *statistical dimension* $\delta(\Omega)$ is defined as

$$\delta(\Omega) := \mathbb{E}\left[\left(\sup_{\boldsymbol{t}\in\Omega} \boldsymbol{g}\cdot\boldsymbol{t}\right)_+^2\right] \tag{3.3}$$

In addition, define $\delta(\varnothing) := 0$. We extend the statistical dimension to a general subset T in $\mathbb{R}^D$ using the spherical retraction:

$$\delta(T) := \delta(\boldsymbol{\theta}(T)). \tag{3.4}$$

Recall that the random vector $\boldsymbol{g} \in \mathbb{R}^D$ has the standard normal distribution.

The statistical dimension has a number of striking properties. We include a short summary; see [ALMT14, Prop. 3.1] and the citations there for further information.

- For a set T in $\mathbb{R}^D$, the statistical dimension $\delta(T)$ takes values in the range $[0, D]$.
- The statistical dimension is increasing with respect to set inclusion: $S \subset T$ implies that $\delta(S) \le \delta(T)$.
- The statistical dimension agrees with the linear dimension on subspaces:

$$\delta(L) = \dim(L) \quad \text{for each subspace } L \text{ in } \mathbb{R}^D.$$

- The statistical dimension interacts nicely with polarity:

$$\delta(\Omega^\circ) = D - \delta(\Omega) \quad \text{for each spherically convex set } \Omega \in \mathbb{R}^D. \tag{3.5}$$

  The same relation holds if we replace Ω by a convex cone K in $\mathbb{R}^D$ and use conic polarity.

As a specific example of (3.5), we can evaluate the statistical dimension of the nonnegative orthant $\mathbb{R}^D_+$. Indeed, the orthant is a self-dual cone, so

$$\delta(\mathbb{R}^D_+) = D/2. \tag{3.6}$$

There is also powerful machinery, developed in [Sto09, OH10, CRPW12, ALMT14, FM14], for computing the statistical dimension of a descent cone of a convex function. Finally, we mention that the statistical dimension can often be evaluated by sampling Gaussian vectors and approximating the expectation in (3.3) with an empirical average.

Remark 3.5 (Gaussian Width)**.** The statistical dimension is related to the *Gaussian width* functional. For a bounded set T in $\mathbb{R}^D$, the Gaussian width $\mathscr{W}(T)$ is defined as

$$\mathscr{W}(T) := \mathbb{E}\sup_{\boldsymbol{t}\in T} \boldsymbol{g}\cdot\boldsymbol{t}. \tag{3.7}$$

For a subset Ω of the unit sphere in $\mathbb{R}^D$, we have the inequalities

$$\mathscr{W}^2(\Omega) \le \delta(\Omega) \le \mathscr{W}^2(\Omega) + 1. \tag{3.8}$$

See [ALMT14, Prop. 10.3] for a proof. The relation (3.8) allows us to pass between the squared width and the statistical dimension.

The Gaussian width of a set is closely related to its mean width [Ver15, Sec. 1.3.5]. The mean width is a canonical functional in the Euclidean setting [Gru07, Sec. 7.3], but it is not quite the right choice for spherical geometry. We have chosen to work with the statistical dimension because it has many geometric properties that the width lacks in the spherical setting.

Remark 3.6 (Convex Cones)**.** The papers [ALMT14, MT14a] define the statistical dimension of a convex cone using intrinsic volumes. It can be shown that the statistical dimension is the canonical (additive, continuous) extension of the linear dimension to the class of closed convex cones [ALMT14, Sec. 5.6]. Our general definition (3.4) agrees with the original definition on this class.

3.4. **Theorem I: Universality of the Embedding Dimension.** We are now prepared to state our main result on the probability that a random linear map succeeds or fails for a given set.

Theorem I (Universality of the Embedding Dimension)**.** *Fix the parameters $p > 4$ and $\nu \geq 1$ for Model 2.4. Choose parameters $\varrho \in (0,1)$ and $\varepsilon \in (0,1)$. There is a number $N := N(p,\nu,\varrho,\varepsilon)$ for which the following statement holds. Suppose that*

- *The ambient dimension $D \geq N$.*
- *E is a nonempty, compact subset of $\mathbb{R}^D$ that does not contain the origin.*
- *The statistical dimension of E is proportional to the ambient dimension: $\delta(E) \geq \varrho D$.*
- *The $d \times D$ random linear map $\mathbf{\Pi}$ obeys Model 2.4 with parameters p and ν.*

Then

$$d \geq (1+\varepsilon)\,\delta(E) \quad \textit{implies} \quad \mathbb{P}\{\mathbf{0} \notin \mathbf{\Pi}(E)\} \geq 1 - \mathrm{C}_p D^{1-p/4}. \tag{a}$$

Furthermore, if $\boldsymbol{\theta}(E)$ is spherically convex, then

$$d \leq (1-\varepsilon)\,\delta(E) \quad \textit{implies} \quad \mathbb{P}\{\mathbf{0} \in \mathbf{\Pi}(E)\} \geq 1 - \mathrm{C}_p D^{1-p/4}. \tag{b}$$

The constant C_p depends only on the parameter p in the random matrix model.

Section 3.5 summarizes our strategy for establishing Theorem I. The detailed proof of Theorem I(a) appears in Section 10.3; the detailed proof of Theorem I(b) appears in Section 17.2. Let us mention that stronger probability bounds hold when the random linear map is drawn from Model 2.1.

For any random linear map $\mathbf{\Pi}$ drawn from Model 2.4, Theorem I(a) ensures that the random dimension reduction $\mathbf{\Pi}(E)$ is likely to succeed when the embedding dimension d exceeds the statistical dimension $\delta(E)$. Similarly, Theorem I(b) shows that $\mathbf{\Pi}(E)$ is likely to fail when the embedding dimension d is smaller than the statistical dimension $\delta(E)$, provided that $\boldsymbol{\theta}(E)$ is spherically convex. Note that both of these interpretations require that the statistical dimension $\delta(E)$ is not too small as compared with the ambient dimension D.

We have already seen a concrete example of Theorem I at work. In view of the calculation (3.6) of the statistical dimension of the orthant, the theorem provides a satisfying explanation of Figure 1.2!

Remark 3.7 (Prior Work)**.** When the random linear map $\mathbf{\Pi}$ is Gaussian, the result Theorem I(a) follows from Gordon [Gor88, Cor. 3.4], while the conclusion (b) seems to be more recent [ALMT14, Thm. I]. To our knowledge, the literature contains no precedent for Theorem I for general sets and for non-Gaussian linear maps. We can identify only a few sporadic special cases. Donoho & Tanner [DT10] studied the problem of recovering a "saturated vector" from random measurements via ℓ_∞ minimization. Their work can be interpreted as a statement about random embeddings of the set of unit vectors with nonnegative coordinates. They demonstrate that the phase transition in the embedding dimension is universal when the rows of the linear map matrix are independent, symmetric random vectors with a density; this result actually follows from classical work of Schäfli [Sch50] and Wendel [Wen62]. Let us emphasize that this result does not apply to discrete random linear maps.

Bayati et al. [BLM15] have studied the problem of recovering a sparse vector from random measurements via ℓ_1 minimization. Their work can be interpreted as a result on the embedding dimension of the set of descent directions of the ℓ_1 norm at a sparse vector. They showed that, asymptotically, the phase transition in the embedding dimension is universal. Their result requires the linear map matrix to contain independent, subgaussian entries that are absolutely continuous with respect to the Gaussian density. See Section 5.1 for more discussion of this result.

There are also many papers in asymptotic convex geometry and mathematical signal processing that contain theory about the order of the embedding dimension for linear maps from Model 2.4; see [MPTJ07, Men10, Men14, Tro15b, Ver15]. These results do not allow us to reach any conclusions about the existence of a phase transition or its location. There is also an extensive amount of empirical work, such as [DT09a,

Sto09, OH10, CSPW11, DGM13, MT14b], that suggests that phase transitions are ubiquitous in high-dimensional signal processing problems, but there has been no theoretical explanation of the universality phenomenon until now.

3.5. **Theorem I: Proof Strategy.** The proof of Theorem I depends on converting the geometric question to an analytic problem. First, recall the equivalence (3.2), which allows us to pass from the compact set E to its spherical retraction $\Omega := \boldsymbol{\theta}(E)$. Next, we identify two analytic quantities that determine whether a linear map annihilates a point in the set Ω.

Proposition 3.8 (Analytic Conditions for Embedding)**.** *Let Ω be a nonempty, closed subset of the unit sphere S^{D-1} in $\mathbb{R}^D$, and let $\boldsymbol{A} : \mathbb{R}^D \to \mathbb{R}^d$ be a linear map. Then*

$$\min_{\boldsymbol{t} \in \Omega} \|\boldsymbol{A}\boldsymbol{t}\| > 0 \quad \textit{implies} \quad \mathbf{0} \notin \boldsymbol{A}(\Omega). \tag{3.9}$$

Furthermore, if Ω is spherically convex and $\operatorname{cone}(\Omega)$ *is not a subspace,*

$$\min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in \operatorname{cone}(\Omega^\circ)} \|\boldsymbol{s} - \boldsymbol{A}^* \boldsymbol{t}\| > 0 \quad \textit{implies} \quad \mathbf{0} \in \boldsymbol{A}(\Omega). \tag{3.10}$$

Finally, if $\operatorname{cone}(\Omega)$ *is a subspace, select an arbitrary subset $\Omega_0 \subset \Omega$ with the property that* $\operatorname{cone}(\Omega_0)$ *is a* $(\dim(\Omega) - 1)$*-dimensional subspace. Then*

$$\min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in \operatorname{cone}(\Omega_0^\circ)} \|\boldsymbol{s} - \boldsymbol{A}^* \boldsymbol{t}\| > 0 \quad \textit{implies} \quad \mathbf{0} \in \boldsymbol{A}(\Omega). \tag{3.11}$$

Recall that $\operatorname{cone}(S)$ *is the smallest convex cone containing the set S.*

Proof. The implication (3.9) is quite easy to check. At each point $\boldsymbol{t} \in \Omega$, we have $\|\boldsymbol{A}\boldsymbol{t}\| > 0$, which implies that $\boldsymbol{A}\boldsymbol{t} \neq \mathbf{0}$. In other words, $\mathbf{0} \notin \boldsymbol{A}(\Omega)$.

The second implication (3.10) follows from a spherical duality principle. Suppose that Ω and Υ are closed and spherically convex, and assume $\operatorname{cone}(\Omega)$ is not a subspace. If Ω° and $-\Upsilon^\circ$ do not intersect, then Ω and Υ must intersect; see [Kle55, Thm (2.7)]. The analytic condition

$$\min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in \operatorname{cone}(\Omega^\circ)} \|\boldsymbol{s} - \boldsymbol{A}^* \boldsymbol{t}\| > 0$$

ensures that $\operatorname{cone}(\Omega^\circ)$ lies at a positive distance from $\boldsymbol{\theta}(\operatorname{range}(\boldsymbol{A}^*))$. It follows that Ω° does not intersect $\boldsymbol{\theta}(\operatorname{range}(\boldsymbol{A}^*))$. By duality, we conclude that $\Omega \cap \boldsymbol{\theta}(\operatorname{null}(\boldsymbol{A})) \neq \emptyset$, which yields (3.10).

Finally, if $\operatorname{cone}(\Omega)$ is a subspace, we use a dimension counting argument. As before, the analytic condition (3.11) implies that $\Omega_0^\circ \cap \boldsymbol{\theta}(\operatorname{range}(\boldsymbol{A}^*)) = \emptyset$. Since these sets do not intersect, the sum of their dimensions cannot exceed the ambient dimension:

$$\begin{aligned} D \geq \dim(\Omega_0^\circ) + \dim(\operatorname{range}(\boldsymbol{A}^*)) &= (D - \dim(\Omega_0)) + (D - \dim(\operatorname{null}(\boldsymbol{A}))) \\ &= 2D + 1 - \dim(\Omega) - \dim(\operatorname{null}(\boldsymbol{A})). \end{aligned}$$

We see that $\dim(\Omega) + \dim(\operatorname{null}(\boldsymbol{A})) \geq D + 1$. As a consequence, the subspace $\operatorname{cone}(\Omega)$ and $(\boldsymbol{A})$ must have a nontrivial intersection. □

In view of Proposition 3.8, we can establish Theorem I(a) by showing that

$$d \geq (1 - \varepsilon)\, \delta(\Omega) \quad \text{implies} \quad \min_{\boldsymbol{t} \in \Omega} \|\boldsymbol{\Pi}\boldsymbol{t}\| > 0 \quad \text{with high probability.}$$

This condition follows from a universality result, Corollary 10.1, for the restricted minimum singular value. The proof appears in Section 10.3. Similarly, we can establish Theorem I(b) by showing that

$$d \leq (1 + \varepsilon)\, \delta(\Omega) \quad \text{implies} \quad \min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in \operatorname{cone}(\Omega^\circ)} \|\boldsymbol{s} - \boldsymbol{\Pi}^* \boldsymbol{t}\| > 0 \quad \text{with high probability.}$$

This condition follows from a specialized argument that culminates in Corollary 17.1. The proof appears in Section 17.2. We need not lavish extra attention on the case where $\operatorname{cone}(\Omega)$ is a subspace. Indeed, the dimension of Ω and Ω_0 only differ by one, which is invisible in the final results.

3.6. **Theorem I: Extensions.** It is an interesting challenge to delineate the scope of the universality phenomenon described in Theorem I. We believe that there remain many opportunities for improving on this result.

- Theorem I only shows that the width of the phase transition is $o(D)$. It is known [ALMT14, Thm. 7.1] that the width of the phase transition has order $\min\{\sqrt{\delta(E)}, \sqrt{D-\delta(E)}\}$ for a linear map with standard normal entries. How wide is the phase transition for more general random linear maps?
- A related question is whether Theorem I holds for those sets E whose statistical dimension $\delta(E)$ is much smaller than the ambient dimension.
- There is empirical evidence that the location of the phase transition is universal over a class wider than Model 2.4. In particular, results like Theorem I may be valid for structured random linear maps.
- Figure 1.2 suggests that the *probability* of successful embedding is universal. Under what conditions can this observation be formalized?

In summary, Theorem I is just the first step toward a broader theory of universality in high-dimensional stochastic geometry.

4. A Universality Law for the Restricted Minimum Singular Value

This section describes a quantitative universality law for random linear maps. We show that the restricted minimum singular value of a random linear map takes the same value for every linear map in a substantial class. This type of result provides information about the stability of randomized dimension reduction.

4.1. **Restricted Minimum Singular Value: Problem Formulation.** Let us frame our assumptions:

- Fix the ambient dimension D.
- Let E be a nonempty, compact subset of the unit ball B^D.
- Let $\mathbf{\Pi} : \mathbb{R}^D \to \mathbb{R}^d$ be a random linear map with embedding dimension d.

In this section, our goal is to understand the distance from the random projection $\mathbf{\Pi}(E)$ to the origin. The following definition captures this property.

Definition 4.1 (Restricted Minimum Singular Value)**.** Let $\boldsymbol{A} : \mathbb{R}^D \to \mathbb{R}^d$ be a linear map, and let T be a nonempty subset of $\mathbb{R}^D$. The *restricted minimum singular value* of $\boldsymbol{A}$ with respect to the set T is the quantity

$$\sigma_{\min}(\boldsymbol{A}; T) := \inf_{\boldsymbol{t} \in T} \|\boldsymbol{A}\boldsymbol{t}\|.$$

More briefly, we write *restricted singular value* or *RSV*.

Proposition 3.8 shows that the restricted minimum singular value is a quantity of interest when studying the embedding dimension of a linear map. It is also productive to think about $\sigma_{\min}(\boldsymbol{A}; T)$ as a measure of the stability of inverting the map $\boldsymbol{A}$ on the image $\boldsymbol{A}(T)$. In particular, note that the restricted singular value $\sigma_{\min}(\boldsymbol{A}; T)$ is a generalization of the ordinary minimum singular value $\sigma_{\min}(\boldsymbol{A})$, which we obtain from the selection $T = \mathsf{S}^{D-1}$.

4.2. **The Excess Width Functional.** It is clear that the restricted singular value $\sigma_{\min}(\boldsymbol{A}; \cdot)$ decreases with respect to set inclusion. In other words, the restricted singular value depends on the "content" of the set T. In the case of a random linear map, the following geometric functional provides the correct notion of content.

Definition 4.2 (Excess Width)**.** Let m be a positive number, and let T be a nonempty, bounded subset of $\mathbb{R}^D$. The *m-excess width* of T is the quantity

$$\mathscr{E}_m(T) := \mathbb{E} \inf_{\boldsymbol{t} \in T} \left(\sqrt{m}\, \|\boldsymbol{t}\| + \boldsymbol{g} \cdot \boldsymbol{t} \right). \tag{4.1}$$

Versions of the excess width appear as early as the work of Gordon [Gor88, Cor. 1.1]. It has also come up in recent papers [Sto13, OTH13, TOH15, TAH15] on the analysis of Gaussian random linear maps.

The excess width has a number of useful properties. These results are immediate consequences of the definition.

- For a subset T of the unit ball in $\mathbb{R}^D$, the m-excess width satisfies the bounds
$$\sqrt{m} - \sqrt{D} \le \mathscr{E}_m(T) \le \sqrt{m}.$$
In particular, the excess width can be positive or negative.
- The m-excess width is weakly increasing in m. That is, $m \le n$ implies $\mathscr{E}_m(T) \le \mathscr{E}_n(T)$.
- The m-excess width is decreasing with respect to set inclusion: $S \subset T$ implies $\mathscr{E}_m(S) \ge \mathscr{E}_m(T)$.
- The m-excess width is absolutely homogeneous: $\mathscr{E}_m(\alpha T) = |\alpha|\, \mathscr{E}_m(T)$ for $\alpha \in \mathbb{R}$.
- The excess width (4.1) is related to the Gaussian width (3.7):
$$\mathscr{E}_m(\Omega) = \sqrt{m} - \mathscr{W}(\Omega) \quad \text{for } \Omega \subset \mathsf{S}^{D-1}. \tag{4.2}$$
Using (3.8), we can also relate the excess width to the statistical dimension:
$$\sqrt{m} - \sqrt{1+\delta(\Omega)} \le \mathscr{E}_m(\Omega) \le \sqrt{m} - \sqrt{\delta(\Omega)} \quad \text{for } \Omega \subset \mathsf{S}^{D-1}. \tag{4.3}$$

The term "excess width" is not standard, but the formula (4.2) suggests that this moniker is appropriate. According to Theorem I, the sign $(\pm)$ of the excess width $\mathscr{E}_d(\Omega)$ indicates that random dimension reduction $\mathbf{\Pi}(\Omega)$ with embedding dimension d succeeds $(+)$ or fails $(-)$ with high probability.

The papers [Sto13, OTH13, TOH15] develop methods for computing the excess width in a variety of situations. For instance, if L is a k-dimensional subspace of $\mathbb{R}^D$, then
$$\mathscr{E}_m(L \cap \mathsf{S}^{D-1}) \approx \sqrt{m} - \sqrt{k}.$$
For a more sophisticated example, consider the probability simplex
$$\Delta_D := \left\{ \boldsymbol{t} \in \mathbb{R}^D : \textstyle\sum_{i=1}^D t_i = 1, t_i \ge 0 \text{ for } i = 1, \dots, D \right\}.$$
We can develop an asymptotic expression for its excess width:
$$\lim_{\substack{D,d\to\infty \\ D/d=\varrho}} \mathscr{E}_d(\Delta_D) = \inf_{\alpha \ge 0} \left(\alpha - \inf_{s \in \mathbb{R}} \left(s + \alpha \sqrt{\varrho\, q(s)} \right) \right) \quad \text{where} \quad q(s) := \mathbb{E}\left[(\gamma - s)_+^2 \right]. \tag{4.4}$$
The proof of the formula (4.4) is involved, so we must omit the details; see [TAH15] for a framework for making such computations. See Part II for some other examples. It is also possible to estimate the excess width numerically by approximating the expectation in (4.1) with an empirical average.

4.3. **Theorem II: Universality for the Restricted Minimum Singular Value.** With this preparation, we can present our main result about the universality properties of the restricted minimum singular value of a random linear map.

Theorem II (Universality for the Restricted Minimum Singular Value)**.** *Fix the parameters p and ν for Model 2.4. Choose parameters $\lambda \in (0,1)$ and $\varrho \in (0,1)$ and $\varepsilon \in (0,1)$. There is a number $N := N(p, \nu, \lambda, \varrho, \varepsilon)$ for which the following statement holds. Suppose that*

- *The ambient dimension $D \ge N$.*
- *E is a nonempty, compact subset of the unit ball B^D in $\mathbb{R}^D$.*
- *The embedding dimension d is in the range $\lambda D \le d \le D^{6/5}$.*
- *The d-excess width of E is not too small: $\mathscr{E}_d(E) \ge \varrho\sqrt{d}$.*
- *The random linear map $\mathbf{\Pi} : \mathbb{R}^D \to \mathbb{R}^d$ obeys Model 2.4 with parameters p and ν.*

Then
$$\mathbb{P}\left\{ \sigma_{\min}(\mathbf{\Pi}; E) \ge (1-\varepsilon)\left(\mathscr{E}_d(E) \right)_+ \right\} \ge 1 - \mathrm{C}_p D^{1-p/4}. \tag{a}$$

Furthermore, if E is convex, then

$$\mathbb{P}\left\{\sigma_{\min}(\mathbf{\Pi}; E) \le (1+\varepsilon)\left(\mathscr{E}_d(E)\right)_+\right\} \ge 1 - \mathrm{C}_p D^{1-p/4}. \tag{b}$$

The constant C_p depends only on the parameter p in the random matrix model.

Section 4.4 contains an overview of the argument. The detailed proof of Theorem II appears in Section 10.2. Stronger probability bounds hold when the random linear map is drawn from Model 2.1.

For any random linear map $\mathbf{\Pi}$ drawn from Model 2.4, Theorem II asserts that the distance of the random projection $\mathbf{\Pi}(E)$ from the origin is typically not much smaller than the d-excess width $\mathscr{E}_d(E)$. Similarly, when E is convex, the distance of the random image from the origin is typically not much larger than the d-excess width. These conclusions require that the excess width $\mathscr{E}_d(E)$ is not too small as compared with the root $\sqrt{d}$ of the embedding dimension.

Theorem II is vacuous when $\mathscr{E}_d(E) \le 0$. Nevertheless, in case $\boldsymbol{\theta}(E)$ is spherically convex, the condition $\mathscr{E}_d(\boldsymbol{\theta}(E)) + o(\sqrt{D}) \le 0$ implies that $\sigma_{\min}(\mathbf{\Pi}; E) = 0$ with high probability because of (4.3) and Theorem I.

We have already seen an illustration of Theorem II in action. In view of the expression (4.4) for the excess width of the probability simplex, the theorem explains the experiment documented in Figure 1.4!

Remark 4.3 (Prior Work)**.** When the random linear map $\mathbf{\Pi}$ is Gaussian, Gordon's work [Gor88, Cor. 1.1] yields the conclusion Theorem II(a), while the second conclusion (b) appears to be more recent; see [Sto13, OTH13, TOH15, TAH15].

Korada & Montanari [KM11] have established universality laws for some optimization problems that arise in communications. The mathematical challenge in their work is similar to showing that the RSV of a cube is universal for Model 2.4 with $p \ge 6$. We have adopted some of their arguments in our work. On the other hand, the cube does not present any of the technical difficulties that arise for general sets.

For other types of random linear maps and other types of sets, we are not aware of any research that yields the precise value of the restricted singular value, which is required to assert universality. The literature on asymptotic convex geometry and mathematical signal processing, e.g., [MPTJ07, Men10, Men14, Tro15b, Ver15], contains many results on the restricted singular value that provide bounds of the correct order with unspecified multiplicative constants.

4.4. **Theorem II: Proof Strategy.** One standard template for proving a universality law factors the problem into two parts. First, we compare a general random structure with a canonical structure that has additional properties. Then we use these extra properties to obtain a complete analysis of the canonical structure. Because of the comparison, we learn that the general structure inherits some of the behavior of the canonical structure. This recipe describes Lindeberg's approach [Lin22] to the central limit theorem; see [Tao12, Sec. 2.2.4]. More recently, similar methods have been directed at harder problems, such as the universality of local spectral statistics of a non-symmetric random matrix [TV15]. Our research is closest in spirit to the papers [MOO10, Cha06, KM11].

Let us imagine how we could apply this technique to prove Theorem II. Suppose that E is a compact, convex set. After performing standard discretization and smoothing steps, we might invoke the Lindeberg exchange principle to compare the restricted singular value of the $d \times D$ random linear map $\mathbf{\Pi}$ with that of a $d \times D$ standard normal matrix $\mathbf{\Gamma}$:

$$\mathbb{E}\,\sigma_{\min}(\mathbf{\Pi}; E) \approx \mathbb{E}\,\sigma_{\min}(\mathbf{\Gamma}; E).$$

To evaluate the restricted minimum singular value of a standard normal matrix, we can use the Gaussian Minimax Theorem, as in [Gor88, Cor. 1.1]:

$$\mathbb{E}\,\sigma_{\min}(\mathbf{\Gamma}; E) \gtrsim \left(\mathscr{E}_d(E)\right)_+.$$

Last, we can incorporate a convex duality argument, as in [Sto13, OTH13, TOH15, TAH15], to obtain the reverse inequality:

$$\mathbb{E}\,\sigma_{\min}(\mathbf{\Gamma}; E) \lesssim \left(\mathscr{E}_d(E)\right)_+.$$

Unfortunately, this simple approach is not adequate. Our argument ultimately follows a related pattern, but we have to overcome a number of technical obstacles.

Let us explain the most serious issue and our mechanism for handling it. We would like to apply the Lindeberg exchange principle to $\sigma_{\min}(\mathbf{\Pi}; E)$ to replace each entry of $\mathbf{\Pi}$ with a standard normal variable. The problem is that E may contain a vector $\boldsymbol{t}_0$ with large entries. If we try to replace the columns of $\mathbf{\Pi}$ associated with the large components of $\boldsymbol{t}_0$, we incur an intolerably large error. Moreover, for any given coordinate j, the set E may contain a vector $\boldsymbol{t}_j$ that takes a large value in the distinguished coordinate j. This fact seems to foreclose the possibility of replacing any part of the matrix $\mathbf{\Pi}$.

We address this challenge by dissecting the index set E. For each (small) subset J of $\{1, \dots, D\}$, we define the set E_J of vectors in E whose components in J^c are small. First, we argue that the restricted singular value of a subset does not differ much from the restricted singular value of the entire set:

$$\mathbb{E}\,\sigma_{\min}(\mathbf{\Pi}; E) \approx \min_J \mathbb{E}\,\sigma_{\min}(\mathbf{\Pi}; E_J).$$

We may now use the Lindeberg method to make the comparison

$$\mathbb{E}\,\sigma_{\min}(\mathbf{\Pi}; E_J) \approx \mathbb{E}\,\sigma_{\min}(\mathbf{\Psi}; E_J),$$

where the matrix $\mathbf{\Psi}$ is a hybrid of the form $\mathbf{\Psi}_J = \mathbf{\Pi}_J$ and $\mathbf{\Psi}_{J^c} = \mathbf{\Gamma}_{J^c}$. That is, we only replace the columns of $\mathbf{\Pi}$ listed in the set J^c with independent standard normal variables. At this point, we need to compare the minimum singular value of $\mathbf{\Psi}$ restricted to the subset E_J against the excess width of the full set:

$$\mathbb{E}\,\sigma_{\min}(\mathbf{\Psi}; E_J) \approx \big(\mathscr{E}_d(E)\big)_+.$$

In other words, the remaining part $\mathbf{\Psi}_J$ of the original matrix plays a negligible role in determining the restricted singular value. We perform this estimate using the Gaussian Minimax Theorem [Gor85], some convex duality arguments [TOH15], and some coarse results from nonasymptotic random matrix theory [Ver12, Tro15c].

See Part III for detailed statements of the main technical results that support Theorem II and the proofs of these results. Most of the ingredients are standard tools from modern applied probability, but we have combined them in a subtle way. To make the long argument clearer, we have attempted to present the proof in a multi-resolution fashion where pieces from each level are combined explicitly at the level above.

4.5. **Theorem II: Extensions.** We expect that there are a number of avenues for extending Theorem II.

- Our analysis shows that the error in estimating $\sigma_{\min}(\mathbf{\Pi}; E)$ by the excess width $\mathscr{E}_d(E)$ is at most $o(\sqrt{D})$. In case $\mathbf{\Pi}$ is a standard normal matrix, the error actually has constant order [TOH15, Thm. II.1]. Can we improve the error bound for more general random linear maps?
- A related question is whether the universality of $\sigma_{\min}(\mathbf{\Pi}; E)$ persists when the excess width $\mathscr{E}_d(E)$ is very small in comparison with $\sqrt{d}$.
- There is some empirical evidence that the restricted minimum singular value may have universality properties for a class of random linear maps wider than Model 2.4.
- Our proof can probably be adapted to show that other types of functionals exhibit universal behavior. For example, we can study

  $$\min_{\boldsymbol{t} \in E} \big(\|\mathbf{\Pi}\boldsymbol{t}\|^2 + f(\boldsymbol{t}) \big)$$

  where f is a convex, Lipschitz function. This type of optimization problem plays an important role in statistics and machine learning.

At the same time, we know that many natural functionals do *not* exhibit universal behavior. In particular, consider the restricted *maximum* singular value:

$$\sigma_{\max}(\boldsymbol{A}; T) := \sup_{\boldsymbol{t} \in T} \|\boldsymbol{A}\boldsymbol{t}\|.$$

There are many sets T where the maximum singular value $\sigma_{\max}(\cdot; T)$ does not take a universal value for linear maps in Model 2.4; for example, see Figure 1.6. It is an interesting challenge to determine the full scope of the universality principle uncovered in Theorem II.

Part II. Applications

In this part of the paper, we outline some of the implications of Theorems I and II in the information sciences. We focus on problems involving least squares and ℓ_1 minimization to reduce the number of distinct calculations that we need to perform; nevertheless the techniques apply broadly. In an effort to make the presentation shorter and more intuitive, we have also chosen to sacrifice some precision.

Section 5 discusses structured signal recovery and, in particular, the compressed sensing problem. Section 6 presents an application in coding theory. Section 7 describes how the results allow us to analyze randomized algorithms for numerical linear algebra. Finally, Section 8 gives a universal formula for the prediction error in a sparse regression problem.

5. Recovery of Structured Signals from Random Measurements

In signal processing, dimension reduction arises as a mechanism for signal acquisition. The idea is that the number of measurements we need to recover a structured signal is comparable with the number of degrees of freedom in the signal, rather than the ambient dimension. Given the measurements, we can reconstruct the signal using nonlinear algorithms that take into account our prior knowledge about the structure. The field of compressed sensing is based on the idea that *random* linear measurements offer an efficient way to acquire structured signals. The practical challenge in implementing this proposal is to find technologies that can perform random sampling.

Our universality laws have a significant implication for compressed sensing. We prove that the number of random measurements required to recover a structured signal does not have a strong dependence on the distribution of the random measurements. This result is important because most applications offer limited flexibility in the type of measurement that we can take. Our theory justifies the use of a broad class of measurement ensembles.

5.1. The Phase Transition for Sparse Signal Recovery.

In this section, we study the phase transition that appears when we reconstruct a sparse signal from random measurements via ℓ_1 minimization. For a large class of random measurements, we prove that the distribution does not affect the location of the phase transition. This result resolves a major open question [DT09a] in the theory of compressed sensing.

Let us give a more precise description of the problem. Suppose that $\boldsymbol{x}_\star \in \mathbb{R}^n$ is a fixed vector with precisely s nonzero entries. Let $\boldsymbol{\Phi} : \mathbb{R}^n \to \mathbb{R}^m$ be a random linear map, and suppose that we have access to the image $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}_\star$. We interpret this data as a list of m samples of the unknown signal.

A standard approach [CDS98, DH01, Tro06, CRT06, Don06a] for reconstructing the sparse signal $\boldsymbol{x}_\star$ from the data is to solve a convex optimization problem:

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad \|\boldsymbol{x}\|_{\ell_1} \quad \text{subject to} \quad \boldsymbol{\Phi}\boldsymbol{x} = \boldsymbol{y}. \tag{5.1}$$

Minimizing the ℓ_1 norm promotes sparsity in the optimization variable $\boldsymbol{x}$, and the constraint ensures that the virtual measurements $\boldsymbol{\Phi}\boldsymbol{x}$ are consistent with the observed data $\boldsymbol{y}$. We say that the optimization problem (5.1) *succeeds* if it has a unique solution $\widehat{\boldsymbol{x}}$ that coincides with $\boldsymbol{x}_\star$. Otherwise, it *fails*. In this setting, the following challenge arises.

> **The Compressed Sensing Problem:** Is the optimization problem (5.1) likely to succeed or to fail to recover $\boldsymbol{x}_\star$ as a function of the sparsity s, the number m of measurements, the ambient dimension n, and the distribution of the random measurement matrix $\boldsymbol{\Phi}$?

This question has been a subject of inquiry in thousands of papers over the last 10 years; see the books [EK12, FR13] for more background and references.

In a series of recent papers [DT09b, Sto09, CRPW12, BLM15, ALMT14, Sto13, OTH13, FM14, GNP14], the compressed sensing problem has been solved completely in the case where the random measurement matrix $\boldsymbol{\Phi}$ follows the standard normal distribution. See Remark 5.2 for a narrative of who did what when.

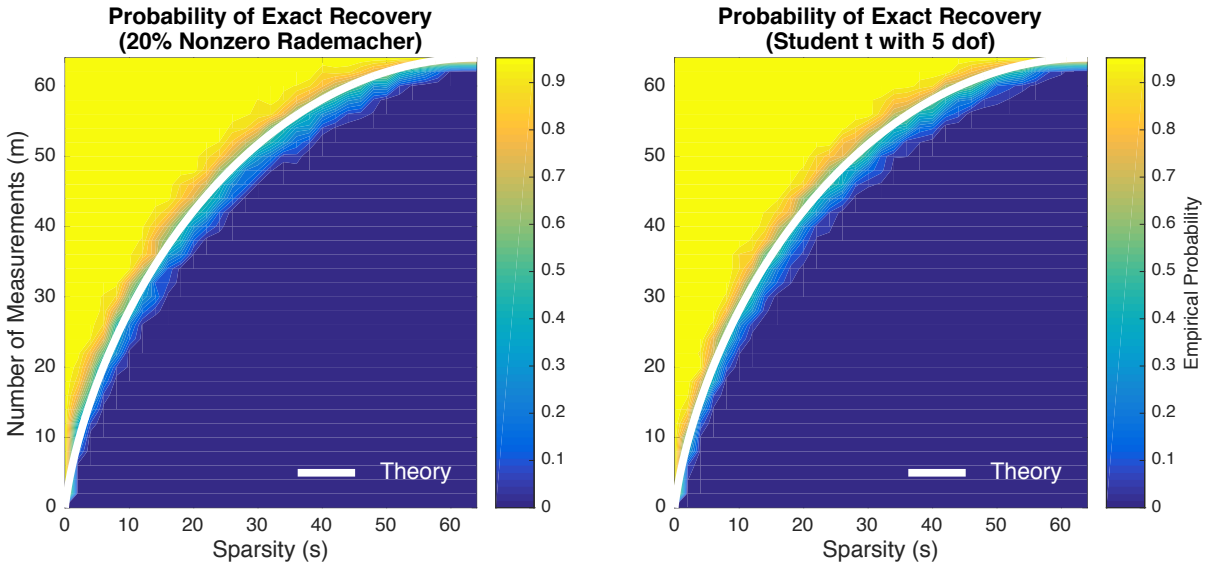

FIGURE 5.1: *Universality of the ℓ_1 Recovery Phase Transition.* These plots depict the empirical probability that the ℓ_1 minimization problem (5.1) recovers a vector $\boldsymbol{x}_\star \in \mathbb{R}^{64}$ with s nonzero entries from a vector of random measurements $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}_\star \in \mathbb{R}^m$. The **heatmap** indicates the empirical probability, computed over 100 trials. The **white curve** is the phase transition $n\psi_{\ell_1}(s/n)$ promised by Proposition 5.1. LEFT: The random measurement matrix $\boldsymbol{\Phi}$ is a sparse Rademacher matrix with an average of 20% nonzero entries. RIGHT: The random measurement matrix $\boldsymbol{\Phi}$ has independent Student t_5 entries. See Section 1.9 for more details.

In brief, there exists a phase transition function $\psi_{\ell_1} : [0,1] \to [0,1]$ defined by

$$\psi_{\ell_1}(\varrho) := \inf_{\tau \geq 0} \left(\varrho(1+\tau^2) + (1-\varrho)\sqrt{\frac{2}{\pi}} \int_\tau^\infty (\zeta - \tau)^2 \mathrm{e}^{-\zeta^2/2} \,\mathrm{d}\zeta \right). \tag{5.2}$$

The phase transition function is increasing and convex, and it satisfies $\psi_{\ell_1}(0) = 0$ and $\psi_{\ell_1}(1) = 1$. When the measurement matrix $\boldsymbol{\Phi}$ is standard normal,

$$\begin{aligned} m/n < \psi_{\ell_1}(s/n) - o(1) \quad &\text{implies} \quad \text{(5.1) fails with probability } 1 - o(1); \\ m/n > \psi_{\ell_1}(s/n) + o(1) \quad &\text{implies} \quad \text{(5.1) succeeds with probability } 1 - o(1). \end{aligned} \tag{5.3}$$

In other words, as the number m of measurements increases, the probability of success jumps from zero to one at the point $n\psi_{\ell_1}(s/n)$ over a range of $o(n)$ measurements. The error terms in (5.3) can be improved substantially, but this presentation suffices for our purposes.

Donoho & Tanner [DT09a] have performed an extensive empirical investigation of the phase transition in the ℓ_1 minimization problem (5.1). Their work suggests that the Gaussian phase transition (5.3) persists for many other types of random measurements. See Figure 5.1 for a small illustration. Our universality results provide the first rigorous explanation of this phenomenon for measurement matrices drawn from Model 2.4.

Proposition 5.1 (Universality of ℓ_1 Phase Transition)**.** *Assume that*

- *The ambient dimension n and the number m of measurements satisfy $m \leq n$.*
- *The vector $\boldsymbol{x}_\star \in \mathbb{R}^n$ has exactly s nonzero entries.*
- *The random measurement matrix $\boldsymbol{\Phi} : \mathbb{R}^n \to \mathbb{R}^m$ follows Model 2.4 with parameters p and ν.*
- *We observe the vector $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}_\star$.*

Then, as the ambient dimension $n \to \infty$,

$$m/n < \psi_{\ell_1}(s/n) - o(1) \quad \textit{implies} \quad (5.1)\ \textit{fails with probability } 1 - o(1);$$
$$m/n > \psi_{\ell_1}(s/n) + o(1) \quad \textit{implies} \quad (5.1)\ \textit{succeeds with probability } 1 - o(1).$$

The little-o suppresses constants that depend only on p and ν.

Proposition 5.1 follows directly from our universality result, Theorem I, and the established calculation (5.3) of the phase transition in the standard normal setting.

Proof. The approach is quite standard. Let Ω be the set of unit-norm descent directions of the ℓ_1 norm at the point $\boldsymbol{x}_\star$. That is,

$$\Omega := \big\{\boldsymbol{u} \in \mathsf{S}^{n-1} : \|\boldsymbol{x}_\star + \lambda \boldsymbol{u}\|_1 \le \|\boldsymbol{x}_\star\|_1 \text{ for some } \lambda > 0\big\}.$$

The primal optimality condition for (5.1) demonstrates that the reconstruction succeeds if and only if $\Omega \cap \operatorname{null}(\boldsymbol{\Phi}) = \emptyset$. Since the ℓ_1 norm is a closed convex function, the set $\boldsymbol{\theta}^{-1}(\Omega) \cup \{\mathbf{0}\}$ of all descent directions forms a closed, convex cone. Therefore, the set Ω is closed and spherically convex. It follows from Theorem I that the behavior of (5.1) undergoes a phase transition at the statistical dimension $\delta(\Omega)$ for any random linear map drawn from Model 2.4.

The result [ALMT14, Prop. 4.5] contains the first complete and accurate computation of the statistical dimension of a descent cone of the ℓ_1 norm:

$$n\psi_{\ell_1}(s/n) - O(\sqrt{n}) \le \delta(\Omega) \le n\psi_{\ell_1}(s/n). \tag{5.4}$$

See also [FM14, Prop. 1]. This fact completes the proof. □

Note that Proposition 5.1 requires the sparsity level s to be proportional to the ambient dimension n before it provides any information. By refining our argument, we can address the case when s is proportional to $n^{1-\varepsilon}$ for a small number ε that depends on the regularity of the random linear map. The empirical work [DT09a] of Donoho & Tanner is unable to provide statistical evidence for the universality hypothesis in the regime where s is very small. It remains an open problem to understand how rapidly s/n can vanish before the universality phenomenon fails.

The paper [DT09a] also contains numerical evidence that the ℓ_1 phase transition persists for random measurement systems that have more structure than Model 2.4. It remains an intriguing open question to understand these experiments.

Remark 5.2 (Prior Work)**.** In early 2005, Donoho [Don06c] and Donoho & Tanner [DT06] observed that there is a phase transition in the number of standard normal measurements needed to reconstruct a sparse signal via ℓ_1 minimization. Using methods from integral geometry, they were able to perform a heuristic computation of the location of the phase transition function (5.2). In subsequent work [DT09b], they proved that the transition (5.3) is valid in some parameter regimes. They later reported extensive empirical evidence [DT09a] that the distribution of the random measurement map has little effect on the location of the phase transition.

In early 2005, Rudelson & Vershynin [RV08] proposed a different approach to studying ℓ_1 minimization by adapting results of Gordon [Gor88] that depend on Gaussian process theory. Stojnic [Sto09] refined this argument to obtain an empirically sharp success condition for standard normal linear maps, but his work did not establish a matching failure condition. Stojnic's calculations were clarified and extended to other signal recovery problems in the papers [OH10, CRPW12, ALMT14, FM14].

Bayati et al. [BLM15] is the first paper to rigorously demonstrate that the phase transition (5.3) is valid for standard normal measurements. The argument is based on a state evolution framework for an iterative algorithm inspired by statistical physics. This work also gives the striking conclusion that the ℓ_1 phase transition is universal over a class of random measurement maps. This result requires the measurement matrix to have independent, standardized, subgaussian entries that are absolutely continuous with respect to the Gaussian distribution. As a consequence, the paper [BLM15] excludes discrete and heavy-tailed

models. Furthermore, it only applies to ℓ_1 minimization. In contrast, Proposition 5.1 and its proof have a much wider compass.

The paper [ALMT14] of Amelunxen et al. contains the first complete integral-geometric proof of (5.3) for standard normal measurements. This work is significant because it established for the first time that phase transitions are ubiquitous in signal reconstruction problems. It also forged the first links between the approach to phase transitions based on integral geometry and those based on Gaussian process theory. The papers [MT14a, GNP14] build on these ideas to obtain precise estimates of the success probability in (5.3).

Subsequently, Stojnic [Sto13] refined the Gaussian process methods to obtain more detailed information about the behavior of the errors in noisy variants of the ℓ_1 minimization problem. His work has been extended in a series [OTH13, TOH15, TAH15] of papers by Abbasi, Oymak, Thrampoulidis, Hassibi, and their collaborators. Because of this research, we now have a very detailed understanding of the behavior of convex signal recovery methods with Gaussian measurements.

To our knowledge, the current paper is the first work that extends the type of general analysis in [ALMT14, OTH13, TAH15] beyond the confines of the standard normal model.

5.2. **Other Signal Recovery Problems.** The compressed sensing problem is the most prominent example from a large class of related questions. Our universality results have implications for this entire class of problems. We include a brief explanation.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a proper[7] convex function whose value increases with the "complexity" of its argument. The ℓ_1 norm is an example of a complexity measure that is appropriate for sparse signals [CDS98]. Similarly, the Schatten 1-norm is a good complexity measure for low-rank matrices [Faz02].

Let $\boldsymbol{x}_\star \in \mathbb{R}^n$ be a vector with "low complexity." Draw a random linear map $\boldsymbol{\Phi} : \mathbb{R}^n \to \mathbb{R}^m$, and suppose we have access to $\boldsymbol{x}_\star$ only through the measurements $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}_\star$. We can attempt to reconstruct $\boldsymbol{x}_\star$ by solving the convex optimization problem

$$\operatorname*{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{\Phi}\boldsymbol{x} = \boldsymbol{y}.$$

In other words, we find the vector with minimum complexity that is consistent with the observed data. We say that (5.2) *succeeds* when it has a unique optimal point that coincides with $\boldsymbol{x}_\star$; otherwise, it *fails*.

The paper [ALMT14] proves that there is a phase transition in the behavior of (5.2) when $\boldsymbol{\Phi}$ is standard normal. Our universality law, Theorem I, allows us to extend this result to include every random linear map from Model 2.4. Define the set Ω of unit-norm descent directions of f at the point $\boldsymbol{x}_\star$:

$$\Omega := \left\{ \boldsymbol{u} \in \mathsf{S}^{n-1} : f(\boldsymbol{x}_\star + \lambda \boldsymbol{u}) \le f(\boldsymbol{x}_\star) \text{ for some } \lambda > 0 \right\}.$$

Then, as the ambient dimension $n \to \infty$,

$$\begin{aligned} m < \delta(\Omega) - o(n) \quad &\text{implies} \quad \text{(5.2) fails with probability } 1 - o(1); \\ m > \delta(\Omega) + o(n) \quad &\text{implies} \quad \text{(5.2) succeeds with probability } 1 - o(1). \end{aligned}$$

In other words, there is a phase transition in the behavior of (5.2) when the number m of measurements equals the statistical dimension $\delta(\Omega)$ of the set of descent directions of f at the point $\boldsymbol{x}_\star$. See the papers [CRPW12, ALMT14, FM14] for some general methods for computing the statistical dimension of a descent cone.

5.3. **Complements.** We conclude this section with a few additional remarks about the scope of our results on signal recovery. First, we discuss some geometric applications. Second, we mention some other signal processing problems that can be studied using the same methods.

[7] A proper convex function takes at least one finite value and never takes the value $-\infty$.

5.3.1. *Geometric Implications.* Proposition 5.1 can be understood as a statement about the facial structure of a random projection of the ℓ_1 ball. Equivalently, it provides information about the facial structure of the convex hull of random points.

Suppose that B_1^n is the n-dimensional ℓ_1 ball, and fix an $(s-1)$-dimensional face F of B_1^n. Let $\mathbf{\Phi} : \mathbb{R}^n \to \mathbb{R}^m$ be a random linear map from Model 2.4. We say that $\mathbf{\Phi}$ *preserves* the face F when $\mathbf{\Phi}(F)$ is an $(s-1)$-dimensional face of the projection $\mathbf{\Phi}(\mathsf{B}_1^n)$. We can reinterpret Proposition 5.1 as saying that

$$\begin{aligned} m/n < \psi_{\ell_1}(s/n) - o(1) \quad &\text{implies} \quad \mathbf{\Phi} \text{ is unlikely to preserve } F; \\ m/n > \psi_{\ell_1}(s/n) + o(1) \quad &\text{implies} \quad \mathbf{\Phi} \text{ is likely to preserve } F. \end{aligned}$$

The connection between ℓ_1 minimization and the facial structure of the ℓ_1 ball was identified in [Don06b]; see also [ALMT14, Sec. 10.1.1].

Here is another way of framing the same result. Fix an index set $J \subset \{1, \dots, n\}$ with cardinality $\#J = s$ and a vector $\boldsymbol{\eta} \in \mathbb{R}^n$ with ± 1 entries. Let $\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_n \in \mathbb{R}^m$ be independent random vectors, drawn from Model 2.4, and consider the absolute convex hull $E := \operatorname{conv}\{\pm\boldsymbol{\varphi}_1, \dots, \pm\boldsymbol{\varphi}_n\}$. The question is whether the set $F := \operatorname{conv}\{\eta_j \boldsymbol{\varphi}_j : j \in J\}$ is an $(s-1)$-dimensional face of E. We have the statements

$$\begin{aligned} s/n < \psi_{\ell_1}^{-1}(m/n) - o(1) \quad &\text{implies} \quad F \text{ is likely to be an } (s-1)\text{-dimensional face of } E; \\ s/n > \psi_{\ell_1}^{-1}(m/n) + o(1) \quad &\text{implies} \quad F \text{ is unlikely to be an } (s-1)\text{-dimensional face of } E. \end{aligned}$$

See the paper [DT09b] for more discussion of the connection between the facial structure of polytopes and signal recovery. Some universality results of this type also appear in Bayati et al. [BLM15].

5.3.2. *Other Signal Processing Applications.* We often want to perform signal processing tasks on data after reducing its dimension. In this section, we have focused on the problem of reconstructing a sparse signal from random measurements. Here are some related problems:

- **Detection.** Does an observed signal consist of a template corrupted with noise? Or is it just noise?
- **Classification.** Does an observed signal belong to class A or to class B?

The literature contains many papers that propose methods for solving these problems after dimension reduction; for example, see [DDW+07]. The existing analysis is either qualitative or it assumes that the dimension reduction map is Gaussian. Our universality laws can be used to study the precise behavior of compressed detection and classification with more general types of random linear maps. For brevity, we omit the details.

6. Decoding with Structured Errors

One of the goals of coding theory is to design codes and decoding algorithms that can correct gross errors in transmission. In particular, it is common that some proportion of the received symbols are corrupted. In this section, we show that a large family of random codes can be decoded in the presence of structured errors. The number of errors that we can correct is universal over this family. This result is valuable because it applies to random codebooks that are closer to realistic coding schemes.

The result on random decoding can also be interpreted as a statement about the behavior of the least-absolute-deviation (LAD) method for regression. We also discuss how our universality results apply to a class of demixing problems.

6.1. **Decoding with Sparse Errors.** We work with a random linear code over the real field. Consider a fixed message $\boldsymbol{x}_\star \in \mathbb{R}^m$. Let $\mathbf{\Phi} \in \mathbb{R}^{n \times m}$ be a random matrix, which is called a *codebook* in this context. Instead of transmitting the original message $\boldsymbol{x}_\star$, we transmit the coded message $\mathbf{\Phi}\boldsymbol{x}_\star$. Suppose that we receive a version $\boldsymbol{y}$ of the message where some number s of the entries are corrupted. That is, $\boldsymbol{y} = \mathbf{\Phi}\boldsymbol{x}_\star + \boldsymbol{z}_\star$ where the error vector $\boldsymbol{z}_\star$ has at most s nonzero entries. For simplicity, we assume that $\boldsymbol{z}_\star$ does not depend on the codebook $\mathbf{\Phi}$.

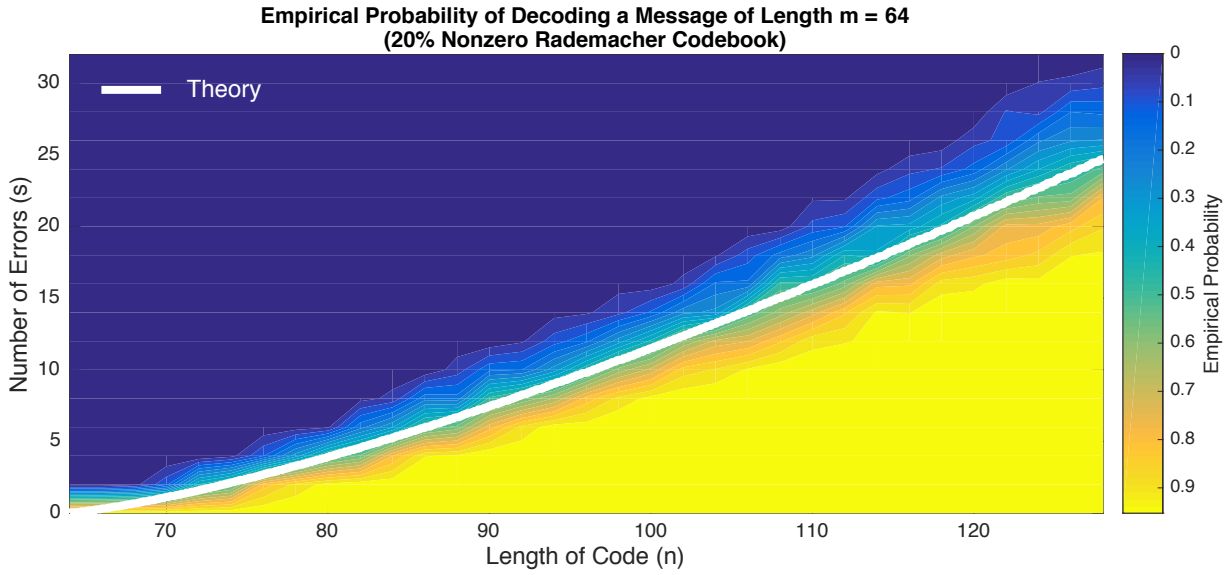

FIGURE 6.1: *Universality of the ℓ_1 Decoding Phase Transition.* This plot shows the empirical probability that the ℓ_1 method (6.1) decodes the message $\boldsymbol{x}_\star \in \mathbb{R}^{64}$ from the received message $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}_\star + \boldsymbol{z}_\star \in \mathbb{R}^n$ where the corruption $\boldsymbol{z}_\star \in \mathbb{R}^n$ has exactly s nonzero entries. The codebook $\boldsymbol{\Phi} \in \mathbb{R}^{n\times 64}$ is a sparse Rademacher matrix with an average of 20% nonzero entries. The **heatmap** gives the empirical probability of correct decoding, computed over 100 trials. The **white curve** is the exact phase transition $n\psi_{\ell_1}^{-1}(1-64/n)$ promised by Proposition 6.1. See Section 1.9 for more details.

In this setting, one can attempt to decode the message using an ℓ_1 minimization method [DH01, CRTV05, DT06, MT14b]. We solve the optimization problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^m, \boldsymbol{z}\in\mathbb{R}^n}{\text{minimize}} \quad \|\boldsymbol{z}\|_{\ell_1} \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{z}. \tag{6.1}$$

In other words, we search for a message $\boldsymbol{x}$ and a sparse corruption $\boldsymbol{z}$ that match the received data. We say that the optimization (6.1) *succeeds* if it has a unique optimal point $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{z}})$ that coincides with $(\boldsymbol{x}_\star, \boldsymbol{z}_\star)$; otherwise it *fails*.

The question is when the optimization problem (6.1) is effective at decoding the received transmission. That is, how many errors s can we correct as a function of the message length m and the code length n? The following result gives a solution to this problem for any codebook drawn from Model 2.4.

Proposition 6.1 (Universality of Sparse Error Correction)**.** *Assume that*

- *The message length m and the code length n satisfy $m \le n$.*
- *The message $\boldsymbol{x}_\star \in \mathbb{R}^m$ is arbitrary.*
- *The error vector $\boldsymbol{z}_\star \in \mathbb{R}^n$ has exactly s nonzero entries, where $s \le (1-\xi)n$ for some $\xi > 0$.*
- *The random codebook $\boldsymbol{\Phi} \in \mathbb{R}^{n\times m}$ follows Model 2.4 with parameters p and ν.*
- *We observe the vector $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}_\star + \boldsymbol{z}_\star$.*

Then, as the message length $m \to \infty$,

$$\begin{aligned} s/n < \psi_{\ell_1}^{-1}(1 - m/n - o(1)) \quad &\textit{implies} \quad (6.1) \textit{ succeeds with probability } 1 - o(1); \\ s/n > \psi_{\ell_1}^{-1}(1 - m/n + o(1)) \quad &\textit{implies} \quad (6.1) \textit{ fails with probability } 1 - o(1). \end{aligned}$$

The function ψ_{ℓ_1} is defined in (5.2). The little-o suppresses constants that depend only on ξ and p and ν.

This result is significant because it allows us to understand the behavior of this method for a sparse, discrete codebook. This type of code is somewhat closer to a practical coding mechanism than the ultra-random codebooks that have been studied in the past; see Remark 6.3. Figure 6.1 contains an illustration of how the theory compares with the actual performance of this coding scheme.

Proof. To analyze the decoding problem (6.1), we change variables: $\boldsymbol{u} := \boldsymbol{x} - \boldsymbol{x}_\star$ and $\boldsymbol{v} := \boldsymbol{z} - \boldsymbol{z}_\star$. We obtain the equivalent optimization problem

$$\operatorname*{minimize}_{\boldsymbol{u}\in\mathbb{R}^m, \boldsymbol{v}\in\mathbb{R}^n} \quad \|\boldsymbol{z}_\star + \boldsymbol{v}\|_{\ell_1} \quad \text{subject to} \quad \boldsymbol{\Phi}\boldsymbol{u} = \boldsymbol{v}. \tag{6.2}$$

The decoding procedure (6.1) succeeds if and only if the unique optimal point $(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}})$ of the problem (6.2) is the pair $(\mathbf{0}, \mathbf{0})$.

Introduce the set Ω of unit-norm descent directions of the ℓ_1 norm at $\boldsymbol{z}_\star$:

$$\Omega := \big\{\boldsymbol{v} \in \mathsf{S}^{n-1} : \|\boldsymbol{z}_\star + \lambda\boldsymbol{v}\|_{\ell_1} \le \|\boldsymbol{z}_\star\|_{\ell_1} \text{ for some } \lambda > 0\big\}.$$

The primal optimality condition for (6.2) shows that decoding succeeds if and only if $\Omega \cap \operatorname{range}(\boldsymbol{\Phi}) = \varnothing$.

Let us compute the statistical dimension of Ω°:

$$\delta(\Omega^\circ) = n - \delta(\Omega) = n - (n\psi_{\ell_1}(s/n) + o(n)) = n(1 - \psi_{\ell_1}(s/n) + o(1)). \tag{6.3}$$

The first relation is the polar identity (3.5) for the statistical dimension, and the value of the statistical dimension appears in (5.4). Since $s \le (1-\xi)n$, the properties of the function ψ_{ℓ_1} ensure that $\delta(\Omega^\circ) \ge \varrho n$ for some $\varrho > 0$.

First, we demonstrate that decoding fails when the number s of errors is too large. To do so, we must show that $\Omega \cap \operatorname{range}(\boldsymbol{\Phi}) \ne \varnothing$. By polarity [Kle55, Thm. (2.7)], it suffices to check that

$$\Omega^\circ \cap \operatorname{null}(\boldsymbol{\Phi}^*) = \varnothing.$$

With probability at least $1 - o(1)$, this relation follows from Theorem I(a), provided that

$$m > \delta(\Omega^\circ) + o(n) = n(1 - \psi_{\ell_1}(s/n)) + o(n).$$

Finally, revert the inequality so that it is expressed in terms of s.

Last, we must check that decoding succeeds when the number s of errors is sufficiently small. To do so, we must verify that $\Omega \cap \operatorname{range}(\boldsymbol{\Phi}) = \varnothing$ with high probability. This relation follows from ideas closely related to the proof of Theorem I, but it is not a direct consequence. See Section 17.3 for the details. □

Remark 6.2 (Least-Absolute-Deviation Regression)**.** Proposition 6.1 can also be interpreted as a statement about the performance of the least-absolute deviation method for fitting models with outliers. Suppose that we observe the data $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_\star + \boldsymbol{z}$. Each of the n rows of $\boldsymbol{X}$ is interpreted as a vector of m measured variables for an independent subject in an experiment. The vector $\boldsymbol{\beta}_\star$ lists the coefficients in the true linear model, and the sparse vector $\boldsymbol{z}$ contains a small number s of arbitrary statistical errors. The least-absolute deviation method fits a model by solving

$$\operatorname*{minimize}_{\boldsymbol{\beta}\in\mathbb{R}^m} \quad \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_{\ell_1}.$$

Proposition 6.1 shows that the procedure identifies the true model $\boldsymbol{\beta}_\star$ exactly, provided that the number s of contaminated data points satisfies $s/n < \psi_{\ell_1}^{-1}(1 - m/n) - o(1)$.

Remark 6.3 (Prior Work)**.** The idea of using ℓ_1 minimization for decoding in the presence of sparse errors dates at least as far back as the paper [DH01]. This scheme received further attention in the work [CRTV05]. Later, Donoho & Tanner [DT06] applied phase transition calculations to assess the precise performance of this coding scheme for a standard normal codebook; the least-absolute-deviation interpretation of this result appears in [DT09a, Sec. 1.3]. The paper [MT14b] revisits the coding problem and develops a sharp analysis in the case where the codebook is a random orthogonal matrix. The current paper contains the first precise result that extends to codebooks with more general distributions.

6.2. **Other Demixing Problems.** The decoding problem (6.1) is an example of a convex demixing problem [MT14b, ALMT14]. Our universality results can be used to study other questions of this species.

Let $f_0 : \mathbb{R}^n \to \mathbb{R}$ and $f_1 : \mathbb{R}^n \to \mathbb{R}$ be proper convex functions that measure the complexity of a signal. Suppose that $\boldsymbol{x}_\star^0 \in \mathbb{R}^n$ and $\boldsymbol{x}_\star^1 \in \mathbb{R}^n$ are signals with "low complexity." Draw random matrices $\boldsymbol{\Phi}_0 : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{\Phi}_1 : \mathbb{R}^n \to \mathbb{R}^m$ from Model 2.4. Suppose that we observe the vector $\boldsymbol{y} = \boldsymbol{\Phi}_0 \boldsymbol{x}_\star^1 + \boldsymbol{\Phi}_1 \boldsymbol{x}_\star^1$. We interpret the random matrices as known transformations of the unknown signals. For example, the matrices $\boldsymbol{\Phi}_i$ might denote dictionaries in which the two components of $\boldsymbol{y}$ are sparse.

We can attempt to reconstruct the original signal pair by solving

$$\underset{\boldsymbol{z}^0 \in \mathbb{R}^n, \boldsymbol{z}^1 \in \mathbb{R}^n}{\text{minimize}} \quad \max\{f_0(\boldsymbol{z}^0), f_1(\boldsymbol{z}^1)\} \quad \text{subject to} \quad \boldsymbol{\Phi}_0 \boldsymbol{z}^0 + \boldsymbol{\Phi}_1 \boldsymbol{z}^1 = \boldsymbol{y}. \tag{6.4}$$

In other words, we witness a superposition of two structured signals, and we attempt to find the lowest complexity pair $(\boldsymbol{z}^0, \boldsymbol{z}^1)$ that reproduces the observed data. The demixing problem *succeeds* if it has a unique optimal point that equals $(\boldsymbol{x}_\star^0, \boldsymbol{x}_\star^1)$.

To analyze this problem, we introduce two descent sets:

$$\Omega_i := \left\{ \boldsymbol{u} \in \mathsf{S}^{n-1} : f_i(\boldsymbol{x}_\star^i + \lambda \boldsymbol{u}) \le f_i(\boldsymbol{x}_\star^i) \text{ for some } \lambda > 0 \right\} \quad \text{for } i = 1, 2.$$

Up to scaling, the descent directions of $\max\{f_0(\cdot), f_1(\cdot\cdot)\}$ at the pair $(\boldsymbol{x}_\star^0, \boldsymbol{x}_\star^1)$ coincide with the direct product $\Omega_0 \times \Omega_1$. The statistical dimension of a direct product of two spherical sets satisfies $\delta(\Omega_0 \times \Omega_1) = \delta(\Omega_0) + \delta(\Omega_1)$. Therefore, Theorem I demonstrates that

$$\begin{aligned} m < \delta(\Omega_0) + \delta(\Omega_1) - o(n) \quad &\text{implies} \quad (6.4) \text{ fails with probability } 1 - o(1); \\ m > \delta(\Omega_0) + \delta(\Omega_1) + o(n) \quad &\text{implies} \quad (6.4) \text{ succeeds with probability } 1 - o(1). \end{aligned}$$

In other words, the amount of information needed to extract a pair of signals from the superposition equals the total complexity of the two signals. This result holds true for a wide class of distributions on $\boldsymbol{\Phi}_0$ and $\boldsymbol{\Phi}_1$.

7. Randomized Numerical Linear Algebra

Numerical linear algebra (NLA) is the study of computational methods for problems in linear algebra, including the solution of linear systems, spectral calculations, and matrix approximations. Over the last 15 years, researchers have developed many new algorithms for NLA that exploit randomness to perform these computations more efficiently. See the surveys [Mah11, HMT11, Woo14] for an overview of this field.

In this section, we apply our universality techniques to obtain new results on dimension reduction in randomized NLA. This discussion shows that a broad class of dimension reduction methods share the same quantitative behavior. Therefore, within some limits, we can choose the random linear map that is most computationally appealing when we design numerical algorithms based on dimension reduction.

As an added bonus, the arguments here lead to a new proof of the Bai–Yin limit for the minimum singular value of a random matrix drawn from Model 2.4.

7.1. **Subspace Embeddings.** In randomized NLA, one of the key primitives is a *subspace embedding*. A subspace embedding is nothing more than a randomized linear map that does not annihilate any point in a fixed subspace.

Definition 7.1 (Subspace Embedding)**.** Fix a natural number k, and let L be an arbitrary k-dimensional subspace. We say that a randomized linear map $\boldsymbol{\Pi} : \mathbb{R}^D \to \mathbb{R}^d$ is an *oblivious subspace embedding of order* k if

$$\boldsymbol{0} \notin \boldsymbol{\Pi}\big(L \cap \mathsf{S}^{D-1}\big) \quad \text{with high probability.}$$

The term "oblivious" indicates that the linear map $\boldsymbol{\Pi}$ is chosen without knowledge of the subspace L.

In the definition of a subspace embedding, some authors include quantitative bounds on the stability of the embedding. These estimates are useful for analyzing certain algorithms, but we have left them out because they are not essential.

A standard normal matrix provides an important theoretical and practical example of a subspace embedding.

Example 7.2 (Gaussian Subspace Embedding)**.** For any natural number k, a standard normal matrix $\mathbf{\Gamma} \in \mathbb{R}^{d\times D}$ is a subspace embedding with probability one when the embedding dimension $d \geq k$. In practice, it is preferable to select the embedding dimension $d \geq k+10$ to ensure that the restricted singular value $\sigma_{\min}(\mathbf{\Gamma}; L \cap \mathsf{S}^{D-1})$ is sufficiently positive, which makes the embedding more stable. See [HMT11] for more details.

A Gaussian subspace embedding has superb dimension reduction properties. On the other hand, standard normal matrices are expensive to generate, to store, and to perform arithmetic with. Therefore, in most randomized NLA algorithms, it is better to use subspace embeddings that are discrete or sparse.

Our universality results demonstrate that, in a certain range of parameters, every matrix that follows Model 2.4 enjoys the same subspace embedding properties as a Gaussian matrix.

Proposition 7.3 (Universality for Subspace Embedding)**.** *Suppose that*

- *The ambient dimension D is sufficiently large.*
- *The embedding dimension satisfies $d \leq D^{6/5}$.*
- *The random linear map $\mathbf{\Pi} : \mathbb{R}^D \to \mathbb{R}^d$ follows Model 2.4 with parameters p and ν.*

Then, for each k-dimensional subspace L of $\mathbb{R}^D$,

$$\sigma_{\min}(\mathbf{\Pi}; L \cap \mathsf{S}^{D-1}) \geq \sqrt{d} - \sqrt{k} - o(\sqrt{D}) \quad \text{with probability } 1 - o(1).$$

In particular, $\mathbf{\Pi}$ is a subspace embedding of order k whenever $d \geq k + o(D)$. In these expressions, the little-o suppresses constants that depend only on p and ν.

Proof. Proposition 7.3 is a consequence of Theorem II(a) and (4.3) because

$$\mathscr{E}_d(L \cap \mathsf{S}^{D-1}) \geq \sqrt{d} - \sqrt{\delta(L \cap \mathsf{S}^{D-1})} = \sqrt{d} - \sqrt{\delta(L)} = \sqrt{d} - \sqrt{k}.$$

The last identity holds because the k-dimensional subspace L has statistical dimension k. We introduce the error term $o\big(\sqrt{D}\big)$ to make sure that the stated result is only valid when the hypotheses of Theorem II are in force. □

Note that Proposition 7.3 applies to a sparse Rademacher linear map with a fixed, but arbitrarily small, proportion of nonzero entries. This particular example has received extensive attention in recent years [CW13, NN13, KN14, BDN15], although these works typically focus on the regime where the subspace dimension k is small and the sparsity level of the random linear map is a vanishing proportion of the embedding dimension d.

Remark 7.4 (Prior Work)**.** For the simple problem considered in Proposition 7.3, much sharper results are available in the random matrix literature. See the paper [KY14] for a recent analysis, as well as additional references.

Remark 7.5 (The Bai–Yin Limit for the Minimum Singular Value)**.** One of the most important problems in random matrix theory is to obtain bounds for the extreme singular values of a random matrix. The Bai–Yin law [BY93] gives a near-optimal result in case the entries of the random matrix are independent and standardized. We can reproduce a slightly weaker version of the Bai–Yin law for the minimum singular value by modifying the proof of Proposition 7.3.

Fix an aspect ratio $\varrho \in (0,1)$. For each natural number d, define $k := k(d) := \lfloor \varrho d \rfloor$. Draw a $d \times k$ random matrix $\mathbf{\Phi}^{(d)}$ from Model 2.4 with fixed parameters p and ν. For each $\varepsilon > 0$, we can apply Theorem II(a) with $E = \mathsf{S}^{k-1}$ to see that

$$\mathbb{P}\left\{ d^{-1/2} \sigma_{\min}\big(\mathbf{\Phi}^{(d)}\big) \geq 1 - \sqrt{\varrho} - \varepsilon \right\} \to 1 \quad \text{as } d \to \infty.$$

Here, $\sigma_{\min}$ denotes the kth largest singular value of $\mathbf{\Phi}^{(d)}$.

Under these assumptions, it is known [Yin86] that the empirical distribution of the singular values of $\mathbf{\Phi}^{(d)}$ converges in probability to the Marčenko–Pastur density, whose support is the interval $1 \pm \sqrt{\varrho}$. It follows that

$$\mathbb{P}\left\{ d^{-1/2}\sigma_{\min}\big(\mathbf{\Phi}^{(d)}\big) \le 1 - \sqrt{\varrho} + \varepsilon \right\} \to 1 \quad \text{as } d \to \infty.$$

Therefore, we may conclude that

$$d^{-1/2}\sigma_{\min}\big(\mathbf{\Phi}^{(d)}\big) \to 1 - \sqrt{\varrho} \quad \text{in probability.}$$

In comparison, the Bai–Yin law [BY93, Thm. 2] gives the same conclusion almost surely when the entries of $\mathbf{\Phi}^{(d)}$ have four finite moments. See the recent paper [Tik15] for an optimal result.

7.2. **Sketching and Least Squares.** In randomized NLA, one of the core applications of dimension reduction is to solve over-determined least-squares problems, perhaps with additional constraints. This idea is attributed to Sarlós [Sar06], and it has been studied intensively over the last decade; see the surveys [Mah11, Woo14] for more information. In this section, we develop sharp bounds for the simplest version of this approach.

Suppose that $\boldsymbol{A}$ is a fixed $D \times n$ matrix with full column rank. Let $\boldsymbol{y} \in \mathbb{R}^D$ be a vector, and consider the over-determined least-squares problem

$$\operatorname*{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2. \tag{7.1}$$

This problem can be expensive to solve when $D \gg n$. One remedy is to apply dimension reduction. Draw a random linear map $\mathbf{\Pi} : \mathbb{R}^D \to \mathbb{R}^d$ from Model 2.4, and consider the compressed problem

$$\operatorname*{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} \quad \|\mathbf{\Pi}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})\|^2. \tag{7.2}$$

The question is how the quality of the solution of (7.2) depends on the embedding dimension d. The following result provides an optimal estimate.

Proposition 7.6 (Randomized Least Squares: Error Bound)**.** *Instate the prevailing notation. Fix parameters $\lambda \in (0,1)$ and $\varrho \in (0,1)$ and $\iota \in (0,1)$. Assume that*

- *The number D of constraints is sufficiently large as a function of the parameters.*
- *The embedding dimension d is comparable with the number D of constraints: $\lambda D \le d \le D$.*
- *The embedding dimension d is somewhat larger than the number n of variables: $d \ge (1+\varrho)n$.*

With high probability, the solution $\widehat{\boldsymbol{x}}$ to the reduced least-squares problem (7.2) satisfies

$$\frac{\|\boldsymbol{A}(\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star)\|^2}{\|\boldsymbol{A}\boldsymbol{x}_\star - \boldsymbol{y}\|^2} \le \frac{n + \iota d}{d - n}, \tag{7.3}$$

where $\boldsymbol{x}_\star$ is the solution to the original least-squares problem (7.1). In particular,

$$\frac{\|\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{y}\|^2}{\|\boldsymbol{A}\boldsymbol{x}_\star - \boldsymbol{y}\|^2} \le (1+\iota)\frac{d}{d-n}$$

In other words, the excess error $\|\boldsymbol{A}(\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star)\|$ incurred in solving the compressed least-squares problem (7.2) is negligible as compared with the optimal value of the least-squares problem (7.1) if we choose the embedding dimension d sufficiently large. Proposition 7.6 improves substantially on the most recent work [PW15, Cor. 2(a)], both in terms of the error bound and in terms of the assumptions on the randomized linear map.

Let us remark that there is nothing special about ordinary least squares. We can also solve least-squares problems with a convex constraint set by dimension reduction. For this class of problems, we can also obtain optimal bounds by adapting the argument below. For example, see the results in Section 8.

Proof. Let $\boldsymbol{x}_\star \in \mathbb{R}^n$ be the solution to the original least-squares problem (7.1). Define the optimal residual $\boldsymbol{z}_\star := \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\star \in \mathbb{R}^D$, and recall that $\boldsymbol{z}_\star$ is orthogonal to range$(\boldsymbol{A})$. Moreover,

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 = \|\boldsymbol{A}\boldsymbol{x}_\star - \boldsymbol{y}\|^2 = \|\boldsymbol{z}_\star\|^2.$$

Without loss of generality, we may scale the problem so that $\|\boldsymbol{z}_\star\|^2 = 1$.

Next, change variables. Define $\boldsymbol{w} := \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_\star)$, and note that $\boldsymbol{w}$ is orthogonal to $\boldsymbol{z}_\star$. We can write the reduced least-squares problem (7.2) as

$$\operatorname*{minimize}_{\boldsymbol{w}\in\operatorname{range}(\boldsymbol{A})} \quad \|\boldsymbol{\Pi}(\boldsymbol{w} - \boldsymbol{z}_\star)\|. \tag{7.4}$$

When dimension reduction is effective, we expect the solution $\widehat{\boldsymbol{w}}$ to (7.4) to be close to zero.

Since $d \geq (1+\rho)n$, we can use the fact (Proposition 7.3) that $\boldsymbol{\Pi}$ is a subspace embedding to obtain an a priori bound $\|\widehat{\boldsymbol{w}}\| \leq R_\infty$ that holds with high probability. The number R_∞ is a constant that depends on nothing but ε. We only need this observation to ensure that we are optimizing over a compact set with constant radius, so we omit the details.

Next, we invoke Theorem II(b) to bound the optimal value of the reduced least-squares problem (7.4). Define the compact, convex set

$$E := \{\boldsymbol{w} \in \operatorname{range}(\boldsymbol{A}) : \|\boldsymbol{w}\| \leq R_\infty\}.$$

Let $\varepsilon > 0$ be a parameter that will depend on the parameter ι. With high probability,

$$\min_{\boldsymbol{w}\in\operatorname{range}(\boldsymbol{A})} \|\boldsymbol{\Pi}(\boldsymbol{w} - \boldsymbol{z}_\star)\| = \min_{\boldsymbol{w}\in E} \|\boldsymbol{\Pi}(\boldsymbol{w} - \boldsymbol{z}_\star)\| \leq (1+\varepsilon)\, \mathscr{E}_d(E - \boldsymbol{z}_\star).$$

By direct calculation, we can bound the excess width above. Let $\boldsymbol{P}$ be the orthogonal projector onto the range of $\boldsymbol{A}$, and observe that

$$\begin{aligned}
\mathscr{E}_d(E - \boldsymbol{z}_\star) &= \mathbb{E} \inf_{\boldsymbol{w}\in E} \left(\sqrt{d}\,\|\boldsymbol{w} - \boldsymbol{z}_\star\| + \boldsymbol{g}\cdot(\boldsymbol{w} - \boldsymbol{z}_\star)\right) \\
&= \mathbb{E} \inf_{\boldsymbol{w}\in E} \left(\sqrt{d}(\|\boldsymbol{w}\|^2 + 1)^{1/2} + \boldsymbol{g}\cdot\boldsymbol{w}\right) \\
&= \mathbb{E} \inf_{\boldsymbol{w}\in E} \left(\sqrt{d}(\|\boldsymbol{w}\|^2 + 1)^{1/2} - \|\boldsymbol{P}\boldsymbol{g}\|\,\|\boldsymbol{w}\|\right) \\
&\leq \inf_{0\leq\alpha\leq R_\infty} \left(\sqrt{d}(\alpha^2+1)^{1/2} - \sqrt{n}\alpha\right) \\
&= \sqrt{d-n}.
\end{aligned}$$

The first line is Definition 4.2, of the excess width. Next, simplify via the orthogonality of $\boldsymbol{w}$ and $\boldsymbol{z}_\star$ and the scaling $\|\boldsymbol{z}_\star\|^2 = 1$. Use translation invariance of the infimum to remove $\boldsymbol{z}_\star$ from the Gaussian term. Apply Jensen's inequality to draw the expectation inside the infimum, and note that $\mathbb{E}\,\|\boldsymbol{P}\boldsymbol{g}\| \leq \sqrt{n}$ because rank$(\boldsymbol{P}) = n$. Then make the change of variables $\alpha = \|\boldsymbol{w}\|$, and solve the scalar convex optimization problem. The infimum occurs at the value

$$\alpha_{\mathrm{opt}}^2 := \frac{n}{d-n}.$$

Since $d \geq (1+\rho)n$, we may be confident that $\alpha_{\mathrm{opt}} < R_\infty$.

We have shown that, with high probability,

$$\min_{\boldsymbol{w}\in\operatorname{range}(\boldsymbol{A})} \|\boldsymbol{\Pi}(\boldsymbol{w} - \boldsymbol{z}_\star)\| \leq (1+\varepsilon)\sqrt{d-n}. \tag{7.5}$$

Furthermore, we have evidence that the norm of the minimizer $\|\widehat{\boldsymbol{w}}\| \approx \alpha_{\mathrm{opt}}$. To prove the main result, we compute the value of the optimization problem (7.4) restricted to points with $\|\boldsymbol{w}\| \geq \alpha_{\mathrm{opt}}\sqrt{1+\iota_0}$, where ι_0 is a small positive number to be chosen later. Then we verify that the optimal value of the restricted problem is usually larger than the bound (7.5) for the optimal value of (7.2). This argument implies that $\|\widehat{\boldsymbol{w}}\| < \alpha_{\mathrm{opt}}\sqrt{1+\iota_0}$ with high probability.

To that end, define $R_+ := \alpha_{\mathrm{opt}}\sqrt{1+\iota_0}$, and introduce the compact set

$$E_+ := \{\boldsymbol{w} \in \operatorname{range}(\boldsymbol{A}) : R_+ \leq \|\boldsymbol{w}\| \leq R_\infty\}.$$

Theorem II(a) shows that, with high probability,

$$\min_{\boldsymbol{w}\in E_+} \|\boldsymbol{\Pi}(\boldsymbol{w}-\boldsymbol{z}_\star)\| \geq (1-\varepsilon)\,\mathscr{E}_d(E_+ - \boldsymbol{z}_\star). \tag{7.6}$$

Calculating the excess width as before,

$$\begin{aligned}\mathscr{E}_d(E_+ - \boldsymbol{z}_\star) &= \mathbb{E} \inf_{R_+\leq\alpha\leq R_\infty} \left(\sqrt{d}(\alpha^2+1)^{1/2} - \|\boldsymbol{P}\boldsymbol{g}\|\,\alpha\right) \\ &\geq \inf_{R_+\leq\alpha\leq R_\infty} \left(\sqrt{d}(\alpha^2+1)^{1/2} - \sqrt{n}\alpha\right) - R_\infty\,\mathbb{E}\left(\|\boldsymbol{P}\boldsymbol{g}\| - \sqrt{n}\right)_+ \\ &\geq \left(\sqrt{d}(R_+^2+1)^{1/2} - \sqrt{n}R_+\right) - R_\infty.\end{aligned} \tag{7.7}$$

Add and subtract $\sqrt{n}\alpha$ to reach the second line, and use the fact that $\mathbb{E}\|\boldsymbol{P}\boldsymbol{g}\| \leq \sqrt{n}$. Then apply the Gaussian variance inequality, Fact A.1, to bound the expectation by one. The infimum occurs at R_+ because the objective is convex and R_+ exceeds the unconstrained minimizer α_{opt}.

Next, we simplify the expression involving R_+. Setting $\iota_0 := \iota d/n$, we find that

$$\begin{aligned}\sqrt{d}(R_+^2+1)^{1/2} - \sqrt{n}R_+ &= \frac{d\sqrt{1+\iota_0 n/d} - n\sqrt{1+\iota_0}}{\sqrt{d-n}} \\ &= \sqrt{\frac{1+\iota}{d-n}}\left(d - n\sqrt{1+\frac{\iota(d/n-1)}{1+\iota}}\right) \\ &\geq \sqrt{\frac{1+\iota}{d-n}}\left(d - n - \frac{\iota(d-n)}{2(1+\iota)}\right) = \frac{1+\iota/2}{\sqrt{1+\iota}}\sqrt{d-n}.\end{aligned} \tag{7.8}$$

The inequality follows from the linear upper bound for the square root at one.

Combine (7.6), (7.7), and (7.8) to arrive at

$$\min_{\boldsymbol{w}\in E_+} \|\boldsymbol{\Pi}(\boldsymbol{w}-\boldsymbol{z}_\star)\| \geq \frac{(1-\varepsilon)(1+\iota/2)}{\sqrt{1+\iota}}\sqrt{d-n}.$$

Comparing (7.5) with the last display, we discover that the choice $\varepsilon := c\iota^2$ is sufficient to ensure that, with high probability,

$$\min_{\boldsymbol{w}\in E_+} \|\boldsymbol{\Pi}(\boldsymbol{w}-\boldsymbol{z}_\star)\| > \min_{\boldsymbol{w}\in E} \|\boldsymbol{\Pi}(\boldsymbol{w}-\boldsymbol{z}_\star)\|\,.$$

It follows that the minimum of (7.4) usually occurs on the set $E \setminus E_+$. We determine that

$$\|\widehat{\boldsymbol{w}}\|^2 \leq (1+\iota d/n)\,\alpha_{\mathrm{opt}}^2.$$

Reinterpret this inequality to obtain the stated result (7.3).

We can obtain a matching lower bound for $\|\widehat{\boldsymbol{w}}\|$ by considering the set of vectors $E_- := \{\boldsymbol{w} \in \mathrm{range}(\boldsymbol{A}) : \|\boldsymbol{w}\| \leq R_-\}$ where $R_- := \alpha_{\mathrm{opt}}\sqrt{1-\iota_0}$. We omit the details. □

Remark 7.7 (Prior Work)**.** The idea of using random linear maps to accelerate the solution of least-squares problems appears in the work of Sarlós [Sar06]. This approach has been extended and refined in the literature on randomized NLA; see the surveys [Mah11, Woo14] for an overview. Most of this research is concerned with randomized linear maps that have favorable computational properties, but the results are much less precise than Proposition 7.6. Recently, Pilanci & Wainright [PW15] have offered a more refined analysis of randomized dimension reduction for constrained least-squares problems, but it still falls short of describing the actual behavior of these methods. Parts of the argument here are adapted from the work of Oymak, Thrampoulidis, and Hassibi [OTH13, TOH15].

8. The Prediction Error for LASSO

Universality results have always played an important role in statistics. The most fundamental example is the law of large numbers, which justifies the use of the sample average to estimate the mean of a general distribution. Similarly, the central limit theorem permits us to build a confidence interval for the mean of a distribution.

High-dimensional statistics relies on more sophisticated methods, often based on optimization, to estimate population parameters. In particular, applied statisticians frequently employ the LASSO estimator [Tib96] to perform regression and variable selection in linear models. It is only recently that researchers have developed theory [KM11, BM12, OTH13, DJM13, JM14, TAH15] that can predict the precise behavior of the LASSO when the data are assumed to be Gaussian. It is a critical methodological challenge to develop universality results that expand the range of models in which we can make confident assertions about the performance of the LASSO.

In this section, we prove the first general universality result for the prediction error using a LASSO model estimate. This theory offers a justification for using a LASSO model to make predictions when the data derives from a sparse model. We expect that further developments in this direction will play an important role in applied statistics.

8.1. The Sparse Linear Model and the LASSO.

The LASSO is designed to perform simultaneous regression and variable selection in a linear model. Let us present a simple statistical model in which to study the behavior of the LASSO estimator. Suppose that the random variable Y takes the form

$$Y = \boldsymbol{x} \cdot \boldsymbol{\beta}_\star + \sigma Z \tag{8.1}$$

where

- The deterministic vector $\boldsymbol{\beta}_\star \in \mathbb{R}^p$ of model parameters has at most s nonzero entries.
- The random vector $\boldsymbol{x} \in \mathbb{R}^p$ of predictor variables is drawn from Model 2.4.
- The noise variance $\sigma^2 > 0$ is known.
- The statistical error Z is drawn from Model 2.4, independent of $\boldsymbol{x}$.

We interpret $\boldsymbol{x} = (X_1, \dots, X_p)$ as a family of predictor variables that we want to use to predict the value of the response variable Y. Only the variables X_j where $(\boldsymbol{\beta}_\star)_j \neq 0$ are relevant to the prediction, while the others are confounding. The observed value Y of the response is contaminated with a statistical error σZ. These assumptions are idealized, but let us emphasize that our analysis holds even when the predictors and the noise are heavy-tailed.

Suppose that we observe independent pairs $(\boldsymbol{x}_1, Y_1), \dots, (\boldsymbol{x}_n, Y_n)$ drawn from the model above, and let $\boldsymbol{z} \in \mathbb{R}^n$ be the unknown vector of statistical errors. One of the goals of sparse regression is to use this data to construct an estimate $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$ of the model coefficients so that we can predict future responses. That is, given a fresh random vector $\boldsymbol{x}_0$ of predictor variables, we can predict the (unknown) response $Y_0 = \boldsymbol{x}_0 \cdot \boldsymbol{\beta}_\star + \sigma Z_0$ using the linear estimate

$$\widehat{Y}_0 := \boldsymbol{x}_0 \cdot \widehat{\boldsymbol{\beta}}.$$

We want to control the *mean squared error in prediction*, which is defined as

$$\mathsf{MSEP} := \mathbb{E}\left[|\widehat{Y}_0 - Y_0|^2 \,|\, \boldsymbol{X}, \boldsymbol{z} \right]. \tag{8.2}$$

Using the statistical model (8.1), it is easy to verify that

$$\mathsf{MSEP} = \mathbb{E}\left[|\boldsymbol{x}_0 \cdot (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star) + \sigma Z_0|^2 \,|\, \boldsymbol{X}, \boldsymbol{z} \right] = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\star\|^2 + \sigma^2. \tag{8.3}$$

In other words, the prediction error is controlled by the squared error in estimating the model coefficients.

The LASSO uses convex optimization to produce an estimate $\widehat{\boldsymbol{\beta}}$ of the model coefficients. The estimator is chosen arbitrarily from the set of solutions to the problem

$$\operatorname*{minimize}_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_{\ell_1} \leq \|\boldsymbol{\beta}_\star\|_{\ell_1}. \tag{8.4}$$

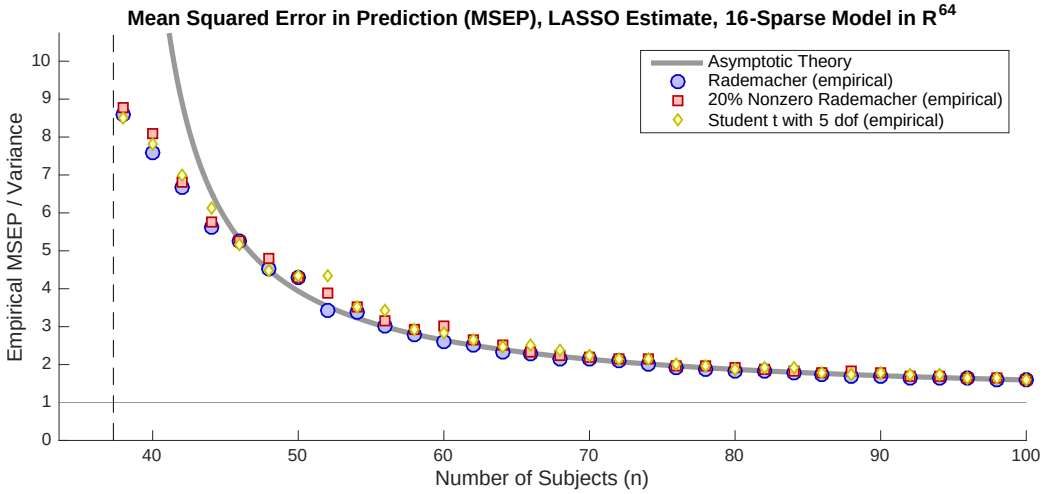

FIGURE 8.1: *Universality of the LASSO Prediction Error.* This plot shows the MSEP (8.2) obtained with the LASSO estimator (8.4), averaged over design matrices $\boldsymbol{X}$ and statistical errors $\boldsymbol{z}$. In the linear model (8.1), the number of predictors $p = 64$; the number of active predictors $s = 16$; each nonzero coefficient $(\boldsymbol{\beta}_\star)_j$ in the model has unit magnitude; the statistical error Z is Gaussian; the variance $\sigma^2 = 1$; and the number n of subjects varies. The **dashed line** marks the location of the phase transition for the number of subjects required to identify the model exactly when the noise variance is zero. The **gray curve** delineates the asymptotic upper bound $n/(n - p\psi_{\ell_1}(s/p))$ for the normalized MSEP from Proposition 8.1. The **markers** give an empirical estimate (over 100 trials) for the MSEP when the design matrix $\boldsymbol{X}$ has the specified distribution.

In this formula, the rows of the $n \times p$ matrix $\boldsymbol{X}$ are the observed predictor vectors $\boldsymbol{x}_i$. The entries of the vector $\boldsymbol{y} \in \mathbb{R}^n$ are the measured responses. For simplicity, we also assume that we have the exact side information $\|\boldsymbol{\beta}_\star\|_{\ell_1}$.

We can prove the following result on the squared error in the LASSO estimate of the sparse coefficient model. This is the first universal statement that offers a precise analysis for this class of statistical models.

Proposition 8.1 (Universality of LASSO Prediction Error)**.** *Instate the prevailing notation. Choose parameters $\lambda \in (0,1)$ and $\varrho \in (0,1)$ and $\iota \in (0,1)$. Assume that*

- *The number n of subjects is sufficiently large as a function of the parameters.*
- *The number p of predictors satisfies $\lambda p \le n \le p^{6/5}$.*
- *The number n of subjects satisfies $n \ge (1+\varrho)\, p\psi_{\ell_1}(s/p)$.*

With high probability over the observed data, the mean squared error in prediction (8.2) satisfies

$$\mathsf{MSEP} \le (1+\iota)\frac{\sigma^2 n}{n - p\psi_{\ell_1}(s/p)}. \tag{8.5}$$

The function ψ_{ℓ_1} is defined in (5.2). Furthermore, the bound (8.5) is sharp when $n \gg p\psi_{\ell_1}(s/p)$ or when $\sigma^2 \to 0$.

Proposition 8.1 gives an upper bound for the MSEP, which matches the low-noise limit ($\sigma \to 0$) obtained in the Gaussian case [OTH13]. See Figure 8.1 for a numerical experiment that confirms our theoretical predictions. The proof of the result appears below in Section 8.2.

The assumptions in Proposition 8.1 are somewhat restrictive, in that the number n of subjects must be roughly comparable with the number p of predictors. This condition can probably be relaxed, but the error bound in Theorem II does not allow for a stronger conclusion. The argument can also be extended to give even more precise formulas for the MSEP under the same assumptions. We also note that there is nothing special about the ℓ_1 constraint in (8.4); similar results are valid for many other convex constraints.

8.2. **Proof of Proposition 8.1.** Without loss of generality, we may assume that the statistical model is scaled so the noise level $\sigma = 1$. Define the sublevel set E of the ℓ_1 norm at the true parameter vector $\boldsymbol{\beta}_\star$:

$$E := \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{\beta}_\star + \boldsymbol{u}\|_{\ell_1} \le \|\boldsymbol{\beta}_\star\|_{\ell_1}\}.$$

Note that E is a compact, convex set that contains the origin. Define the $n \times (p+1)$ random matrix $\boldsymbol{\Phi} := \begin{bmatrix}\boldsymbol{X} & \boldsymbol{z}\end{bmatrix}$, and note that $\boldsymbol{\Phi}$ also follows Model 2.4. Making the change of variables $\boldsymbol{u} := \boldsymbol{\beta} - \boldsymbol{\beta}_\star$, we can rewrite the LASSO problem (8.4) in the form

$$\operatorname*{minimize}_{\boldsymbol{u} \in E} \quad \left\| \boldsymbol{\Phi} \begin{bmatrix} \boldsymbol{u} \\ -1 \end{bmatrix} \right\|. \tag{8.6}$$

This expression depends on the assumption that $\sigma = 1$. Let $\widehat{\boldsymbol{u}}$ be an optimal point of the problem (8.6). Referring to (8.3), we see that a formula for $\|\widehat{\boldsymbol{u}}\|^2$ leads to a formula for the MSEP.

It will be helpful to introduce some additional sets. For a parameter $\alpha > 0$, define the compact (but typically nonconvex) set

$$E_\alpha := \{\boldsymbol{u} \in E : \|\boldsymbol{u}\| = \alpha\}.$$

We also define the compact and convex set

$$E_{\le\alpha} := \{\boldsymbol{u} \in E : \|\boldsymbol{u}\| \le \alpha\}.$$

Observe that $E_\alpha \subset E_{\le\alpha}$. Furthermore,

$$\alpha \le \alpha_+ \quad \text{implies} \quad (1/\alpha_+)E_{\alpha_+} \subset (1/\alpha)E_\alpha. \tag{8.7}$$

The inclusion (8.7) holds because E is convex and contains the origin.

Let R_∞ be a constant that depends only on the parameters ρ and ι. Suppose that $0 \le R \le R_+ \le R_\infty$. To prove that $\|\widehat{\boldsymbol{u}}\| < R_+$, it suffices to establish the inequality

$$\min_{\boldsymbol{u} \in E_{\le R}} \left\| \boldsymbol{\Phi} \begin{bmatrix} \boldsymbol{u} \\ -1 \end{bmatrix} \right\| < \min_{\boldsymbol{u} \in E_{R_+}} \left\| \boldsymbol{\Phi} \begin{bmatrix} \boldsymbol{u} \\ -1 \end{bmatrix} \right\|. \tag{8.8}$$

Indeed, recall that $\widehat{\boldsymbol{u}}$ is the minimizer of the objective over E, and let $\boldsymbol{u}_0$ be the point in $E_{\le R}$ where the left-hand minimum in (8.8) is attained. The objective is a convex function of $\boldsymbol{u}$, so it does not decrease as $\boldsymbol{u}$ traverses the line segment from $\widehat{\boldsymbol{u}}$ to $\boldsymbol{u}_0$. If $\|\widehat{\boldsymbol{u}}\| \ge R_+$, this line segment muss pass through E_{R_+}, which is impossible because the ordering (8.8) forces the objective to decrease on the way from E_{R_+} to $\boldsymbol{u}_0$.

Fix the parameter R in the range $0 \le R \le R_\infty$; we will select a suitable value later. Theorem II(b) demonstrates that, with high probability,

$$\min_{\boldsymbol{u} \in E_{\le R}} \left\| \boldsymbol{\Phi} \begin{bmatrix} \boldsymbol{u} \\ -1 \end{bmatrix} \right\| \le \mathscr{E}_n\big(E_{\le R} \times \{-1\}\big) + o(\sqrt{n}) \le \mathscr{E}_n\big(E_R \times \{-1\}\big) + o(\sqrt{n}).$$

The second relation holds because the excess width decreases with respect to set inclusion. Observe that

$$\begin{aligned} \mathscr{E}_n\big(E_R \times \{-1\}\big) &= \mathbb{E} \inf_{\boldsymbol{t} \in E_R} \big(\sqrt{n}(\|\boldsymbol{t}\|^2 + 1)^{1/2} + \boldsymbol{g} \cdot \boldsymbol{t}\big) \\ &= \sqrt{n}(R^2 + 1)^{1/2} - \mathbb{E} \sup_{\boldsymbol{t} \in E_R} \boldsymbol{g} \cdot \boldsymbol{t} \\ &= \sqrt{n}(R^2 + 1)^{1/2} - R\mathscr{W}\big((1/R)E_R\big). \end{aligned}$$

The last identity holds when we factor out $\|\boldsymbol{t}\|$ and identify the Gaussian width (3.7).

Fix the second parameter R_+, such that $R \le R_+ \le R_\infty$. Theorem II(a) demonstrates that, with high probability,

$$\min_{\boldsymbol{u} \in E_{R_+}} \left\| \boldsymbol{\Phi} \begin{bmatrix} \boldsymbol{u} \\ -1 \end{bmatrix} \right\| \ge \mathscr{E}_n\big(E_{R_+} \times \{-1\}\big) - o(\sqrt{n}).$$

Much as before, we calculate the excess width:

$$\begin{aligned} \mathscr{E}_n\big(E_{R_+} \times \{-1\}\big) &= \sqrt{n}(R_+^2 + 1)^{1/2} - R_+\mathscr{W}\big((1/R_+)E_{R_+}\big) \\ &\ge \sqrt{n}(R_+^2 + 1)^{1/2} - R_+\mathscr{W}\big((1/R)E_R\big). \end{aligned}$$

The last inequality holds because of (8.7) and the fact that the Gaussian width is increasing with respect to set inclusion.

Combine the last four displays to reach

$$\min_{\boldsymbol{u}\in E_{\le R}} \left\| \boldsymbol{\Phi} \begin{bmatrix} \boldsymbol{u} \\ -1 \end{bmatrix} \right\| \le \sqrt{n}(R^2+1)^{1/2} - R\,\mathscr{W}\big((1/R)E_R\big) + o(\sqrt{n})$$
$$\le \sqrt{n}(R_+^2+1)^{1/2} - R_+\,\mathscr{W}\big((1/R)E_R\big) - o(\sqrt{n}) \le \min_{\boldsymbol{u}\in E_{R_+}} \left\| \boldsymbol{\Phi} \begin{bmatrix} \boldsymbol{u} \\ -1 \end{bmatrix} \right\|.$$

It follows that we can establish (8.8) by finding parameters for which $0 \le R \le R_+ \le R_\infty$ and

$$\big[\sqrt{n}(R_+^2+1)^{1/2} - R_+\,\mathscr{W}\big((1/R)E_R\big)\big] - \big[\sqrt{n}(R^2+1)^{1/2} - R\,\mathscr{W}\big((1/R)E_R\big)\big] \ge o(\sqrt{n}).$$

To that end, we replace the Gaussian width by a number that does not depend on the parameter R:

$$\mathscr{W}^2\big((1/R)E_R\big) \le \delta\big((1/R)E_R\big) \le \delta\big(\operatorname{cone}(E)\big) \le p\psi_{\ell_1}(s/p) =: d.$$

The first relation is (3.8); the second follows from Definition 3.4; the last is the estimate (5.4). Moreover, these bounds are sharp when R is sufficiently close to zero. Therefore, we just need to verify that

$$\left[\sqrt{n}(R_+^2+1)^{1/2} - R_+\sqrt{d}\right] - \left[\sqrt{n}(R^2+1)^{1/2} - R\sqrt{d}\right] \ge o(\sqrt{n}). \tag{8.9}$$

Once we choose R and R_+ appropriately, we can adapt the analysis in the proof of Proposition 7.6.

For $\alpha \ge 0$, introduce the function

$$f(\alpha) := \sqrt{n}(\alpha^2+1)^{1/2} - \alpha\sqrt{d}.$$

As in the proof of Proposition 7.6, by direct calculation, f is minimized at the value

$$R := \frac{d}{n-d}.$$

Note that R is very close to zero when $n \gg d$, in which case d is an accurate bound for $\mathscr{W}^2\big((1/R)E_R\big)$. Furthermore,

$$f(R) = \sqrt{n-d}. \tag{8.10}$$

Now, make the selection

$$R_+^2 := (1+\iota n/d)R^2 = \frac{d+\iota n}{n-d}. \tag{8.11}$$

Since $n \ge (1+\varrho)\,d$, we see that R_+ is bounded by a constant R_∞ that depends only on ι and ϱ. Repeating the calculation in (7.8), *mutatis mutandis*, we have

$$f(R_+) \ge \frac{1+\iota/2}{\sqrt{1+\iota}}\sqrt{n-d}. \tag{8.12}$$

Combining (8.10) and (8.12), we determine that

$$f(R_+) - f(R) \ge \frac{1+\iota/2-\sqrt{1+\iota}}{\sqrt{1+\iota}}\sqrt{n-d} \ge o(\sqrt{n}).$$

The last inequality holds because ι is a fixed positive constant and we have assumed that $n \ge (1+\varrho)d$.

In conclusion, we have confirmed the claim (8.9) for R and the value R_+ designated in (8.11). It follows that (8.8) holds with high probability, and so the optimizer of (8.6) satisfies $\|\widehat{\boldsymbol{u}}\| \le R_+$ with high probability. Therefore, the formula (8.2) for the mean squared error yields the bound

$$\mathsf{MSEP} = \|\widehat{\boldsymbol{u}}\|^2 + 1 \le R_+^2 + 1 = (1+\iota)\frac{n}{n-d}.$$

Once again, we have used the assumption that $\sigma = 1$. By homogeneity, the MSEP must be proportional to σ^2, which leads to the stated result (8.5). Finally, if we allow $\sigma \to 0$ with the other parameters fixed, the analysis here can be adapted to show that the error bound (8.5) is sharp.

Remark 8.2 (Prior Work)**.** In the special case of a standard normal design $\boldsymbol{X}$ and a standard normal error $\boldsymbol{z}$, the result of Proposition 8.1 appeared in the paper [OTH13]. Our extension to more general random models is new. Nevertheless, our proof has a lot in common with the analysis in [OTH13, TAH15].

Part III. Universality of the Restricted Minimum Singular Value: Proofs of Theorem II and Theorem I(a)

In this part of the paper, we present a detailed proof of the universality law for the restricted singular value of a random matrix, Theorem II, and the first part of the universality law for the embedding dimension, Theorem I(a).

Section 9 contains our main technical result, which establishes universality for the bounded random matrix model, Model 2.1. Section 10 extends the result for bounded random matrices to the heavy-tailed random matrix model, Model 2.4. In Section 10.2, we obtain Theorem II as an immediate consequence of the result for heavy-tailed matrices. Section 10.3 shows how to derive Theorem I(a) as an additional consequence. The remaining sections in this part lay out the calculations that undergird the result for bounded random matrices.

9. The Restricted Singular Values of a Bounded Random Matrix

The key challenges in establishing universality for the restricted minimum singular value are already present in the case where the random matrix is drawn from the bounded model, Model 2.1. This section presents a universality law for bounded random matrices, and it gives an overview of the calculations that are required to establish this result.

9.1. Theorem 9.1: Main Result for the Bounded Random Matrix Model.

The main technical result in this paper is a theorem on the behavior of the restricted minimum singular value of a bounded random matrix.

Theorem 9.1 (RSV: Bounded Random Matrix Model)**.** *Place the following assumptions:*

- *Let m and n be natural numbers with $m \le n^{6/5}$.*
- *Let T be a closed subset of the unit ball B^n in $\mathbb{R}^n$.*
- *Draw an $m \times n$ random matrix $\mathbf{\Phi}$ from Model 2.1 with bound B.*

Then the squared restricted singular value $\sigma^2_{\min}(\mathbf{\Phi}; T)$ has the following properties:

(1) The squared restricted singular value concentrates about its mean on a scale of $B^2\sqrt{m+n}$. For each $\zeta \ge 0$,

$$\mathbb{P}\left\{\sigma^2_{\min}(\mathbf{\Phi}; T) \le \mathbb{E}\,\sigma^2_{\min}(\mathbf{\Phi}; T) - \mathrm{C}B^2\zeta\right\} \le \mathrm{e}^{-\zeta^2/m}, \quad \text{and}$$
$$\mathbb{P}\left\{\sigma^2_{\min}(\mathbf{\Phi}; T) \ge \mathbb{E}\,\sigma^2_{\min}(\mathbf{\Phi}; T) + \mathrm{C}B^2\zeta\right\} \le \mathrm{C}\mathrm{e}^{-\zeta^2/(m+\zeta\sqrt{n})}.$$

(2) The expectation of the squared restricted singular value is bounded below in terms of the excess width:

$$\mathbb{E}\,\sigma^2_{\min}(\mathbf{\Phi}; T) \ge \left(\mathscr{E}_m(T)\right)_+^2 - \mathrm{C}B^2(m+n)^{0.92}.$$

(3) If T is a convex set, the squared restricted singular value is bounded above in terms of the excess width:

$$\mathbb{E}\,\sigma^2_{\min}(\mathbf{\Phi}; T) \le \left(\mathscr{E}_m(T)\right)_+^2 + \mathrm{C}B^4(m+n)^{0.94}.$$

Furthermore, the entries of $\mathbf{\Phi}$ need not be symmetric for this result to hold.

The proof of this result will occupy us for the rest of this part of the paper. This section summarizes the required calculations, with cross-references to the detailed arguments.

9.2. Proof of Theorem 9.1(1): Concentration.

Theorem 9.1(1) states that the squared restricted singular value $\sigma^2_{\min}(\mathbf{\Phi}; T)$ concentrates around its mean. We prove this claim in Proposition 11.1, which appears below. The argument depends on some concentration inequalities that are derived using the entropy method. This approach is more or less standard, so we move on to the more interesting part of the proof.

9.3. **Setup for Proof of Theorem 9.1(2) and (3): Dissection of the Index Set.** Let us continue with the proof of Theorem 9.1, conclusions (2) and (3). The overall approach is the same in both cases, but some of the details differ.

The first step is to dissect the index set T into appropriate subsets. For each set $J \subset \{1,\dots,n\}$, we introduce a closed subset T_J of T via the rule

$$T_J := \big\{ \boldsymbol{t} \in T : |t_j| \le (\#J)^{-1/2} \text{ for all } j \in J^c \big\}, \tag{9.1}$$

where $J^c := \{1,\dots,n\} \setminus J$. In other words, T_J contains all the vectors in T where the coordinates listed in J^c are sufficiently small. Note that convexity of the set T implies convexity of each subset T_J.

Fix a number $k \in \{1,\dots,n\}$. Since T is a subset of the unit ball B^n, every vector $\boldsymbol{t} \in T$ satisfies

$$\#\big\{ j : |t_j| > k^{-1/2} \big\} \le k.$$

Therefore, $\boldsymbol{t}$ belongs to some subset T_J where the cardinality $\#J = k$, and we have the decomposition

$$T = \bigcup\nolimits_{\#J=k} T_J. \tag{9.2}$$

It is clear that the number of subsets T_J in this decomposition satisfies

$$\#\{T_J : \#J = k\} = \binom{n}{k} \le \left(\frac{\mathrm{e}n}{k}\right)^k. \tag{9.3}$$

We must limit the cardinality k of the subsets, so we can control the number of subsets we need to examine.

9.4. **Proof of Theorem 9.1(2): Lower Bound for the RSV.** Let us proceed with the proof of Theorem 9.1(2), the lower bound for the RSV. For the time being, we fix the parameter k that designates the cardinality of the sets J. We require that $k \ge m^{2/3}$.

First, we pass from the restricted singular value $\sigma_{\min}(\boldsymbol{\Phi}; T)$ over the whole set T to a bound that depends on $\sigma_{\min}(\boldsymbol{\Phi}; T_J)$ for the subsets T_J. Proposition 12.2 gives the comparison

$$\mathbb{E}\,\sigma_{\min}^2(\boldsymbol{\Phi}; T) \ge \min_{\#J=k} \mathbb{E}\,\sigma_{\min}^2(\boldsymbol{\Phi}; T_J) - CB^2\sqrt{km\log(n/k)}. \tag{9.4}$$

We have used the decomposition (9.2) and the bound (9.3) on the number of subsets in the decomposition to invoke the proposition. The main ingredient in the proof of this estimate is the concentration inequality for restricted singular values, Proposition 11.1.

Let us focus on a specific subset T_J. To study the restricted singular value $\sigma_{\min}(\boldsymbol{\Phi}; T_J)$, we want to replace the entries of the random matrix $\boldsymbol{\Phi}$ with standard normal random variables. Proposition 13.1 allows us to make partial progress toward this goal. Let $\boldsymbol{\Gamma}$ be an $m \times n$ standard normal matrix, independent from $\boldsymbol{\Phi}$. Define an $m \times n$ random matrix $\boldsymbol{\Psi} := \boldsymbol{\Psi}(J)$ where $\boldsymbol{\Psi}_J = \boldsymbol{\Phi}_J$ and $\boldsymbol{\Psi}_{J^c} = \boldsymbol{\Gamma}_{J^c}$. Then

$$\mathbb{E}\,\sigma_{\min}^2(\boldsymbol{\Phi}; T_J) \ge \mathbb{E}\,\sigma_{\min}^2(\boldsymbol{\Psi}; T_J) - \frac{CB^2 m^{1/3} n\log(mn)}{k^{1/2}}. \tag{9.5}$$

This bound requires the assumption that $k \ge m^{2/3}$. The argument is based on the Lindeberg exchange principle, but we have used this method in an unusual way. For a vector $\boldsymbol{t} \in T_J$, the definition (9.1) gives us control on the magnitude of the entries listed in J^c, which we can exploit to replace the column submatrix $\boldsymbol{\Phi}_{J^c}$ with $\boldsymbol{\Gamma}_{J^c}$. The coordinates of $\boldsymbol{t}$ listed in J may be large, so we cannot replace the column submatrix $\boldsymbol{\Phi}_J$ without incurring a significant penalty. Instead, we just leave it in place.

Next, we want to compare the expected RSV on the right-hand side of (9.5) with the excess width of the set T_J. Proposition 14.1 provides the bound

$$\mathbb{E}\,\sigma_{\min}^2(\boldsymbol{\Psi}; T_J) \ge \Big(\mathscr{E}_m(T_J) - CB^2\sqrt{k} \Big)_+^2 \ge \Big(\mathscr{E}_m(T) - CB^2\sqrt{k} \Big)_+^2. \tag{9.6}$$

The second inequality is a consequence of the facts that $T_J \subset T$ and that the excess width is decreasing with respect to set inclusion. The calculation in Proposition 14.1 uses the Gaussian Minimax Theorem (see

Fact A.3) to simplify the average with respect to the standard normal matrix $\mathbf{\Gamma}$. We also invoke standard results from nonasymptotic random matrix theory to control the expectation over $\mathbf{\Phi}_J$.

We can obtain a simple lower bound on the last term in (9.6) by linearizing the convex function $(\cdot)_+^2$ at the point $\mathscr{E}_m(T)$:

$$\left(\mathscr{E}_m(T) - CB^2\sqrt{k}\right)_+^2 \geq \left(\mathscr{E}_m(T)\right)_+^2 - CB^2\sqrt{k}\left(\mathscr{E}_m(T)\right)_+ \geq \left(\mathscr{E}_m(T)\right)_+^2 - CB^2\sqrt{km}. \tag{9.7}$$

The last estimate follows from the observation that

$$\mathscr{E}_m(T) = \mathbb{E}\inf_{\boldsymbol{t}\in T}\left(\sqrt{m}\,\|\boldsymbol{t}\| + \boldsymbol{g}\cdot\boldsymbol{t}\right) \leq \mathbb{E}\left(\sqrt{m}\,\|\boldsymbol{t}_0\| + \boldsymbol{g}\cdot\boldsymbol{t}_0\right) \leq \sqrt{m}. \tag{9.8}$$

In this calculation, $\boldsymbol{t}_0$ is an arbitrary point in T.

Finally, we sequence the four displays (9.4), (9.5), (9.6), and (9.7) and combine error terms to obtain

$$\mathbb{E}\,\sigma_{\min}^2(\mathbf{\Phi}; T) \geq \left(\mathscr{E}_m(T)\right)_+^2 - \frac{CB^2 m^{1/3} n\log(mn)}{k^{1/2}} - CB^2\sqrt{km\log(n/k)}.$$

Up to logarithms, an optimal choice of the cardinality parameter is $k = \lceil m^{-1/6}n\rceil$. Since $m \leq n^{6/5}$, this choice ensures that $k \geq m^{2/3}$. We conclude that

$$\begin{aligned}\mathbb{E}\,\sigma_{\min}^2(\mathbf{\Phi}; T) &\geq \left(\mathscr{E}_m(T)\right)_+^2 - CB^2 m^{5/12} n^{1/2}\log(mn)\\ &\geq \left(\mathscr{E}_m(T)\right)_+^2 - CB^2(m+n)^{11/12}\log(m+n).\end{aligned}$$

Combine the logarithm with the power, and adjust the constants to complete the proof of Theorem 9.1(2).

9.5. **Proof of Theorem 9.1(3): Upper Bound for the RSV of a Convex Set.** At a high level, the steps in the proof of Theorem 9.1(3) are similar with the argument in the last section. Many of the technical details, however, depend on convex duality arguments.

As before, we fix the cardinality parameter k, with the requirement $k \geq m^{2/3}$. The first step is to pass from $\sigma_{\min}(\mathbf{\Phi}; T)$ to $\sigma_{\min}(\mathbf{\Phi}; T_J)$. We have

$$\mathbb{E}\,\sigma_{\min}^2(\mathbf{\Phi}; T) \leq \min_{\#J=k}\mathbb{E}\,\sigma_{\min}^2(\mathbf{\Phi}; T_J). \tag{9.9}$$

This bound is a trivial consequence of the facts that $T_J \subset T$ for each subset J of $\{1,\dots,n\}$ and that $\sigma_{\min}(\mathbf{\Phi};\cdot)$ is decreasing with respect to set inclusion.

Select any subset T_J. We invoke Proposition 13.1 to exchange most of the entries of the random matrix $\mathbf{\Phi}$ for standard normal variables. Using the same random matrix $\mathbf{\Psi}$ from the last section, we have

$$\mathbb{E}\,\sigma_{\min}^2(\mathbf{\Phi}; T_J) \leq \mathbb{E}\,\sigma_{\min}^2(\mathbf{\Psi}; T_J) + \frac{CB^2 m^{1/3} n\log(mn)}{k^{1/2}}. \tag{9.10}$$

The bound (9.10) requires the assumption $k \geq m^{2/3}$, and the proof is identical with the proof of the lower bound (9.5).

Next, we compare the expected RSV on the right-hand side of (9.10) with the excess width. Proposition 14.1 delivers

$$\mathbb{E}\,\sigma_{\min}^2(\mathbf{\Psi}; T_J) \leq \left(\mathscr{E}_m(T_J) + CB^2\sqrt{k}\right)_+^2. \tag{9.11}$$

This argument is more complicated than the analogous lower bound (9.6), and it depends on the convexity of T_J. We also apply the assumption that $k \geq m^{2/3}$ here.

Proposition 12.1 allows us to replace the excess width of T_J in (9.11) with the excess width of T:

$$\mathscr{E}_m(T_J) \leq \mathscr{E}_m(T) + C\sqrt{k\log(n/k)}. \tag{9.12}$$

We use the decomposition (9.2) and the bound (9.3) on the number of sets T_J to invoke the proposition. This proof depends on the Gaussian concentration inequality.

Sequence the bounds (9.9), (9.10), (9.11), and (9.12), and combine the error terms to arrive at

$$\mathbb{E}\,\sigma_{\min}^2(\mathbf{\Phi}; T) \leq \left(\mathscr{E}_m(T) + CB^2\sqrt{k\log(n/k)}\right)_+^2 + \frac{CB^2 m^{1/3} n\log(mn)}{k^{1/2}}.$$

Expand the square using (9.8) to reach

$$\mathbb{E}\,\sigma_{\min}^2(\mathbf{\Phi}; T) \le \big(\mathscr{E}_m(T)\big)_+^2 + CB^2\left(\frac{m^{1/3}n\log(mn)}{k^{1/2}} + \sqrt{km\log(n/k)} + B^2 k\log(n/k)\right).$$

Select $k = \lceil n^{4/5}\rceil$ to arrive at

$$\begin{aligned}\mathbb{E}\,\sigma_{\min}^2(\mathbf{\Phi}; T) &\le \big(\mathscr{E}_m(T)\big)_+^2 + CB^2(m^{1/3}n^{3/5} + n^{2/5}m^{1/2})\log(n) + CB^4 n^{4/5}\log(mn)\\ &\le \big(\mathscr{E}_m(T)\big)_+^2 + CB^4(m+n)^{14/15}\log(m+n).\end{aligned}$$

Combine the power with the logarithm, and adjust constants to finish the proof of Theorem 9.1(3).

10. The Restricted Singular Values of a Heavy-Tailed Random Matrix

In this section, we present an extension of Theorem 9.1 to the heavy-tailed matrix model, Model 2.4. In Section 10.2, we explain how the universality result for the restricted singular value, Theorem II, is an immediate consequence. In Section 10.3, we show how to derive the first half of the universality result for the embedding dimension, Theorem I(a).

10.1. Corollary 10.1: Main Result for the p-Moment Random Matrix Model.

We can extend Theorem 9.1 to the heavy-tailed random matrix model, Model 2.4 using a truncation argument. The following corollary contains a detailed statement of the result.

Corollary 10.1 (RSV: p-Moment Random Matrix Model)**.** *Fix parameters $p > 4$ and $\nu \ge 1$. Place the following assumptions:*

- *Let m and n be natural numbers with $m \le n^{6/5}$.*
- *Let T be a closed subset of the unit ball B^n in $\mathbb{R}^n$.*
- *Draw an $m \times n$ random matrix $\mathbf{\Phi}$ that satisfies Model 2.4 with given p and ν.*

Then the restricted singular value $\sigma_{\min}(\mathbf{\Phi}; T)$ has the following properties:

(1) With high probability, the restricted singular value is bounded below by the excess width:

$$\mathbb{P}\left\{\sigma_{\min}(\mathbf{\Phi}; T) \le \big(\mathscr{E}_m(T)\big)_+ - C_p\nu(m+n)^{1/2-\kappa(p)}\right\} \le C_p(m+n)^{1-p/4}. \tag{10.1}$$

(2) If T is a convex set, with high probability, the restricted singular value is bounded above by the excess width:

$$\mathbb{P}\left\{\sigma_{\min}(\mathbf{\Phi}; T) \ge \big(\mathscr{E}_m(T)\big)_+ + C_p\nu^2(m+n)^{1/2-\kappa(p)}\right\} \le C_p(m+n)^{1-p/4}. \tag{10.2}$$

The function $\kappa(p)$ is strictly positive for $p > 4$, and the constant C_p depends only on p.

The proof of Corollary 10.1 appears below in Section 15. The main idea is to truncate each entry of the heavy-tailed random matrix $\mathbf{\Phi}$ individually. We can treat the bounded part of the random matrix using Theorem 9.1. We show that the tails are negligible by means of a relatively simple norm bound for random matrices, Fact B.2.

10.2. Proof of Theorem II from Corollary 10.1.

Theorem II is an easy consequence of Corollary 10.1. Recall the assumptions of the theorem:

- The embedding dimension satisfies $\lambda D \le d \le D^{6/5}$.
- E is a closed subset of the unit ball in $\mathbb{R}^D$.
- The d-excess width of E satisfies $\mathscr{E}_d(E) \ge \varrho\sqrt{d}$.
- The $d \times D$ random linear map $\mathbf{\Pi}$ follows Model 2.4 with parameters $p > 4$ and $\nu \ge 1$.

Therefore, we may apply Corollary 10.1 with $\mathbf{\Phi} = \mathbf{\Pi}$ and $T = E$ to see that

$$\begin{aligned}&\mathbb{P}\left\{\sigma_{\min}(\mathbf{\Pi}; E) \le \big(\mathscr{E}_d(E)\big)_+ - C_p\nu(d+D)^{1/2-\kappa(p)}\right\} \le C_p D^{1-p/4}, \quad \text{and}\\ &\mathbb{P}\left\{\sigma_{\min}(\mathbf{\Pi}; E) \ge \big(\mathscr{E}_d(E)\big)_+ + C_p\nu^2(d+D)^{1/2-\kappa(p)}\right\} \le C_p D^{1-p/4} \quad \text{if } E \text{ is convex.}\end{aligned}$$

For simplicity, we have dropped the embedding dimension d from the right-hand side of the bounds in the last display.

To prove Theorem II, it suffices to check that we can make the error term $C_p \nu^2 (d+D)^{1/2-\kappa(p)}$ smaller than $\varepsilon \mathscr{E}_d(E)$ if we select the ambient dimension D large enough. Indeed, the conditions $D \le \lambda^{-1} d$ and $d^{1/2} \le \varrho^{-1} \mathscr{E}_d(E)$ ensure that

$$\begin{aligned}(d+D)^{1/2-\kappa(p)} &\le (1+\lambda^{-1})^{1/2} d^{1/2-\kappa(p)} \\ &\le (1+\lambda^{-1})^{1/2} \varrho^{-1} \mathscr{E}_d(E) d^{-\kappa(p)} \\ &\le \frac{(1+\lambda^{-1})^{1/2} \varrho^{-1} \lambda^{-\kappa(p)}}{D^{\kappa(p)}} \mathscr{E}_d(E)\end{aligned}$$

Since $\kappa(p)$ is positive, there is a number $N := N(p, \nu, \lambda, \varrho, \varepsilon)$ for which $D \ge N$ implies

$$C_p \nu^2 (d+D)^{1/2-\kappa(p)} \le \varepsilon \mathscr{E}_d(E).$$

This is what we needed to show.

10.3. **Proof of Theorem I(a) from Corollary 10.1.** Theorem I(a) is also an easy consequence of Corollary 10.1. Recall the assumptions of the theorem:

- E is a compact subset of $\mathbb{R}^D$ that does not contain the origin.
- The statistical dimension of E satisfies $\delta(E) \ge \varrho D$.
- The $d \times D$ random linear map $\mathbf{\Pi}$ follows Model 2.4 with parameters $p > 4$ and $\nu \ge 1$.

In this section, we consider the regime where the embedding dimension $d \ge (1+\varepsilon)\,\delta(E)$. We need to demonstrate that

$$\mathbb{P}\{\mathbf{0} \notin \mathbf{\Pi}(E)\} = \mathbb{P}\{E \cap \operatorname{null}(\mathbf{\Pi}) = \emptyset\} \ge 1 - C_p D^{1-p/4}. \tag{10.3}$$

We begin with a reduction to a specific choice of the embedding dimension d. Let $\mathbf{\Pi}_m$ be the $m \times D$ matrix formed from the first m rows of the random linear map $\mathbf{\Pi}$. The function $m \mapsto \mathbb{P}\{E \cap \operatorname{null}(\mathbf{\Pi}_m) = \emptyset\}$ is weakly increasing because $m \mapsto \operatorname{null}(\mathbf{\Pi}_m)$ is a decreasing sequence of sets. Therefore, it suffices to verify (10.3) in the case where $d = \lceil (1+\varepsilon)\,\delta(E) \rceil$. Note that $d \le 2D+1$ because the statistical dimension $\delta(E) \le D$ and $\varepsilon < 1$.

Introduce the spherical retraction $\Omega := \boldsymbol{\theta}(E)$. Proposition 3.8 and (3.2) demonstrate that

$$\sigma_{\min}(\mathbf{\Pi}; \Omega) > 0 \quad \text{implies} \quad \mathbf{0} \notin \mathbf{\Pi}(\Omega) \quad \text{implies} \quad \mathbf{0} \notin \mathbf{\Pi}(E).$$

Therefore, to check (10.3), it suffices to produce a high-probability lower bound on the restricted singular value $\sigma_{\min}(\mathbf{\Pi}; \Omega)$. With the choices $\mathbf{\Phi} = \mathbf{\Pi}$ and $T = \Omega$, Corollary 10.1 yields

$$\mathbb{P}\left\{\sigma_{\min}(\mathbf{\Pi}; \Omega) \ge \left(\mathscr{E}_d(\Omega)\right)_+ - C_p \nu (d+D)^{1/2-\kappa(p)}\right\} \ge 1 - C_p D^{1-p/4}.$$

To complete the proof, we need to verify that our hypotheses imply

$$\mathscr{E}_d(\Omega) > C_p \nu (d+D)^{1/2-\kappa(p)}. \tag{10.4}$$

This point follows from two relatively short calculations.

Since Ω is a subset of the unit sphere, we quickly compute ite excess width:

$$\begin{aligned}\mathscr{E}_d(\Omega) = \mathbb{E} \inf_{\boldsymbol{x} \in \Omega} \left(\sqrt{d}\,\|\boldsymbol{x}\| + \boldsymbol{g} \cdot \boldsymbol{x}\right) &\ge \sqrt{d} - \sqrt{\delta(\Omega)} \\ &= \sqrt{d} - \sqrt{\delta(E)} \ge \left(\sqrt{1+\varepsilon} - 1\right)\sqrt{\delta(E)}.\end{aligned} \tag{10.5}$$

The second identity holds because Ω is a subset of the unit sphere, and we have used the relation (4.3). Recall that the statistical dimension of a general set E is defined as $\delta(E) = \delta(\boldsymbol{\theta}(E))$, and then introduce the value $d = \lceil (1+\varepsilon)\,\delta(E) \rceil$ of the embedding dimension.

Meanwhile, we can bound the dimensional term in (10.4) above:

$$(d+D)^{1/2-\kappa(p)} \le C D^{1/2-\kappa(p)} \le C \varrho^{-1/2} D^{-\kappa(p)} \sqrt{\delta(E)}.$$

The first inequality holds because $d \le 2D+1$, and the second inequality uses the assumption $D \le \varrho^{-1}\delta(E)$. Since $\kappa(p)$ is positive, we can find a number $N := N(p,\nu,\varrho,\varepsilon)$ for which $D \ge N$ implies that

$$\mathrm{C}_p \nu (d+D)^{1/2-\kappa(p)} < \left(\sqrt{1+\varepsilon} - 1\right)\sqrt{\delta(E)}.$$

Combine the last display with (10.5) to determine that the claim (10.4) is valid.

11. Theorem 9.1: Concentration for Restricted Singular Values

In this section, we demonstrate that the restricted minimum singular value of a bounded random matrix concentrates about its mean value. This result yields Theorem 9.1(1).

Proposition 11.1 (Theorem 9.1: Concentration)**.** *Let S be a closed subset of B^n. Let $\mathbf{\Phi}$ be an $m \times n$ random matrix drawn from Model 2.1 with uniform bound B. For all $\zeta \ge 0$,*

$$\mathbb{P}\left\{\sigma_{\min}^2(\mathbf{\Phi};S) \le \mathbb{E}\,\sigma_{\min}^2(\mathbf{\Phi};S) - \mathrm{C}B^2\zeta\right\} \le \mathrm{e}^{-\zeta^2/m} \tag{11.1}$$

$$\mathbb{P}\left\{\sigma_{\min}^2(\mathbf{\Phi};S) \ge \mathbb{E}\,\sigma_{\min}^2(\mathbf{\Phi};S) + \mathrm{C}B^2\zeta\right\} \le \mathrm{C}\mathrm{e}^{-\zeta^2/(m+\zeta\sqrt{n})}. \tag{11.2}$$

The proof relies on some modern concentration inequalities that are derived using the entropy method. We establish the bound (11.1) on the lower tail in Section 11.1; the bound (11.2) on the upper tail appears in Section 11.2.

In several other parts of the paper, we rely on variants of Proposition 11.1 that follow from essentially the same arguments. We omit the details of these proofs to avoid repetition.

11.1. Proposition 11.1: The Lower Tail of the RSV.

First, we establish that the restricted singular value is unlikely to be much smaller than its mean. The proof depends on a lower tail inequality [BLM13, Thm. 6.27] derived from Massart's modified logarithmic Sobolev inequality [Mas00].

Fact 11.2 (Self-Bounded Random Variable: Lower Tail)**.** *Let $(X_1,\dots,X_p)$ be an independent sequence of real random variables, and let $(X_1',\dots,X_p')$ be an independent copy of this sequence. For a nonnegative function $f : \mathbb{R}^p \to \mathbb{R}_+$, define*

$$\begin{aligned} Y &:= \quad f(X_1,\dots,X_{i-1},X_i,X_{i+1},\dots,X_p), \quad \text{and} \\ Y_i' &:= \quad f(X_1,\dots,X_{i-1},X_i',X_{i+1},\dots,X_p) \quad \text{for } i = 1,\dots,p. \end{aligned}$$

Suppose that

$$V_- := \sum\nolimits_{i=1}^p \mathbb{E}'\left(Y_i' - Y\right)_+^2 \le aY + b \quad \text{where } a \ge 0.$$

Then

$$\mathbb{P}\{Y \le \mathbb{E}\,Y - \zeta\} \le \exp\left(\frac{-\zeta^2/4}{a\,\mathbb{E}\,Y + b}\right).$$

The operator $\mathbb{E}'$ integrates out the random variables marked with a prime.

With this fact at hand, we may derive the lower tail bound.

Proof of Proposition 11.1, Eqn. (11.1). Introduce the random variable

$$Y := \sigma_{\min}^2(\mathbf{\Phi};S) = \min_{\boldsymbol{s}\in S} \|\mathbf{\Phi}\boldsymbol{s}\|^2 = \min_{\boldsymbol{s}\in S} \sum\nolimits_{i=1}^m \left(\sum\nolimits_{j=1}^n \varphi_{ij} s_j\right)^2.$$

For each index pair (i,j), define a random matrix $\mathbf{\Phi}_{ij}'$ by replacing the (i,j) entry φ_{ij} of $\mathbf{\Phi}$ with an independent copy φ_{ij}'. Now, the random variable

$$Y_{ij}' := \sigma_{\min}^2\left(\mathbf{\Phi}_{ij}';S\right) = \min_{\boldsymbol{s}\in S} \|\mathbf{\Phi}_{ij}'\boldsymbol{s}\|^2.$$

We must evaluate the lower variance proxy

$$V_- := \sum\nolimits_{i=1}^m \sum\nolimits_{j=1}^n \mathbb{E}'\left(Y_{ij}' - Y\right)^2.$$

To that end, select a point $\boldsymbol{t} \in \arg\min_{\boldsymbol{s}\in S} \|\boldsymbol{\Phi}\boldsymbol{s}\|^2$. For each index pair (i, j),

$$Y'_{ij} - Y = \min_{\boldsymbol{s}\in S} \|\boldsymbol{\Phi}'_{ij}\boldsymbol{s}\|^2 - \min_{\boldsymbol{s}\in S} \|\boldsymbol{\Phi}\boldsymbol{s}\|^2 \le \|\boldsymbol{\Phi}'_{ij}\boldsymbol{t}\|^2 - \|\boldsymbol{\Phi}\boldsymbol{t}\|^2.$$

The matrix $\boldsymbol{\Phi}'_{ij}$ differs from $\boldsymbol{\Phi}$ only in the ith row. Therefore, when we expand the squared norms into components, all of the components cancel except for the ith one. We discover that

$$\begin{aligned} Y'_{ij} - Y &\le \big((\varphi'_{ij} - \varphi_{ij})\, t_j + \textstyle\sum_{\ell=1}^n \varphi_{i\ell}\, t_\ell\big)^2 - \big(\sum_{\ell=1}^n \varphi_{i\ell}\, t_\ell\big)^2 \\ &= (\varphi'_{ij} - \varphi_{ij})^2\, t_j^2 + 2(\varphi'_{ij} - \varphi_{ij})\, t_j \textstyle\sum_{\ell=1}^n \varphi_{i\ell}\, t_\ell \\ &\le (\varphi'_{ij} - \varphi_{ij})^2\, t_j^2 + 2\,|\varphi'_{ij} - \varphi_{ij}|\,|t_j|\,\big|\textstyle\sum_{\ell=1}^n \varphi_{i\ell}\, t_\ell\big| \\ &\le 4B^2 t_j^2 + 4B\,|t_j|\,\big|\textstyle\sum_{\ell=1}^n \varphi_{i\ell}\, t_\ell\big|. \end{aligned} \tag{11.3}$$

The last inequality holds because $|\varphi_{ij}| \le B$ and $|\varphi'_{ij}| \le B$. As a consequence,

$$\mathbb{E}'\big(Y'_{ij} - Y\big)_+^2 \le \Big(4B^2 t_j^2 + 4B\,|t_j|\,\big|\textstyle\sum_{\ell=1}^n \varphi_{i\ell}\, t_\ell\big|\Big)^2 \le 8B^4 t_j^4 + 8B^2 t_j^2 \big(\textstyle\sum_{\ell=1}^n \varphi_{i\ell}\, t_\ell\big)^2.$$

Sum over pairs (i, j) to arrive at

$$\begin{aligned} \textstyle\sum_{i=1}^m \sum_{j=1}^n \mathbb{E}'\big(Y'_{ij} - Y\big)_+^2 &\le \textstyle\sum_{i=1}^m \sum_{j=1}^n \Big(8B^4 t_j^4 + 8B^2 t_j^2 \big(\sum_{\ell=1}^n \varphi_{i\ell}\, t_\ell\big)^2\Big) \\ &\le 8B^4 m + 8B^2 \textstyle\sum_{i=1}^m \big(\sum_{\ell=1}^n \varphi_{i\ell}\, t_\ell\big)^2 \\ &= 8B^4 m + 8B^2\, \|\boldsymbol{\Phi}\boldsymbol{t}\|^2 \\ &= 8B^4 m + 8B^2 \min_{\boldsymbol{s}\in S} \|\boldsymbol{\Phi}\boldsymbol{s}\|^2. \end{aligned}$$

To reach the second line, we used the fact that $\|\boldsymbol{t}\|_{\ell_4} \le \|\boldsymbol{t}\| \le 1$. The last line depends on the definition of $\boldsymbol{t}$.

In summary, we have demonstrated that

$$V_- \le 8B^2 Y + 8B^4 m.$$

To apply Fact 11.2, we need a bound for the expectation of Y. Designate a point $\boldsymbol{s}_0 \in S$, and calculate that

$$\mathbb{E}\, Y \le \mathbb{E}\, \|\boldsymbol{\Phi}\boldsymbol{s}_0\|^2 = m\, \|\boldsymbol{s}_0\|^2 \le m.$$

The identity holds because $\boldsymbol{\Phi}$ has independent, standardized entries. Fact 11.2 ensures that

$$\mathbb{P}\{Y \le \mathbb{E}\, Y - \zeta\} \le \exp\left(\frac{-\zeta^2/4}{8B^2\, \mathbb{E}\, Y + 8B^4 m}\right) \le \exp\left(\frac{-\zeta^2}{64B^4 m}\right).$$

Rewrite this formula to complete the proof of (11.1). □

11.2. **Proposition 11.1: The Upper Tail of the RSV.** Next, we establish that the restricted singular value of a bounded random matrix is unlikely to be much larger than its mean. The proof depends on another tail inequality for self-bounded random variables. This result follows by combining the moment inequality [BLM13, Thm. 15.7] with a standard moment comparison argument.

Fact 11.3 (Self-Bounded Random Variable: Upper Tail)**.** *Let $(X_1, \dots, X_p)$ be an independent sequence of real random variables, and let $(X'_1, \dots, X'_p)$ be an independent copy of this sequence. For a nonnegative function $f : \mathbb{R}^p \to \mathbb{R}_+$, define*

$$\begin{aligned} Y &:= \quad f(X_1, \dots, X_{i-1}, X_i, X_{i+1}, \dots, X_p), \quad \text{and} \\ Y'_i &:= \quad f(X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots, X_p) \quad \text{for } i = 1, \dots, p. \end{aligned}$$

Suppose that $Y'_i - Y \le L$ uniformly for a fixed value L, and assume that

$$V_- := \textstyle\sum_{i=1}^p \mathbb{E}'\big(Y'_i - Y\big)_+^2 \le aY + b \quad \text{where } a \ge 0.$$

Then

$$\mathbb{P}\{Y \ge \mathbb{E}\, Y + \zeta\} \le \mathrm{C} \exp\left(\frac{-\mathrm{c}\zeta^2}{(a\,\mathbb{E}\, Y + b) + (a + L)\zeta}\right).$$

The operator $\mathbb{E}'$ integrates out the random variables marked with a prime.

Proof of Proposition 11.1, Eqn. (11.2). We proceed as in the proof of (11.1). It just remains to verify the uniform bound on $Y'_{ij} - Y$. Starting from (11.3), we calculate that

$$\begin{aligned} Y'_{ij} - Y &\le 4B^2 t_j^2 + 4B\,|t_j| \left|\textstyle\sum_{\ell=1}^n \varphi_{il}\, t_\ell\right| \\ &\le 4B^2 + 4B \left(\textstyle\sum_{\ell=1}^n \varphi_{i\ell}^2\right)^{1/2} \\ &\le 4B^2 + 4B^2\sqrt{n} \\ &\le 8B^2\sqrt{n}. \end{aligned}$$

This calculation relies on the fact that $\|\boldsymbol{t}\| \le 1$ and the entries of $\boldsymbol{\Phi}$ have magnitude bounded by B. Apply Fact 11.3 to complete the argument. □

12. Theorem 9.1: Probability Bounds for Dissections

In this section, we establish some results that compare the expectation of a minimum of random variables indexed by a set with the expectations of the minima indexed by subsets. These facts are easy consequences of concentration phenomena.

12.1. Dissection of the Excess Width.

First, we show that the excess width of a set can be related to the excess width of a collection of subsets.

Proposition 12.1 (Theorem 9.1: Dissection of Excess Width)**.** *Consider a closed subset T of the unit ball B^n that has been decomposed into a finite number of closed subsets T_J:*

$$T = \textstyle\bigcup_{J \in \mathscr{J}} T_J.$$

For each $m \ge 0$, it holds that

$$\Big(\mathbb{E} \min_{\boldsymbol{t} \in T} \big(\sqrt{m}\,\|\boldsymbol{t}\| + \boldsymbol{g}\cdot\boldsymbol{t}\big)\Big)_+ \ge \min_{J \in \mathscr{J}} \mathbb{E} \min_{\boldsymbol{t} \in T_J} \big(\sqrt{m}\,\|\boldsymbol{t}\| + \boldsymbol{g}\cdot\boldsymbol{t}\big) - \mathrm{C}\sqrt{\log(\#\,\mathscr{J})}.$$

In other words,

$$\min_{J \in \mathscr{J}} \big(\mathscr{E}_m(T_J)\big)_+ \le \big(\mathscr{E}_m(T)\big)_+ + \mathrm{C}\sqrt{\log(\#\,\mathscr{J})}.$$

Proof. Since $T = \bigcup_{J \in \mathscr{J}} T_J$, we can stratify the minimum over T:

$$\min_{\boldsymbol{t} \in T} \big(\sqrt{m}\,\|\boldsymbol{t}\| + \boldsymbol{g}\cdot\boldsymbol{t}\big) = \min_{J \in \mathscr{J}} \min_{\boldsymbol{t} \in T_J} \big(\sqrt{m}\,\|\boldsymbol{t}\| + \boldsymbol{g}\cdot\boldsymbol{t}\big).$$

We will obtain a lower tail bound for the minimum over T by combining lower tail bounds for each minimum over T_J. Then we integrate the tail probability to obtain the required expectation bound.

Each subset T_J is contained in B^n, so the map

$$\boldsymbol{g} \mapsto \min_{\boldsymbol{t} \in T_J} \big(\sqrt{m}\,\|\boldsymbol{t}\| + \boldsymbol{g}\cdot\boldsymbol{t}\big)$$

is 1-Lipschitz. The Gaussian concentration inequality, Fact A.2, provides a tail bound. For each $J \in \mathscr{J}$ and for each $\lambda \ge 0$,

$$\mathbb{P}\left\{\min_{\boldsymbol{t} \in T_J} \big(\sqrt{m}\,\|\boldsymbol{t}\| + \boldsymbol{g}\cdot\boldsymbol{t}\big) \le \mathbb{E} \min_{\boldsymbol{t} \in T_J} \big(\sqrt{m}\,\|\boldsymbol{t}\| + \boldsymbol{g}\cdot\boldsymbol{t}\big) - \lambda\right\} \le \mathrm{e}^{-\lambda^2/2}.$$

Apply the union bound over $J \in \mathscr{J}$ to see that, for all $\zeta \ge 0$,

$$\begin{aligned} \mathbb{P}\left\{\min_{\boldsymbol{t} \in T} \big(\sqrt{m}\,\|\boldsymbol{t}\| + \boldsymbol{g}\cdot\boldsymbol{t}\big) \le \min_{J \in \mathscr{J}} \mathbb{E} \min_{\boldsymbol{t} \in T_J} \big(\sqrt{m}\,\|\boldsymbol{t}\| + \boldsymbol{g}\cdot\boldsymbol{t}\big) - \sqrt{2\log(\#\,\mathscr{J})} - \zeta\right\} \\ \le (\#\,\mathscr{J})\,\mathrm{e}^{-(\sqrt{2\log(\#\,\mathscr{J})}+\zeta)^2/2} \le \mathrm{e}^{-\zeta^2/2}. \end{aligned}$$

Using the latter estimate, we quickly compute the excess width of T. Abbreviate

$$Y := \min_{\boldsymbol{t}\in T} \left(\sqrt{m}\,\|\boldsymbol{t}\| + \boldsymbol{g}\cdot\boldsymbol{t}\right) \quad \text{and} \quad \alpha := \min_{J\in\mathscr{J}} \mathbb{E} \min_{\boldsymbol{t}\in T_J} \left(\sqrt{m}\,\|\boldsymbol{t}\| + \boldsymbol{g}\cdot\boldsymbol{t}\right) - \sqrt{2\log(\#\mathscr{J})}.$$

If $\alpha \le 0$, the stated result is trivial, so we may assume that $\alpha > 0$. The integration by parts representation of the expectation yields

$$(\mathbb{E}\,Y)_+ \ge \int_0^\alpha \mathbb{P}\{Y > \zeta\}\,\mathrm{d}\zeta = \int_0^\alpha \mathbb{P}\{Y > \alpha - \zeta\}\,\mathrm{d}\zeta = \alpha - \int_0^\alpha \mathbb{P}\{Y \le \alpha - \zeta\}\,\mathrm{d}\zeta \ge \alpha - \int_0^\alpha \mathrm{e}^{-\zeta^2/2}\,\mathrm{d}\zeta > \alpha - \sqrt{\pi/2}.$$

Reintroduce the values of Y and α, and combine constants to complete the argument. □

12.2. **Dissection of the Restricted Singular Value.** Next, we show that the minimum singular value of a random matrix, restricted to a set, is controlled by the minimum singular value, restricted to subsets.

Proposition 12.2 (Theorem 9.1: Dissection of RSV)**.** *Consider a closed subset T of the unit ball B^n, and assume that it has been decomposed into a finite number of closed subsets T_J:*

$$T = \bigcup\nolimits_{J\in\mathscr{J}} T_J.$$

Let $\boldsymbol{\Phi}$ be an $m \times n$ random matrix that satisfies Model 2.1 with uniform bound B. Then

$$\min_{J\in\mathscr{J}} \mathbb{E}\,\sigma_{\min}^2(\boldsymbol{\Phi}; T_J) \le \mathbb{E}\,\sigma_{\min}^2(\boldsymbol{\Phi}; T) + \mathrm{C}B^2\sqrt{m\log(\#\mathscr{J})}.$$

Proof. The argument is similar with the proof of Proposition 12.1. Since $T = \bigcup_{J\in\mathscr{J}} T_J$,

$$\sigma_{\min}^2(\boldsymbol{\Phi}; T) = \min_{\boldsymbol{t}\in T} \|\boldsymbol{\Phi}\boldsymbol{t}\|^2 = \min_{J\in\mathscr{J}} \min_{\boldsymbol{t}\in T_J} \|\boldsymbol{\Phi}\boldsymbol{t}\|^2 = \min_{J\in\mathscr{J}} \sigma_{\min}^2(\boldsymbol{\Phi}; T_J).$$

The lower tail bound (11.1) for restricted singular values implies that, for all $\lambda \ge 0$,

$$\mathbb{P}\left\{\sigma_{\min}^2(\boldsymbol{\Phi}; T_J) \le \mathbb{E}\,\sigma_{\min}^2(\boldsymbol{\Phi}; T_J) - \lambda\right\} \le \mathrm{e}^{-\lambda^2/(\mathrm{C}B^4 m)}.$$

Next, we take a union bound. For all $\zeta \ge 0$,

$$\mathbb{P}\left\{\sigma_{\min}^2(\boldsymbol{\Phi}; T) \le \min_{J\in\mathscr{J}} \mathbb{E}\,\sigma_{\min}^2(\boldsymbol{\Phi}; T_J) - \sqrt{\mathrm{C}B^4 m\log(\#\mathscr{J})} - \zeta\right\}$$
$$\le (\#\mathscr{J})\,\mathrm{e}^{-(\sqrt{\mathrm{C}B^4 m\log(\#\mathscr{J})}+\zeta)^2/(\mathrm{C}B^4 m)} \le \mathrm{e}^{-\zeta^2/(\mathrm{C}B^4 m)}.$$

Define a random variable Y and a constant α:

$$Y := \sigma_{\min}^2(\boldsymbol{\Phi}; T) \quad \text{and} \quad \alpha := \min_{J\in\mathscr{J}} \mathbb{E}\,\sigma_{\min}^2(\boldsymbol{\Phi}; T_J) - \sqrt{\mathrm{C}B^4 m\log(\#\mathscr{J})}.$$

If $\alpha \le 0$, the stated result is trivial, so we may assume that $\alpha > 0$. Calculating as in Proposition 12.1,

$$\mathbb{E}\,Y \ge \int_0^\alpha \mathbb{P}\{Y > \zeta\}\,\mathrm{d}\zeta = \int_0^\alpha \mathbb{P}\{Y > \alpha - \zeta\}\,\mathrm{d}\zeta$$
$$= \alpha - \int_0^\alpha \mathbb{P}\{Y \le \alpha - \zeta\}\,\mathrm{d}\zeta \ge \alpha - \int_0^\alpha \mathrm{e}^{-\zeta^2/(\mathrm{C}B^4 m)}\,\mathrm{d}\zeta \ge \alpha - \mathrm{C}B^2\sqrt{m}.$$

Simplify this bound to complete the proof. □

13. Theorem 9.1: Replacing Most Entries of the Random Matrix

In this section, we show that it is possible to replace most of the entries of a random matrix $\mathbf{\Phi}$ with standard normal random variables without changing the restricted singular value $\sigma_{\min}(\mathbf{\Phi}; T_J)$ very much.

Proposition 13.1 (Theorem 9.1: Partial Replacement)**.** *Let $\mathbf{\Phi}$ be an $m \times n$ random matrix that satisfies Model 2.1 with magnitude bound B. Let J be a subset of $\{1, \dots, n\}$ with cardinality k, and let T_J be a closed subset of B^n for which*

$$\boldsymbol{t} \in T_J \quad \text{implies} \quad |t_j| \le k^{-1/2} \quad \text{for each index } j \in J^c. \tag{13.1}$$

Define an $m \times n$ random matrix $\mathbf{\Psi}$ where

$$\mathbf{\Psi}_J = \mathbf{\Phi}_J \quad \text{and} \quad \mathbf{\Psi}_{J^c} = \mathbf{\Gamma}_{J^c}. \tag{13.2}$$

Assuming that $k \ge m^{2/3}$, we have

$$\left| \mathbb{E} \min_{\boldsymbol{t} \in T_J} \|\mathbf{\Phi}\boldsymbol{t}\|^2 - \mathbb{E} \min_{\boldsymbol{t} \in T_J} \|\mathbf{\Psi}\boldsymbol{t}\|^2 \right| \quad \le \quad \frac{\mathrm{C} B^2 m^{1/3} n \log(mn)}{k^{1/2}}. \tag{13.3}$$

Equivalently,

$$\left| \mathbb{E}\, \sigma_{\min}^2(\mathbf{\Phi}; T_J) - \mathbb{E}\, \sigma_{\min}^2(\mathbf{\Psi}; T_J) \right| \quad \le \quad \frac{\mathrm{C} B^2 m^{1/3} n \log(mn)}{k^{1/2}}.$$

As usual, $\mathbf{\Gamma}$ is an $m \times n$ standard normal matrix.

The hypothesis (13.1) is an essential ingredient in the proof of Proposition 13.1. Indeed, we can only exchange the elements of the random matrix $\mathbf{\Phi}$ in the columns indexed by J^c because we depend on the uniform bound $k^{-1/2}$ to control the replacement errors.

13.1. **Proof of Proposition 13.1.** The proof of Proposition 13.1 involves several steps, so it is helpful to give an overview before we begin in earnest. Throughout this discussion, the index set J is fixed.

Let $\varepsilon \in (0, 1)$ be a discretization parameter. The first step in the argument is to replace the index set T_J by a finite subset T_J^ε with cardinality $\log(\# T_J^\varepsilon) \le n \log(3/\varepsilon)$. We obtain the bounds

$$\left| \mathbb{E} \min_{\boldsymbol{t} \in T_J} \|\mathbf{\Phi}\boldsymbol{t}\|^2 - \mathbb{E} \min_{\boldsymbol{t} \in T_J^\varepsilon} \|\mathbf{\Phi}\boldsymbol{t}\|^2 \right| \le 2mn\varepsilon \quad \text{and} \quad \left| \mathbb{E} \min_{\boldsymbol{t} \in T_J} \|\mathbf{\Psi}\boldsymbol{t}\|^2 - \mathbb{E} \min_{\boldsymbol{t} \in T_J^\varepsilon} \|\mathbf{\Psi}\boldsymbol{t}\|^2 \right| \le 2mn\varepsilon. \tag{13.4}$$

See Lemma 13.2 for details, which are quite standard.

Next, we introduce a smoothing parameter $\beta > 0$, and we define the soft-min function:

$$F : \mathbb{R}^{m \times n} \to \mathbb{R} \quad \text{where} \quad F(\boldsymbol{A}) := -\frac{1}{\beta} \log \sum\nolimits_{\boldsymbol{t} \in T_J^\varepsilon} \mathrm{e}^{-\beta \|\boldsymbol{A}\boldsymbol{t}\|^2}. \tag{13.5}$$

It is advantageous to work with the soft-min because it is differentiable. Lemma 13.3 demonstrates that we pay only a small cost for passing to the soft-min:

$$\left| \mathbb{E} \min_{\boldsymbol{t} \in T_J^\varepsilon} \|\mathbf{\Phi}\boldsymbol{t}\|^2 - \mathbb{E}\, F(\mathbf{\Phi}) \right| \le \beta^{-1} \log(\# T_J^\varepsilon) \quad \text{and} \quad \left| \mathbb{E} \min_{\boldsymbol{t} \in T_J^\varepsilon} \|\mathbf{\Psi}\boldsymbol{t}\|^2 - \mathbb{E}\, F(\mathbf{\Psi}) \right| \le \beta^{-1} \log(\# T_J^\varepsilon). \tag{13.6}$$

This argument is also standard.

Finally, we compare the expectation of the soft-min, evaluated at each of the random matrices:

$$|\mathbb{E}\, F(\mathbf{\Phi}) - \mathbb{E}\, F(\mathbf{\Psi})| \le \frac{\mathrm{C}(\beta B^4 + \beta^2 B^6) mn}{k^{3/2}}. \tag{13.7}$$

Lemma 13.4 encapsulates the this argument, which is based on the Lindeberg principle (Fact 13.6, below). The idea is to replace one entry of $\mathbf{\Phi}_{J^c}$ at a time with the corresponding entry of $\mathbf{\Psi}_{J^c}$, which is a Gaussian random variable. We lose very little with each exchange because the function F is smooth and the vectors in T_J^ε are bounded on the coordinates in J^c.

Combine the relations (13.4), (13.6), and (13.7) to obtain an estimate for the total error:

$$\left| \mathbb{E} \min_{\boldsymbol{t} \in T_J} \|\boldsymbol{\Phi t}\|^2 - \mathbb{E} \min_{\boldsymbol{t} \in T_J} \|\boldsymbol{\Psi t}\|^2 \right| \quad \leq \quad Cmn\varepsilon \; + \; C\beta^{-1} n \log(1/\varepsilon) \; + \; \frac{C(\beta B^4 + \beta^2 B^6) mn}{k^{3/2}}.$$

We used the fact that $\log \#(T_J^\varepsilon) \leq n \log(3/\varepsilon)$ in the smoothing term.

It remains to identify appropriate values for the parameters. Select $\varepsilon = (mn)^{-1}$ so the discretization error is negligible. We choose the smoothing parameter so that $\beta^3 = k^{3/2}/(B^6 m)$. In summary,

$$\left| \mathbb{E} \min_{\boldsymbol{t} \in T_J} \|\boldsymbol{\Phi t}\|^2 - \mathbb{E} \min_{\boldsymbol{t} \in T_J} \|\boldsymbol{\Psi t}\|^2 \right| \quad \leq \quad C \; + \; \frac{CB^2 m^{2/3} n}{k} \; + \; \frac{CB^2 m^{1/3} n \log(mn)}{k^{1/2}}.$$

Since $B \geq 1$ and $m^{2/3} \leq k \leq n$, the third term dominates. This is the bound stated in (13.3).

13.2. **Proposition 13.1: Discretizing the Index Set.** The first step in the proof of Proposition 13.1 is to replace the set T_J with a finite subset T_J^ε without changing the restricted singular values of $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ substantially.

Lemma 13.2 (Proposition 13.1: Discretization)**.** *Adopt the notation and hypotheses of Proposition 13.1. Fix a parameter $\varepsilon \in (0, 1]$. We can construct a subset T_J^ε of T_J with cardinality $\log(\#T_J^\varepsilon) \leq n \log(3/\varepsilon)$ that has the property*

$$\left| \mathbb{E} \min_{\boldsymbol{t} \in T_J} \|\boldsymbol{\Phi t}\|^2 - \mathbb{E} \min_{\boldsymbol{t} \in T_J^\varepsilon} \|\boldsymbol{\Phi t}\|^2 \right| \leq 2mn\varepsilon. \tag{13.8}$$

Furthermore, the bound (13.8) holds if we replace $\boldsymbol{\Phi}$ by $\boldsymbol{\Psi}$.

Proof. We choose the discretization T_J^ε to be an ε-covering of T_J with minimal cardinality. That is,

$$T_J^\varepsilon \subset T_J \quad \text{and} \quad \max_{\boldsymbol{t} \in T_J} \min_{\boldsymbol{t}_\varepsilon \in T_J^\varepsilon} \|\boldsymbol{t}_\varepsilon - \boldsymbol{t}\| \leq \varepsilon.$$

Since $\mathrm{Vol}(T_J) \leq \mathrm{Vol}(\mathsf{B}^n)$, we can use a classic volumetric argument to ensure that the covering satisfies $\#T_J^\varepsilon \leq (3/\varepsilon)^n$. For example, see [Ver12, Lem. 5.2].

When we replace the set T_J with the covering T_J^ε, we incur a discretization error. We establish the error bound for $\boldsymbol{\Phi}$; the argument for $\boldsymbol{\Psi}$ is identical. Since $T_J^\varepsilon \subset T_J$, it is immediate that

$$\mathbb{E} \min_{\boldsymbol{t} \in T_J} \|\boldsymbol{\Phi t}\|^2 \leq \mathbb{E} \min_{\boldsymbol{t} \in T_J^\varepsilon} \|\boldsymbol{\Phi t}\|^2.$$

We claim that

$$\mathbb{E} \min_{\boldsymbol{t} \in T_J^\varepsilon} \|\boldsymbol{\Phi t}\|^2 \leq \mathbb{E} \min_{\boldsymbol{t} \in T_J} \|\boldsymbol{\Phi t}\|^2 + 2\varepsilon \left(\mathbb{E} \|\boldsymbol{\Phi}\|^2 \right). \tag{13.9}$$

Since $\boldsymbol{\Phi}$ has standardized entries,

$$\mathbb{E} \|\boldsymbol{\Phi}\|^2 \leq \mathbb{E} \|\boldsymbol{\Phi}\|_{\mathrm{F}}^2 = mn,$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. Combine the last three displays to verify (13.8).

It is quite easy to establish the claim (13.9). For all $\boldsymbol{s}, \boldsymbol{t} \in T_J$, we have

$$\begin{aligned} \|\boldsymbol{\Phi s}\|^2 - \|\boldsymbol{\Phi t}\|^2 &= \big(\|\boldsymbol{\Phi s}\| + \|\boldsymbol{\Phi t}\| \big) \cdot \big(\|\boldsymbol{\Phi s}\| - \|\boldsymbol{\Phi t}\| \big) \\ &\leq \|\boldsymbol{\Phi}\| \big(\|\boldsymbol{s}\| + \|\boldsymbol{t}\| \big) \cdot \|\boldsymbol{\Phi}(\boldsymbol{s} - \boldsymbol{t})\| \\ &\leq 2 \|\boldsymbol{\Phi}\|^2 \|\boldsymbol{s} - \boldsymbol{t}\|. \end{aligned}$$

We have used standard norm inequalities and the fact that T_J is a subset of the unit ball. Now, let $\boldsymbol{t}_\star \in \arg\min_{\boldsymbol{t} \in T_J} \|\boldsymbol{\Phi t}\|^2$, and let $\boldsymbol{t}_\varepsilon$ be a point in T_J^ε for which $\|\boldsymbol{t}_\varepsilon - \boldsymbol{t}_\star\| \leq \varepsilon$. Then

$$\begin{aligned} \min_{\boldsymbol{t} \in T} \|\boldsymbol{\Phi t}\|^2 = \|\boldsymbol{\Phi t}_\star\|^2 &= \|\boldsymbol{\Phi t}_\varepsilon\|^2 - \big(\|\boldsymbol{\Phi t}_\varepsilon\|^2 - \|\boldsymbol{\Phi t}_\star\|^2 \big) \\ &\geq \min_{\boldsymbol{t} \in T_J^\varepsilon} \|\boldsymbol{\Phi t}\|^2 - 2 \|\boldsymbol{\Phi}\|^2 \|\boldsymbol{t}_\varepsilon - \boldsymbol{t}_\star\| \\ &\geq \min_{\boldsymbol{t} \in T_J^\varepsilon} \|\boldsymbol{\Phi t}\|^2 - 2\varepsilon \|\boldsymbol{\Phi}\|^2. \end{aligned}$$

Taking expectations, we arrive at (13.9). □

13.3. **Proposition 13.1: Smoothing the Minimum.** The second step in the proof of Proposition 13.1 is to pass from the restricted minimum singular value over the discrete set T_J^ε to a smooth function. We rely on an exponential smoothing technique that is common in the mathematical literature on statistical physics.

Lemma 13.3 (Proposition 13.1: Smoothing the Minimum)**.** *Adopt the notation and hypotheses of Proposition 13.1, and let T_J^ε be the set introduced in Lemma 13.2. Fix a parameter $\beta > 0$, and define the soft-min function F via* (13.5) *Then*

$$\left| \mathbb{E} \min_{\boldsymbol{t} \in T_J^\varepsilon} \|\boldsymbol{\Phi}\boldsymbol{t}\|^2 - \mathbb{E}\, F(\boldsymbol{\Phi}) \right| \le \frac{1}{\beta} \log(\# T_J^\varepsilon). \tag{13.10}$$

The bound (13.10) *also holds if we replace $\boldsymbol{\Phi}$ by $\boldsymbol{\Psi}$.*

Proof. This result follows from trivial bounds on the sum in the definition (13.5) of the soft-min function F. The summands are nonnegative, so the sum exceeds its maximum term. Thus,

$$\log \sum\nolimits_{\boldsymbol{t} \in T_J^\varepsilon} \mathrm{e}^{-\beta \|\boldsymbol{A}\boldsymbol{t}\|^2} \ge \log \max_{\boldsymbol{t} \in T_J^\varepsilon} \mathrm{e}^{-\beta \|\boldsymbol{A}\boldsymbol{t}\|^2} = -\beta \min_{\boldsymbol{t} \in T_J^\varepsilon} \|\boldsymbol{A}\boldsymbol{t}\|^2 .$$

Similarly, the sum does not exceed the number of summands times the maximum term, so

$$\log \sum\nolimits_{\boldsymbol{t} \in T_J^\varepsilon} \mathrm{e}^{-\beta \|\boldsymbol{A}\boldsymbol{t}\|^2} \le \log \Big((\# T_J^\varepsilon) \max_{\boldsymbol{t} \in T_J^\varepsilon} \mathrm{e}^{-\beta \|\boldsymbol{A}\boldsymbol{t}\|^2} \Big) = \log(\# T_J^\varepsilon) - \beta \min_{\boldsymbol{t} \in T_J^\varepsilon} \|\boldsymbol{A}\boldsymbol{t}\|^2 .$$

Combine the last two displays and multiply through by the negative number $-1/\beta$ to reach

$$\min_{\boldsymbol{t} \in T_J^\varepsilon} \|\boldsymbol{A}\boldsymbol{t}\|^2 - \frac{1}{\beta} \log(\# T_J^\varepsilon) \le -\frac{1}{\beta} \log \sum\nolimits_{\boldsymbol{t} \in T_J^\varepsilon} \mathrm{e}^{-\beta \|\boldsymbol{A}\boldsymbol{t}\|^2} \le \min_{\boldsymbol{t} \in T_J^\varepsilon} \|\boldsymbol{A}\boldsymbol{t}\|^2 .$$

Replace $\boldsymbol{A}$ by the random matrix $\boldsymbol{\Phi}$ and take the expectation to reach (13.10). Similarly, we can take $\boldsymbol{A} = \boldsymbol{\Psi}$ to obtain the result for $\boldsymbol{\Psi}$. □

13.4. **Proposition 13.1: Exchanging the Entries of the Random Matrix.** The last step in the proof of Proposition 13.1 is the hardest. We must demonstrate that the expectation $\mathbb{E}\, F(\boldsymbol{\Phi})$ of the soft-min function does not change very much when we replace the submatrix $\boldsymbol{\Phi}_{J^c}$ with the submatrix $\boldsymbol{\Psi}_{J^c}$.

Lemma 13.4 (Proposition 13.1: Exchanging Entries)**.** *Adopt the notation and hypotheses of Proposition 13.1; let T_J^ε be the set described in Lemma 13.2; and let F be the soft-min function* (13.5) *with parameter $\beta > 0$. Then*

$$|\mathbb{E}\, F(\boldsymbol{\Phi}) - \mathbb{E}\, F(\boldsymbol{\Psi})| \le \frac{\mathrm{C}(\beta B^4 + \beta^2 B^6) mn}{k^{3/2}}. \tag{13.11}$$

The random matrix $\boldsymbol{\Psi}$ is defined in (13.2).

Proof. We establish the lemma by replacing the rows of the random matrix $\boldsymbol{\Phi}$ by the rows of the random matrix $\boldsymbol{\Psi}$ one at a time. For each $i = 1, 2, \dots, m+1$, let $\boldsymbol{\Xi}(i)$ be the random matrix whose first $i-1$ rows are drawn from $\boldsymbol{\Psi}$ and whose remaining rows are drawn from $\boldsymbol{\Phi}$. By construction, $\boldsymbol{\Xi}(1) = \boldsymbol{\Phi}$ and $\boldsymbol{\Xi}(m+1) = \boldsymbol{\Psi}$. It follows that

$$|\mathbb{E}\, F(\boldsymbol{\Phi}) - \mathbb{E}\, F(\boldsymbol{\Psi})| \le \sum\nolimits_{i=1}^{m} \big| \mathbb{E}\, F\big(\boldsymbol{\Xi}(i)\big) - \mathbb{E}\, F\big(\boldsymbol{\Xi}(i+1)\big) \big| .$$

We will demonstrate that

$$\big| \mathbb{E}\, F\big(\boldsymbol{\Xi}(i)\big) - \mathbb{E}\, F\big(\boldsymbol{\Xi}(i+1)\big) \big| \le \frac{\mathrm{C}(\beta B^4 + \beta^2 B^6) n}{k^{3/2}} \quad \text{for } i = 1, \dots, m. \tag{13.12}$$

Combining the last two displays, we arrive at the bound (13.11).

Fix an index i. By construction, the random matrices $\boldsymbol{\Xi}(i)$ and $\boldsymbol{\Xi}(i+1)$ are identical, except in the ith row. To perform the estimate (13.12), it will be convenient to suppress the dependence of the function F on the remaining rows. To that end, define the functions

$$f_i : \mathbb{R}^n \to \mathbb{R} \quad \text{via} \quad f_i(\boldsymbol{a}) := -\frac{1}{\beta} \log \sum\nolimits_{\boldsymbol{t} \in T_J^\varepsilon} \mathrm{e}^{-\beta (\boldsymbol{a} \cdot \boldsymbol{t})^2 + q_i(\boldsymbol{t})}$$

where

$$q_i : \mathbb{R}^n \to \mathbb{R} \quad \text{via} \quad q_i(\boldsymbol{t}) := -\beta \left[\sum\nolimits_{j<i} (\boldsymbol{\psi}^j \cdot \boldsymbol{t})^2 + \sum\nolimits_{j>i} (\boldsymbol{\varphi}^j \cdot \boldsymbol{t})^2 \right].$$

We have written $\boldsymbol{\varphi}^j$ and $\boldsymbol{\psi}^j$ for the jth rows of $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$. With these definitions,

$$F\big(\Xi(i)\big) = f_i\big(\boldsymbol{\varphi}^i\big) \quad \text{and} \quad F\big(\Xi(i+1)\big) = f_i\big(\boldsymbol{\psi}^i\big).$$

Therefore, the inequality (13.12) is equivalent with

$$\left| \mathbb{E} f_i\big(\boldsymbol{\varphi}^i\big) - \mathbb{E} f_i\big(\boldsymbol{\psi}^i\big) \right| \le \frac{\mathrm{C}(\beta B^4 + \beta^2 B^6) n}{k^{3/2}} \quad \text{for } i = 1, \dots, m. \tag{13.13}$$

Since the matrix $\boldsymbol{\Phi}$ is drawn from Model 2.1, the random vector $\boldsymbol{\varphi}^i$ contains independent, standardized entries whose magnitudes are bounded by B. Meanwhile, the form (13.2) of the matrix $\boldsymbol{\Psi}$ shows that the random vector $\boldsymbol{\psi}^i$ coincides with $\boldsymbol{\varphi}^i$ on the components indexed by J, while the entries of $\boldsymbol{\psi}^i$ indexed by J^c are independent standard normal variables. Sublemma 13.5, below, contains the proof of the inequality (13.13). □

13.4.1. *Lemma 13.4: Comparison Principle for One Row.* To complete the argument in Lemma 13.4, we need to control how much a certain function changes when we replace some of the entries of its argument with standard normal variables. The following result contains the required calculation.

Sublemma 13.5 (Lemma 13.4: Comparison for One Row)**.** *Adopt the notation and hypotheses of Proposition 13.1, and let T_J^ϵ be the set defined in Lemma 13.3. Introduce the function*

$$f : \mathbb{R}^n \to \mathbb{R} \quad \textit{given by} \quad f(\boldsymbol{a}) := -\frac{1}{\beta} \log \sum\nolimits_{\boldsymbol{t} \in T_J^\epsilon} \mathrm{e}^{-\beta(\boldsymbol{a}\cdot\boldsymbol{t})^2 + q(\boldsymbol{t})},$$

where $q : T_J^\epsilon \to \mathbb{R}$ is an arbitrary function. Suppose that $\boldsymbol{\varphi} \in \mathbb{R}^n$ is a random vector with independent, standardized entries that are bounded in magnitude by B. Suppose that $\boldsymbol{\psi} \in \mathbb{R}^n$ is a random vector with

$$\boldsymbol{\psi}_J = \boldsymbol{\varphi}_J \quad \textit{and} \quad \boldsymbol{\psi}_{J^c} = \boldsymbol{\gamma}_{J^c},$$

where $\boldsymbol{\gamma} \in \mathbb{R}^n$ is a standard normal vector. Then

$$\left| \mathbb{E} f(\boldsymbol{\varphi}) - \mathbb{E} f(\boldsymbol{\psi}) \right| \le \frac{\mathrm{C}(\beta B^4 + \beta^2 B^6) n}{k^{3/2}}. \tag{13.14}$$

The proof of Sublemma 13.5 is based on a modern interpretation [MOO10, Cha06, KM11] of the Lindeberg exchange principle [Lin22, Tro59, Rot73]. It is similar in spirit with examples from the paper [KM11]. We apply the following version of the Lindeberg principle, which is adapted from these works.

Fact 13.6 (Lindeberg Exchange Principle)**.** *Let $r : \mathbb{R} \to \mathbb{R}$ be a function with three continuous derivatives. Let φ and ψ be standardized random variables, not necessarily independent, with three finite moments. Then*

$$\left| \mathbb{E} r(\varphi) - \mathbb{E} r(\psi) \right| \le \frac{1}{6} \mathbb{E} \left[|\varphi|^3 \max_{|\alpha| \le |\varphi|} \left| r'''(\alpha) \right| + |\psi|^3 \max_{|\alpha| \le |\psi|} \left| r'''(\alpha) \right| \right].$$

The proof of Fact 13.6 involves nothing more than a third-order Taylor expansion of r around the origin. The first- and second-order terms cancel because the random variables φ and ψ have mean zero and variance one. With this inequality at hand, we may proceed with the proof of the sublemma.

Proof of Sublemma 13.5. Without loss of generality, assume that the index set $J = \{1, \dots, k\}$. Indeed, the entries of the random matrix $\boldsymbol{\Phi}$ are independent, standardized, and uniformly bounded so it does not matter which set J of columns we distinguish. This choice allows more intuitive indexing.

For each fixed index $j \in \{k+1, \dots, n\}$, define the interpolating vector

$$\boldsymbol{\xi}_j : \mathbb{R} \to \mathbb{R}^n \quad \text{where} \quad \boldsymbol{\xi}_j(\alpha) := (\varphi_1, \dots, \varphi_{j-1}, \alpha, \psi_{j+1}, \dots, \psi_n).$$

Introduce the function

$$r_j : \mathbb{R} \to \mathbb{R} \quad \text{where} \quad r_j(\alpha) := f\big(\boldsymbol{\xi}_j(\alpha)\big).$$

Observe that

$$\left|\mathbb{E} f(\boldsymbol{\varphi}) - \mathbb{E} f(\boldsymbol{\psi})\right| = \left|\sum\nolimits_{j=k+1}^{n} \mathbb{E} f\big(\boldsymbol{\xi}_j(\varphi_j)\big) - \mathbb{E} f\big(\boldsymbol{\xi}_j(\psi_j)\big)\right| \le \sum\nolimits_{j=k+1}^{n} \left|\mathbb{E}\, r_j(\varphi_j) - \mathbb{E}\, r_j(\psi_j)\right|.$$

The first identity holds because the sum telescopes. Fact 13.6 shows that

$$\left|\mathbb{E}\, r_j(\varphi_j) - \mathbb{E}\, r_j(\psi_j)\right| \le \frac{1}{6}\mathbb{E}\left[|\varphi_j|^3 \max_{|\alpha|\le|\varphi_j|}\left|r_j'''(\alpha)\right| + |\psi_j|^3 \max_{|\alpha|\le|\psi_j|}\left|r_j'''(\alpha)\right|\right].$$

We claim that, for each index $j \in \{k+1, \dots, n\}$,

$$\begin{aligned} \mathbb{E}\left[|\varphi_j|^3 \max_{|\alpha|\le|\varphi_j|}\left|r_j'''(\alpha)\right|\right] &\le \mathrm{C}(\beta B^4 + \beta^2 B^6) k^{-3/2}, \quad \text{and} \\ \mathbb{E}\left[|\psi_j|^3 \max_{|\alpha|\le|\psi_j|}\left|r_j'''(\alpha)\right|\right] &\le \mathrm{C}(\beta B^4 + \beta^2 B^6) k^{-3/2}. \end{aligned} \tag{13.15}$$

Once we establish the bound (13.15), we can combine the last three displays to reach

$$\left|\mathbb{E} f(\boldsymbol{\varphi}) - \mathbb{E} f(\boldsymbol{\psi})\right| \le \mathrm{C}(\beta B^4 + \beta^2 B^6) k^{-3/2}(n-k).$$

The main result (13.14) follows.

To establish (13.15), fix an index $j \in \{k+1, \dots, n\}$. The forthcoming Sublemma 13.7 will demonstrate that

$$\left|r_j'''(\alpha)\right| \le \mathrm{C}\left(\max_{\boldsymbol{t}\in T_J^\varepsilon}\left|\boldsymbol{\xi}'(\alpha)\cdot\boldsymbol{t}\right|^3\right)\left(\beta\,\mathbb{E}_{\boldsymbol{v}}\left|\boldsymbol{\xi}_j(\alpha)\cdot\boldsymbol{v}\right| + \beta^2\,\mathbb{E}_{\boldsymbol{v}}\left|\boldsymbol{\xi}_j(\alpha)\cdot\boldsymbol{v}\right|^3\right) \tag{13.16}$$

In this expression, $\boldsymbol{v} \in T_J^\varepsilon$ is a random vector that is independent from $\boldsymbol{\xi}_j$; the precise distribution of $\boldsymbol{v}$ is immaterial. Note that

$$\max_{\boldsymbol{t}\in T_J^\varepsilon}\left|\boldsymbol{\xi}'(\alpha)\cdot\boldsymbol{t}\right|^3 = \max_{\boldsymbol{t}\in T_J^\varepsilon}|t_j|^3 \le k^{-3/2}. \tag{13.17}$$

Indeed, the derivative $\boldsymbol{\xi}_j'(\alpha) = \mathbf{e}_j$, where $\mathbf{e}_j$ is the jth standard basis vector. The last inequality holds because the conditions $T_J^\varepsilon \subset T_J$ and $j \in J^c$ allow us to invoke the assumption (13.1).

We arrive at the following bound on the quantity of interest from (13.15):

$$\mathbb{E}\left[|\varphi_j|^3 \max_{|\alpha|\le|\varphi_j|}\left|r_j'''(\alpha)\right|\right] \le \mathrm{C}k^{-3/2}\left(\beta\mathbb{E}\left[|\varphi_j|^3 \max_{|\alpha|\le|\varphi_j|}\left|\boldsymbol{\xi}_j(\alpha)\cdot\boldsymbol{v}\right|\right] + \beta^2\,\mathbb{E}\left[|\varphi_j|^3 \max_{|\alpha|\le|\varphi_j|}\left|\boldsymbol{\xi}_j(\alpha)\cdot\boldsymbol{v}\right|^3\right]\right). \tag{13.18}$$

We have merged the bounds (13.16) and (13.17) to control $\left|r_j(\alpha)\right|$. Next, we invoked Jensen's inequality to draw the expectation over $\boldsymbol{v}$ out of the maximum over α. Last, we combined the expectations to reach (13.18). It remains to bound the expectations on the right-hand side.

Let us begin with the term in (13.18) that is linear in β. In view of the identity $\boldsymbol{\xi}_j(\alpha) = \alpha\mathbf{e}_j + \boldsymbol{\xi}_j(0)$,

$$\begin{aligned} \mathbb{E}\left[|\varphi_j|^3 \max_{|\alpha|\le|\varphi_j|}\left|\boldsymbol{\xi}_j(\alpha)\cdot\boldsymbol{v}\right|\right] &\le \mathbb{E}\left[|\varphi_j|^3 \max_{|\alpha|\le|\varphi_j|}\left(|\alpha v_j| + \left|\boldsymbol{\xi}_j(0)\cdot\boldsymbol{v}\right|\right)\right] \\ &= \mathbb{E}\left[|\varphi_j|^4 |v_j|\right] + \left(\mathbb{E}|\varphi_j|^3\right)\left(\mathbb{E}\left|\boldsymbol{\xi}_j(0)\cdot\boldsymbol{v}\right|\right) \\ &\le B^4 k^{-1/2} + B^3. \end{aligned} \tag{13.19}$$

We used the assumption that φ_j is independent from $\boldsymbol{\xi}_j(0)$ to factor the expectation in the second term in the second line. The bounds in the third line exploit the fact that $|\varphi_j| \le B$ and, via the hypothesis (13.1), the fact that $\boldsymbol{v} \in T_J$. We also relied on the estimate

$$\mathbb{E}\left|\boldsymbol{\xi}_j(0)\cdot\boldsymbol{v}\right| = \mathbb{E}\left|\sum\nolimits_{i<j}\varphi_i v_i + \sum\nolimits_{i>j}\psi_i v_i\right| \le \|\boldsymbol{v}\| \le 1$$

Indeed, Jensen's inequality allows us to replace the expectation with the second moment, which simplifies to $\|\boldsymbol{v}\|$ because $\{\varphi_1, \dots, \varphi_{j-1}, \psi_{j+1}, \dots, \psi_n\}$ is an independent family of standardized random variables. Since $\boldsymbol{v} \in T_J$, the norm $\|\boldsymbol{v}\|$ does not exceed one.

Continuing to the term in (13.18) that is quadratic in β^2, we pursue the same approach to see that

$$\begin{aligned}\mathbb{E}\left[|\varphi_j|^3 \max_{|\alpha|\le|\varphi_j|} \left|\boldsymbol{\xi}_j(\alpha)\cdot\boldsymbol{v}\right|^3\right] &\le \mathrm{C}\mathbb{E}\left[|\varphi_j|^3 \max_{|\alpha|\le|\varphi_j|} \left(|\alpha v_j|^3 + \left|\boldsymbol{\xi}_j(0)\cdot\boldsymbol{v}\right|^3\right)\right] \\ &\le \mathrm{C}\left(B^6 k^{-3/2} + B^6\right).\end{aligned} \tag{13.20}$$

This bound relies on the Khintchine-type inequality

$$\left(\mathbb{E}\left|\boldsymbol{\xi}_j(0)\cdot\boldsymbol{v}\right|^3\right)^{1/3} \le \mathrm{C}B\,\|\boldsymbol{v}\| \le \mathrm{C}B.$$

For example, see [Ver12, Cor. 5.12]. We remark that this estimate could be improved if we had additional information about the distribution of the entries of $\boldsymbol{\Phi}$.

Substituting the bounds (13.19) and (13.20) into (13.18), we obtain

$$\begin{aligned}\mathbb{E}\left[|\varphi_j|^3 \max_{|\alpha|\le|\varphi_j|} \left|r_j'''(\alpha)\right|\right] &\le \mathrm{C}k^{-3/2}\left[\beta\left(B^4k^{-1/2} + B^3\right) + \beta^2\left(B^6k^{-3/2} + B^6\right)\right] \\ &\le \mathrm{C}(\beta B^4 + \beta^2 B^6)k^{-3/2}.\end{aligned}$$

This establishes the first branch of the claim (13.15). The second branch follows from a similar argument, where we use explicit values for the moments of a standard normal variable instead of the uniform upper bound B. □

13.4.2. *Sublemma 13.5: Derivative Calculations.* The final obstacle in the proof of Proposition 13.1 is to bound the derivatives that are required in Sublemma 13.5. This argument uses some standard methods from statistical physics, and it is similar with the approach in the paper [KM11].

Sublemma 13.7 (Sublemma 13.5: Derivatives)**.** *Adopt the notation and hypotheses of Proposition 13.1 and Sublemma 13.5. Let $\boldsymbol{\xi}:\mathbb{R}\to\mathbb{R}^n$ be a linear function, so its derivative $\boldsymbol{\xi}'\in\mathbb{R}^n$ is a constant vector. Define the function*

$$r(\alpha) := f\left(\boldsymbol{\xi}(\alpha)\right) = -\frac{1}{\beta}\log\textstyle\sum_{\boldsymbol{t}\in T_J^\epsilon} \mathrm{e}^{-\beta(\boldsymbol{\xi}(\alpha)\cdot\boldsymbol{t})^2 + q(\boldsymbol{t})}$$

where $q: T_J^\epsilon \to \mathbb{R}$ is arbitrary. The third derivative of this function satisfies

$$\left|r'''(\alpha)\right| \le 48\left(\max_{\boldsymbol{t}\in T_J^\epsilon}\left|\boldsymbol{\xi}'\cdot\boldsymbol{t}\right|^3\right)\left(\beta\,\mathbb{E}_{\boldsymbol{v}}\left|\boldsymbol{\xi}(\alpha)\cdot\boldsymbol{v}\right| + \beta^2\,\mathbb{E}_{\boldsymbol{v}}\left|\boldsymbol{\xi}(\alpha)\cdot\boldsymbol{v}\right|^3\right).$$

In this expression, $\boldsymbol{v}\in T_J^\epsilon$ is a random vector that is independent from $\boldsymbol{\xi}(\alpha)$.

Proof. Introduce a (parameterized) probability measure μ_α on the set T_J^ϵ:

$$\mu_\alpha(\boldsymbol{t}) := \frac{1}{Z_\alpha}\mathrm{e}^{-\beta(\boldsymbol{\xi}(\alpha)\cdot\boldsymbol{t})^2 + q(\boldsymbol{t})} \quad \text{where} \quad Z_\alpha := \textstyle\sum_{\boldsymbol{t}\in T_J^\epsilon} \mathrm{e}^{-\beta(\boldsymbol{\xi}(\alpha)\cdot\boldsymbol{t})^2 + q(\boldsymbol{t})}.$$

We treat the normalizing factor Z_α as a function of α, and we write its derivatives as $Z_\alpha', Z_\alpha'', Z_\alpha'''$. It is convenient to use the statistical mechanics notation for expectation with respect to this measure. For any function $h: T_J^\epsilon \to \mathbb{R}$,

$$\langle h(\boldsymbol{t})\rangle := \textstyle\sum_{\boldsymbol{t}\in T_J^\epsilon} h(\boldsymbol{t})\mu_\alpha(\boldsymbol{t}).$$

For brevity, we always suppress the dependence of $\langle\cdot\rangle$ on the parameter α. We often suppress the dependence of $\boldsymbol{\xi}$ on α as well.

The function r is proportional to the logarithm of the normalizing factor Z_α:

$$r(\alpha) = -\frac{1}{\beta}\log Z_\alpha.$$

Thus, it is straightforward to express the third derivative of r in terms of the derivatives of Z_α:

$$r'''(\alpha) = -\frac{1}{\beta}\left(\frac{Z_\alpha'}{Z_\alpha}\right)'' = -\frac{1}{\beta}\left(\frac{Z_\alpha''}{Z_\alpha} - \frac{(Z_\alpha')^2}{Z_\alpha^2}\right)' = -\frac{1}{\beta}\left(\frac{Z_\alpha'''}{Z_\alpha} - 3\frac{Z_\alpha'}{Z_\alpha}\frac{Z_\alpha''}{Z_\alpha} + 2\left(\frac{Z_\alpha'}{Z_\alpha}\right)^3\right).$$

By direct calculation, the derivative Z'_α satisfies

$$\begin{aligned}\frac{Z'_\alpha}{Z_\alpha} &= \frac{1}{Z_\alpha}\sum\nolimits_{\boldsymbol{t}\in T_J^\varepsilon}(-2\beta)(\boldsymbol{\xi}'\cdot\boldsymbol{t})(\boldsymbol{\xi}(\alpha)\cdot\boldsymbol{t})\mathrm{e}^{-\beta(\boldsymbol{\xi}(\alpha)\cdot\boldsymbol{t})^2+q(\boldsymbol{t})}\\ &= -2\beta\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})(\boldsymbol{\xi}\cdot\boldsymbol{t})\right\rangle.\end{aligned}$$

The second derivative is

$$\begin{aligned}\frac{Z''_\alpha}{Z_\alpha} &= \frac{1}{Z_\alpha}\sum\nolimits_{\boldsymbol{t}\in T_J^\varepsilon}\left[(-2\beta)(\boldsymbol{\xi}'\cdot\boldsymbol{t})^2+(2\beta)^2(\boldsymbol{\xi}'\cdot\boldsymbol{t})^2(\boldsymbol{\xi}(\alpha)\cdot\boldsymbol{t})^2\right]\mathrm{e}^{-\beta(\boldsymbol{\xi}(\alpha)\cdot\boldsymbol{t})^2+q(\boldsymbol{t})}\\ &= -2\beta\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})^2\right\rangle+4\beta^2\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})^2(\boldsymbol{\xi}\cdot\boldsymbol{t})^2\right\rangle.\end{aligned}$$

The third derivative is

$$\begin{aligned}\frac{Z'''_\alpha}{Z_\alpha} &= \frac{1}{Z_\alpha}\sum\nolimits_{\boldsymbol{t}\in T_J^\varepsilon}\left[12\beta^2(\boldsymbol{\xi}'\cdot\boldsymbol{t})^3(\boldsymbol{\xi}(\alpha)\cdot\boldsymbol{t})-8\beta^3(\boldsymbol{\xi}'\cdot\boldsymbol{t})^3(\boldsymbol{\xi}(\alpha)\cdot\boldsymbol{t})^3\right]\mathrm{e}^{-\beta(\boldsymbol{\xi}(\alpha)\cdot\boldsymbol{t})^2+q(\boldsymbol{t})}\\ &= 12\beta^2\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})^3(\boldsymbol{\xi}\cdot\boldsymbol{t})\right\rangle-8\beta^3\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})^3(\boldsymbol{\xi}\cdot\boldsymbol{t})^3\right\rangle\end{aligned}$$

We ascertain that

$$\begin{aligned}r'''(\alpha) = \beta\big[&-12\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})^3(\boldsymbol{\xi}\cdot\boldsymbol{t})\right\rangle+12\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})^2\right\rangle\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})(\boldsymbol{\xi}\cdot\boldsymbol{t})\right\rangle\big]\\ &+\beta^2\big[8\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})^3(\boldsymbol{\xi}\cdot\boldsymbol{t})^3\right\rangle-24\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})(\boldsymbol{\xi}\cdot\boldsymbol{t})\right\rangle\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})^2(\boldsymbol{\xi}\cdot\boldsymbol{t})^2\right\rangle+16\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})(\boldsymbol{\xi}\cdot\boldsymbol{t})\right\rangle^3\big].\end{aligned}\tag{13.21}$$

This completes the calculation of the exact form of the third derivative of r.

Next, we simplify the formula (13.21) using basic probability inequalities for the expectation $\langle\cdot\rangle$. First, consider the terms that are linear in β. Observe that

$$\left|\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})^3(\boldsymbol{\xi}\cdot\boldsymbol{t})\right\rangle\right| \le \left\langle\left|\boldsymbol{\xi}'\cdot\boldsymbol{t}\right|^3|\boldsymbol{\xi}\cdot\boldsymbol{t}|\right\rangle \le \left(\max_{\boldsymbol{t}\in T_J^\varepsilon}\left|\boldsymbol{\xi}'\cdot\boldsymbol{t}\right|^3\right)\left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|\right\rangle.$$

Indeed, Jensen's inequality allows us to draw the absolute value inside the average, and we can invoke Hölder's inequality to pull out the maximum of the first term. Similarly, since $\langle 1\rangle = 1$,

$$\left|\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})^2\right\rangle\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})(\boldsymbol{\xi}\cdot\boldsymbol{t})\right\rangle\right| \le \left(\max_{\boldsymbol{t}\in T_J^\varepsilon}\left|\boldsymbol{\xi}'\cdot\boldsymbol{t}\right|^3\right)\left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|\right\rangle$$

Next, consider the terms that are quadratic in β. The simplest is

$$\left|\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})^3(\boldsymbol{\xi}\cdot\boldsymbol{t})^3\right\rangle\right| \le \left(\max_{\boldsymbol{t}\in T_J^\varepsilon}\left|\boldsymbol{\xi}'\cdot\boldsymbol{t}\right|^3\right)\left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|^3\right\rangle.$$

Using Jensen's inequality, we find that

$$\left|\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})(\boldsymbol{\xi}\cdot\boldsymbol{t})\right\rangle^3\right| \le \left(\max_{\boldsymbol{t}\in T_J^\varepsilon}\left|\boldsymbol{\xi}'\cdot\boldsymbol{t}\right|^3\right)\left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|\right\rangle^3 \le \left(\max_{\boldsymbol{t}\in T_J^\varepsilon}\left|\boldsymbol{\xi}'\cdot\boldsymbol{t}\right|^3\right)\left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|^3\right\rangle.$$

Last, using Lyapunov's inequality twice,

$$\left|\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})(\boldsymbol{\xi}\cdot\boldsymbol{t})\right\rangle\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{t})^2(\boldsymbol{\xi}\cdot\boldsymbol{t})^2\right\rangle\right| \le \left(\max_{\boldsymbol{t}\in T_J^\varepsilon}\left|\boldsymbol{\xi}'\cdot\boldsymbol{t}\right|^3\right)\left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|\right\rangle\left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|^2\right\rangle \le \left(\max_{\boldsymbol{t}\in T_J^\varepsilon}\left|\boldsymbol{\xi}'\cdot\boldsymbol{t}\right|^3\right)\left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|^3\right\rangle.$$

Introduce the last five displays into (13.21) to arrive at the bound

$$\left|r'''(\alpha)\right| \le \left(\max_{\boldsymbol{t}\in T_J^\varepsilon}\left|\boldsymbol{\xi}'\cdot\boldsymbol{t}\right|^3\right)\left(24\beta\left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|\right\rangle+48\beta^2\left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|^3\right\rangle\right).\tag{13.22}$$

Last, we want to replace the averages with respect to μ_α by averages with respect to a simpler probability measure that does not depend on $\boldsymbol{\xi}(\alpha)$. This argument relies on a correlation inequality. Define another probability measure μ_* on the set T_J^ε:

$$\mu_*(\boldsymbol{t}) := \frac{1}{Z_*}\mathrm{e}^{q(\boldsymbol{t})} \quad\text{where}\quad Z_* := \sum\nolimits_{\boldsymbol{t}\in T_J^\varepsilon}\mathrm{e}^{q(\boldsymbol{t})}.$$

Write $\langle\cdot\rangle_*$ for averages with respect to this new measure. First, consider the average

$$\begin{aligned}\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|\rangle &= \frac{1}{Z_\alpha}\sum\nolimits_{\boldsymbol{t}\in T_J^\varepsilon}|\boldsymbol{\xi}\cdot\boldsymbol{t}|\,\mathrm{e}^{-\beta(\boldsymbol{\xi}\cdot\boldsymbol{t})^2+q(\boldsymbol{t})}\\ &= \frac{Z_*}{Z_\alpha}\frac{1}{Z_*}\sum\nolimits_{\boldsymbol{t}\in T_J^\varepsilon}|\boldsymbol{\xi}\cdot\boldsymbol{t}|\,\mathrm{e}^{-\beta(\boldsymbol{\xi}\cdot\boldsymbol{t})^2}\mathrm{e}^{q(\boldsymbol{t})}\\ &= \frac{Z_*}{Z_\alpha}\left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|\,\mathrm{e}^{-\beta(\boldsymbol{\xi}\cdot\boldsymbol{t})^2}\right\rangle_*.\end{aligned}$$

Let $X := |\boldsymbol{\xi}\cdot\boldsymbol{t}|$ be the random variable obtained by pushing forward the measure μ_* from T_J^ε to the nonnegative real line. Since $x\mapsto x$ is increasing and $x\mapsto\exp(-\beta x^2)$ is decreasing, Chebyshev's association inequality [BLM13, Thm. 2.14] provides that

$$\left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|\,\mathrm{e}^{-\beta(\boldsymbol{\xi}\cdot\boldsymbol{t})^2}\right\rangle_* \le \langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|\rangle_*\left\langle\mathrm{e}^{-\beta(\boldsymbol{\xi}\cdot\boldsymbol{t})^2}\right\rangle_* = \langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|\rangle_*\frac{1}{Z_*}\sum\nolimits_{\boldsymbol{t}\in T_J^\varepsilon}\mathrm{e}^{-\beta(\boldsymbol{\xi}\cdot\boldsymbol{t})^2+q(\boldsymbol{t})} = \langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|\rangle_*\frac{Z_\alpha}{Z_*}.$$

In summary,

$$\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|\rangle \le \langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|\rangle_*. \tag{13.23}$$

The same argument shows that

$$\left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|^3\right\rangle \le \left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|^3\right\rangle_*. \tag{13.24}$$

Introduce (13.23) and (13.24) into the bound (13.22) to reach the inequality

$$r'''(\alpha) \le 48\Big(\max_{\boldsymbol{t}\in T_J^\varepsilon}\left|\boldsymbol{\xi}'\cdot\boldsymbol{t}\right|^3\Big)\left(\beta\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|\rangle_* + \beta^2\left\langle|\boldsymbol{\xi}\cdot\boldsymbol{t}|^3\right\rangle_*\right).$$

The statement of the result follows when we reinterpret the averages with respect to μ_* as expectations with respect to a random vector $\boldsymbol{\nu}\in T_J^\varepsilon$ with distribution $Z_*^{-1}\mathrm{e}^{-q(\boldsymbol{t})}$. □

14. Theorem 9.1: Bounding the Restricted Singular Value by the Excess Width

In Section 13, we showed that the restricted singular value $\sigma_{\min}(\boldsymbol{\Phi};T_J)$ of the random matrix $\boldsymbol{\Phi}$ does not change very much if we replace $\boldsymbol{\Phi}$ with a hybrid matrix $\boldsymbol{\Psi}$, defined in (13.2), that contains many standard normal random variables. Our next goal is to relate the restricted singular value $\sigma_{\min}(\boldsymbol{\Psi};T_J)$ of the hybrid matrix to the excess width of the set T_J.

Proposition 14.1 (Theorem 9.1: Excess Width Bound)**.** *Let $\boldsymbol{\Phi}$ be an $m\times n$ random matrix from Model 2.1 with magnitude bound B, and let $\boldsymbol{\Gamma}$ be an $m\times n$ random matrix with independent, standard normal entries. Let J be a subset of $\{1,\dots,n\}$ with cardinality k, and let T_J be a closed subset of B^n. Introduce the $m\times n$ random matrix $\boldsymbol{\Psi} := \boldsymbol{\Psi}(J)$ from (13.2). Then*

$$\mathbb{E}\,\sigma_{\min}^2(\boldsymbol{\Psi};T_J) = \mathbb{E}\min_{\boldsymbol{t}\in T_J}\left\|\boldsymbol{\Phi}_J\boldsymbol{t}_J+\boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\|^2 \ge \left(\mathscr{E}_m(T_J)-CB^2\sqrt{k}\right)_+^2. \tag{14.1}$$

Furthermore, if T_J is convex and $k\ge m^{1/2}$,

$$\mathbb{E}\,\sigma_{\min}^2(\boldsymbol{\Psi};T_J) = \mathbb{E}\min_{\boldsymbol{t}\in T_J}\left\|\boldsymbol{\Phi}_J\boldsymbol{t}_J+\boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\|^2 \le \left(\mathscr{E}_m(T_J)+CB^2\sqrt{k}\right)_+^2. \tag{14.2}$$

To obtain this result, we will invoke the Gaussian Minimax Theorem [Gor85, Gor88] to reduce the random matrix bounds to simpler bounds involving random vectors. This approach has been used to study the ordinary singular values of a Gaussian matrix [DS01, Sec. 2.3]. It also plays a role in the analysis of restricted singular values of Gaussian matrices [Sto13, OTH13, TOH15]. The application here is complicated significantly by the presence of the non-Gaussian matrix $\boldsymbol{\Phi}_J$.

14.1. **Proof of Proposition 14.1.** The proof of Proposition 14.1 involves several steps, so it is helpful once again to summarize the calculations that are required.

First, let us explain the process by which we obtain the lower bound for a general closed subset T_J of the Euclidean unit ball B^m. We will argue that

$$\begin{aligned}
\mathbb{E}\min_{\boldsymbol{t}\in T_J}\left\|\boldsymbol{\Phi}_J\boldsymbol{t}_J+\boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\|^2 &\geq \left(\mathbb{E}\min_{\boldsymbol{t}\in T_J}\left\|\boldsymbol{\Phi}_J\boldsymbol{t}_J+\boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\|\right)_+^2 && \text{(Jensen)}\\
&\geq \left(\mathbb{E}\min_{\boldsymbol{t}\in T_J}\left(\left\|\boldsymbol{\Phi}_J\boldsymbol{t}_J+\boldsymbol{h}\,\|\boldsymbol{t}_{J^c}\|\right\|+\boldsymbol{g}_{J^c}\cdot\boldsymbol{t}_{J^c}\right)-2\right)_+^2 && \text{(Lemma 14.5)}\\
&\geq \left(\mathbb{E}\min_{\boldsymbol{t}\in T_J}\left(\sqrt{m}\,\|\boldsymbol{t}\|+\boldsymbol{g}_{J^c}\cdot\boldsymbol{t}_{J^c}\right)-\mathrm{C}B^2\sqrt{k}\right)_+^2 && \text{(Lemma 14.7)}\\
&\geq \left(\mathbb{E}\min_{\boldsymbol{t}\in T_J}\left(\sqrt{m}\,\|\boldsymbol{t}\|+\boldsymbol{g}\cdot\boldsymbol{t}\right)-\mathrm{C}B^2\sqrt{k}\right)_+^2 && \text{(Lemma 14.8)}\\
&= \left(\mathscr{E}_m(T_J)-\mathrm{C}B^2\sqrt{k}\right)_+^2. && \text{(Definition 4.2)}
\end{aligned}$$

In the first line, we pass from the expected square to the square of the expectation, which reduces the technical complexity of the rest of the argument. Next, we replace the $m\times n$ Gaussian matrix $\boldsymbol{\Gamma}$ with an expectation over two independent standard normal vectors $\boldsymbol{h}\in\mathbb{R}^m$ and $\boldsymbol{g}\in\mathbb{R}^n$. Third, we remove the remaining piece $\boldsymbol{\Phi}_J$ of the original random matrix to arrive at a formula involving only the random vector $\boldsymbol{g}$. Finally, we replace the missing coordinates in the random vector $\boldsymbol{g}$ to obtain a bound in terms of the excess width $\mathscr{E}_m(T_J)$. We arrive at the inequality (14.1).

The upper bound follows from a similar calculation. Assuming that T_J is convex, we may calculate that

$$\begin{aligned}
\mathbb{E}\min_{\boldsymbol{t}\in T_J}\left\|\boldsymbol{\Phi}_J\boldsymbol{t}_J+\boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\|^2 &\leq \left(\mathbb{E}\min_{\boldsymbol{t}\in T_J}\left\|\boldsymbol{\Phi}_J\boldsymbol{t}_J+\boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\|+\mathrm{C}Bm^{1/4}\right)_+^2 && \text{(Lemma 14.2)}\\
&\leq \left(\mathbb{E}\min_{\boldsymbol{t}\in T_J}\left(\left\|\boldsymbol{\Phi}_J\boldsymbol{t}_J+\boldsymbol{h}\,\|\boldsymbol{t}_{J^c}\|\right\|+\boldsymbol{g}_{J^c}\cdot\boldsymbol{t}_{J^c}\right)+\mathrm{C}Bm^{1/4}\right)_+^2 && \text{(Lemma 14.6)}\\
&\leq \left(\mathbb{E}\min_{\boldsymbol{t}\in T_J}\left(\sqrt{m}\,\|\boldsymbol{t}\|+\boldsymbol{g}_{J^c}\cdot\boldsymbol{t}_{J^c}\right)+\mathrm{C}B^2\sqrt{k}\right)_+^2 && \text{(Lemma 14.7)}\\
&\leq \left(\mathbb{E}\min_{\boldsymbol{t}\in T_J}\left(\sqrt{m}\,\|\boldsymbol{t}\|+\boldsymbol{g}\cdot\boldsymbol{t}\right)+\mathrm{C}B^2\sqrt{k}\right)_+^2 && \text{(Lemma 14.9)}\\
&= \left(\mathscr{E}_m(T_J)+\mathrm{C}B^2\sqrt{k}\right)_+^2. && \text{(Definition 4.2)}
\end{aligned}$$

The first step is a type of Poincaré inequality, which requires some specialized concentration results for the restricted singular value of the hybrid matrix. The second step of this chain involves a convex duality argument that is not present in the proof of the lower bound (14.1). We have used the assumption that $k\geq m^{1/2}$ to simplify the error term when we pass from the second line to the third. Altogether, this yields the bound (14.2).

14.2. **Proposition 14.1: Moment Comparison Inequality for the Hybrid RSV.** We require a moment comparison inequality to pass from the expected square of the restricted singular value of the hybrid matrix to the expectation of the restricted singular value itself.

Lemma 14.2 (Proposition 14.1: Moment Comparison)**.** *Adopt the notation and hypotheses of Proposition 14.1. Then*

$$\mathbb{E}\min_{\boldsymbol{t}\in T_J}\left\|\boldsymbol{\Phi}_J\boldsymbol{t}_J+\boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\|^2 \leq \left(\mathbb{E}\min_{\boldsymbol{t}\in T_J}\left\|\boldsymbol{\Phi}_J\boldsymbol{t}_J+\boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\|+\mathrm{C}Bm^{1/4}\right)^2.$$

Proof. Define the random variable

$$X := \min_{\boldsymbol{t}\in T_J}\left\|\boldsymbol{\Phi}_J\boldsymbol{t}_J+\boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\|.$$

We need a bound for $\mathbb{E}\, X^2$. We obtain this result by applying a moment comparison inequality for $\mathbf{\Phi}$, conditional on $\mathbf{\Gamma}$, and then we apply a moment comparison inequality for $\mathbf{\Gamma}$.

Sublemma 14.3, applied conditionally, with the choice $\boldsymbol{A} = \mathbf{\Gamma}$, gives

$$\mathbb{E}\, X^2 = \mathbb{E}\left[\mathbb{E}\left[X^2 \,|\, \mathbf{\Gamma}\right]\right] \le \mathbb{E}\left[\left(\mathbb{E}[X \,|\, \mathbf{\Gamma}] + CBm^{1/4}\right)^2\right].$$

The function $\mathbf{\Gamma} \mapsto \mathbb{E}[X \,|\, \mathbf{\Gamma}] + CBm^{1/4}$ is 1-Lipschitz, so the Gaussian variance inequality, Fact A.1, implies that

$$\mathbb{E}\left[\left(\mathbb{E}[X \,|\, \mathbf{\Gamma}] + CBm^{1/4}\right)^2\right] \le \left(\mathbb{E}\, X + CBm^{1/4}\right)^2 + 1.$$

Combine the last two displays, and adjust the constant to complete the proof. □

The proof of Lemma 14.2 requires a separate moment comparison result for a random variable that depends only on the original matrix $\mathbf{\Phi}_J$.

Sublemma 14.3 (Lemma 14.2: Moment Comparison for Original Matrix)**.** *Adopt the notation and hypotheses of Lemma 14.2. For a fixed $m \times n$ matrix $\boldsymbol{A}$,*

$$\mathbb{E} \min_{\boldsymbol{t} \in T_J} \left\| \mathbf{\Phi}_J \boldsymbol{t}_J + \boldsymbol{A}_{J^c} \boldsymbol{t}_{J^c} \right\|^2 \le \left(\mathbb{E} \min_{\boldsymbol{t} \in T_J} \left\| \mathbf{\Phi}_J \boldsymbol{t}_J + \boldsymbol{A}_{J^c} \boldsymbol{t}_{J^c} \right\| + CBm^{1/4} \right)^2. \tag{14.3}$$

Proof. Define the random variable

$$X := \min_{\boldsymbol{t} \in T_J} \left\| \mathbf{\Phi}_J \boldsymbol{t}_J + \boldsymbol{A}_{J^c} \boldsymbol{t}_{J^c} \right\|.$$

First, observe that

$$(\mathbb{E}\, X^2)^{1/2} - \mathbb{E}\, X = \left((\mathbb{E}\, X^2)^{1/2} - \mathbb{E}\, X\right)_+ = \left(\mathbb{E}\, X - (\mathbb{E}\, X^2)^{1/2}\right)_- \le \mathbb{E}\left(X - (\mathbb{E}\, X^2)^{1/2}\right)_-. \tag{14.4}$$

The last inequality is Jensen's. We can use concentration to bound this quantity. *Mutatis mutandis*, repeat the proof of equation (11.1) from Proposition 11.1 to see that, for all $\zeta \ge 0$,

$$\mathbb{P}\left\{X^2 \le \mathbb{E}\, X^2 - CB^2\zeta\right\} \le \mathrm{e}^{-\zeta^2/m}.$$

Invoke the subadditivity of the square root and change variables:

$$\mathbb{P}\left\{X \le (\mathbb{E}\, X^2)^{1/2} - CB\zeta\right\} \le \mathrm{e}^{-\zeta^4/m}. \tag{14.5}$$

Jensen's inequality and integration by parts deliver

$$\mathbb{E}\left(X - (\mathbb{E}\, X^2)^{1/2}\right)_- = \int_0^\infty \mathbb{P}\left\{X - (\mathbb{E}\, X^2)^{1/2} \le -\zeta\right\} \mathrm{d}\zeta \le \int_0^\infty \mathrm{e}^{-\zeta^4/(CB^4 m)} \,\mathrm{d}\zeta \le CBm^{1/4}.$$

The third inequality follows from (14.5). Introduce the last display into (14.4) to reach

$$(\mathbb{E}\, X^2)^{1/2} - \mathbb{E}\, X \le CBm^{1/4}.$$

Rearrange this relation to complete to proof of (14.3). □

14.3. **Proposition 14.1: The Role of the Gaussian Minimax Theorem.** To prove Proposition 14.1, we must replace the Gaussian matrix in the quantity of interest with a pair of Gaussian vectors. The key to this argument is the following technical result.

Lemma 14.4 (Proposition 14.1: Application of Gaussian Minimax Theorem)**.** *Adopt the notation and hypotheses of Proposition 14.1. Let $\boldsymbol{A}$ be a fixed $m \times n$ matrix, and let $\boldsymbol{g} \in \mathbb{R}^n$ and $\boldsymbol{h} \in \mathbb{R}^m$ be independent standard normal random vectors. Then, for all $\zeta \in \mathbb{R}$,*

$$\mathbb{P}\left\{\min_{\boldsymbol{t} \in T_J} \left\| \boldsymbol{A}_J \boldsymbol{t}_J + \mathbf{\Gamma}_{J^c} \boldsymbol{t}_{J^c} \right\| \le \zeta\right\} \le 2\,\mathbb{P}\left\{\min_{\boldsymbol{t} \in T_J} \left(\left\| \boldsymbol{A}_J \boldsymbol{t}_J + \boldsymbol{h} \left\| \boldsymbol{t}_{J^c} \right\| \right\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) \le \zeta\right\}. \tag{14.6}$$

Furthermore, if T_J is convex and $\zeta \ge 0$,

$$\mathbb{P}\left\{\min_{\boldsymbol{t} \in T_J} \left\| \boldsymbol{A}_J \boldsymbol{t}_J + \mathbf{\Gamma}_{J^c} \boldsymbol{t}_{J^c} \right\| \ge \zeta\right\} \le 2\,\mathbb{P}\left\{\min_{\boldsymbol{t} \in T_J} \left(\left\| \boldsymbol{A}_J \boldsymbol{t}_J + \boldsymbol{h} \left\| \boldsymbol{t}_{J^c} \right\| \right\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) \ge \zeta\right\}. \tag{14.7}$$

If $\zeta \le 0$, the left-hand side is trivially equal to one.

This result depends on the Gaussian Minimax Theorem [Gor85]; see Fact A.3 for a statement. Lemma 14.4 is similar with early results of Gordon [Gor88, Cor. 1.2]. The detailed argument here is adapted from [TOH15, Thm. 2.1]; see also Stojnic [Sto13].

Proof. The basic idea is to express the quantity of interest as the value of a saddle-point problem:

$$\min_{\boldsymbol{t}\in T_J} \left\| \boldsymbol{A}_J \boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c} \boldsymbol{t}_{J^c} \right\| = \min_{\boldsymbol{t}\in T_J} \max_{\boldsymbol{u}\in \mathsf{B}^m} \boldsymbol{u} \cdot \left(\boldsymbol{A}_J \boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c} \boldsymbol{t}_{J^c} \right).$$

Then we apply the Gaussian Minimax Theorem to obtain probabilistic lower bounds. When T_J is convex, we can also invoke convex duality to interchange the minimum and maximum, which leads to complementary bounds. To proceed with this approach, however, it is convenient to work with a slightly different minimax problem.

Define the deterministic function

$$\lambda(\boldsymbol{t}, \boldsymbol{u}) = \boldsymbol{u} \cdot (\boldsymbol{A}_J \boldsymbol{t}_J) \quad \text{for } \boldsymbol{t} \in T_J \text{ and } \boldsymbol{u} \in \mathsf{B}^m.$$

Let γ be a standard normal random variable, independent from everything else. Introduce two centered Gaussian processes:

$$X(\boldsymbol{t}, \boldsymbol{u}) := \boldsymbol{u} \cdot (\boldsymbol{\Gamma}_{J^c} \boldsymbol{t}_{J^c}) + \|\boldsymbol{u}\| \, \|\boldsymbol{t}_{J^c}\| \, \gamma \quad \text{and} \quad Y(\boldsymbol{t}, \boldsymbol{u}) := (\boldsymbol{u} \cdot \boldsymbol{h}) \, \|\boldsymbol{t}_{J^c}\| + \|\boldsymbol{u}\| \, (\boldsymbol{g} \cdot \boldsymbol{t}_{J^c})$$

indexed by $\boldsymbol{t} \in T_J$ and $\boldsymbol{u} \in \mathsf{B}^m$. Let us verify that these processes satisfy the conditions required by the Gaussian Minimax Theorem, Fact A.3. First, for all parameters $\boldsymbol{t} \in T_J$ and $\boldsymbol{u} \in \mathsf{B}^m$,

$$\mathbb{E}\, X(\boldsymbol{t}, \boldsymbol{u})^2 = 2 \|\boldsymbol{u}\|^2 \|\boldsymbol{t}_{J^c}\|^2 = \mathbb{E}\, Y(\boldsymbol{t}, \boldsymbol{u})^2. \tag{14.8}$$

Second, for all parameters $\boldsymbol{t}, \boldsymbol{t}' \in T_J$ and $\boldsymbol{u}, \boldsymbol{u}' \in \mathsf{B}^m$,

$$\mathbb{E}\left[X(\boldsymbol{t}, \boldsymbol{u}) X(\boldsymbol{t}', \boldsymbol{u}') \right] - \mathbb{E}\left[Y(\boldsymbol{t}, \boldsymbol{u}) Y(\boldsymbol{t}', \boldsymbol{u}') \right] = \left(\|\boldsymbol{u}\| \, \|\boldsymbol{u}'\| - \boldsymbol{u} \cdot \boldsymbol{u}' \right) \left(\|\boldsymbol{t}_{J^c}\| \, \|\boldsymbol{t}'_{J^c}\| - \boldsymbol{t}_{J^c} \cdot \boldsymbol{t}'_{J^c} \right).$$

By the Cauchy–Schwarz inequality,

$$\mathbb{E}\left[X(\boldsymbol{t}, \boldsymbol{u}) X(\boldsymbol{t}, \boldsymbol{u}') \right] = \mathbb{E}\left[Y(\boldsymbol{t}, \boldsymbol{u}) Y(\boldsymbol{t}, \boldsymbol{u}') \right]; \tag{14.9}$$

$$\mathbb{E}\left[X(\boldsymbol{t}, \boldsymbol{u}) X(\boldsymbol{t}', \boldsymbol{u}) \right] = \mathbb{E}\left[Y(\boldsymbol{t}, \boldsymbol{u}) Y(\boldsymbol{t}', \boldsymbol{u}) \right]; \tag{14.10}$$

$$\mathbb{E}\left[X(\boldsymbol{t}, \boldsymbol{u}) X(\boldsymbol{t}', \boldsymbol{u}') \right] \geq \mathbb{E}\left[Y(\boldsymbol{t}, \boldsymbol{u}) Y(\boldsymbol{t}', \boldsymbol{u}') \right]. \tag{14.11}$$

The formulas (14.8), (14.9), and (14.11) verify the conditions of the Gaussian Minimax Theorem. Fact A.3 delivers the bound

$$\mathbb{P}\left\{ \min_{\boldsymbol{t}\in T_J} \max_{\boldsymbol{u}\in \mathsf{B}^m} \left(\lambda(\boldsymbol{t}, \boldsymbol{u}) + X(\boldsymbol{t}, \boldsymbol{u}) \right) > \zeta \right\} \geq \mathbb{P}\left\{ \min_{\boldsymbol{t}\in T_J} \max_{\boldsymbol{u}\in \mathsf{B}^m} \left(\lambda(\boldsymbol{t}, \boldsymbol{u}) + Y(\boldsymbol{t}, \boldsymbol{u}) \right) > \zeta \right\}. \tag{14.12}$$

This estimate does involve a small technicality. We can only apply the Gaussian Minimax Theorem to a finite subset of $T_J \times \mathsf{B}^m$, so we must make an approximation argument to pass to the entire set. We omit the details.

Now, let us determine the values of the saddle-point problems in (14.12). First,

$$\begin{aligned} \min_{\boldsymbol{t}\in T_J} \max_{\boldsymbol{u}\in \mathsf{B}^m} \left(\lambda(\boldsymbol{t}, \boldsymbol{u}) + X(\boldsymbol{t}, \boldsymbol{u}) \right) &= \min_{\boldsymbol{t}\in T_J} \max_{\boldsymbol{u}\in \mathsf{B}^m} \left(\boldsymbol{u} \cdot (\boldsymbol{A}_J \boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c} \boldsymbol{t}_{J^c}) + \|\boldsymbol{u}\| \, \|\boldsymbol{t}_{J^c}\| \, \gamma \right) \\ &= \min_{\boldsymbol{t}\in T_J} \left(\left\| \boldsymbol{A}_J \boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c} \boldsymbol{t}_{J^c} \right\| + \|\boldsymbol{t}_{J^c}\| \, \gamma \right)_+. \end{aligned} \tag{14.13}$$

Similarly,

$$\min_{\boldsymbol{t}\in T_J} \max_{\boldsymbol{u}\in \mathsf{B}^m} \left(\lambda(\boldsymbol{t}, \boldsymbol{u}) + Y(\boldsymbol{t}, \boldsymbol{u}) \right) \geq \min_{\boldsymbol{t}\in T_J} \left(\left\| \boldsymbol{A}_J \boldsymbol{t}_J + \boldsymbol{h} \, \|\boldsymbol{t}_{J^c}\| \right\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right).$$

Next, we remove the term involving γ from the right-hand side of (14.13). To do so, we condition on $\gamma > 0$ and $\gamma \le 0$ and calculate that

$$\begin{aligned}
\mathbb{P}\left\{\min_{\boldsymbol{t}\in T_J}\max_{\boldsymbol{u}\in \mathsf{B}^m}\big(\lambda(\boldsymbol{t},\boldsymbol{u}) + X(\boldsymbol{t},\boldsymbol{u})\big) > \zeta\right\} &= \mathbb{P}\left\{\min_{\boldsymbol{t}\in T_J}\max_{\boldsymbol{u}\in \mathsf{B}^m}\big(\left\|\boldsymbol{A}_J\boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\| + \|\boldsymbol{t}_{J^c}\|\,\gamma\big)_+ > \zeta\right\} \\
&\le \frac{1}{2} + \frac{1}{2}\mathbb{P}\left\{\min_{\boldsymbol{t}\in T_J}\big(\left\|\boldsymbol{A}_J\boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\| + \|\boldsymbol{t}_{J^c}\|\,\gamma\big)_+ > \zeta \,\middle|\, \gamma \le 0\right\} \\
&\le \frac{1}{2} + \frac{1}{2}\mathbb{P}\left\{\min_{\boldsymbol{t}\in T_J}\big(\left\|\boldsymbol{A}_J\boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\|\big)_+ > \zeta\right\} \\
&= \frac{1}{2} + \frac{1}{2}\mathbb{P}\left\{\min_{\boldsymbol{t}\in T_J}\left\|\boldsymbol{A}_J\boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\| > \zeta\right\}.
\end{aligned}$$

On the other hand,

$$\mathbb{P}\left\{\min_{\boldsymbol{t}\in T_J}\max_{\boldsymbol{u}\in \mathsf{B}^m}\big(\lambda(\boldsymbol{t},\boldsymbol{u}) + Y(\boldsymbol{t},\boldsymbol{u})\big) > \zeta\right\} \ge \mathbb{P}\left\{\min_{\boldsymbol{t}\in T_J}\big(\left\|\boldsymbol{A}_J\boldsymbol{t}_J + \boldsymbol{h}\,\|\boldsymbol{t}_{J^c}\|\right\| + \boldsymbol{g}_{J^c}\cdot\boldsymbol{t}_{J^c}\big) > \zeta\right\}.$$

Introduce the last two displays into (14.12) to reach

$$\frac{1}{2} + \frac{1}{2}\mathbb{P}\left\{\min_{\boldsymbol{t}\in T_J}\left\|\boldsymbol{A}_J\boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\| > \zeta\right\} \ge \mathbb{P}\left\{\min_{\boldsymbol{t}\in T_J}\big(\left\|\boldsymbol{A}_J\boldsymbol{t}_J + \boldsymbol{h}\,\|\boldsymbol{t}_{J^c}\|\right\| + \boldsymbol{g}_{J^c}\cdot\boldsymbol{t}_{J^c}\big) > \zeta\right\}.$$

Take the complements of both probabilities and rearrange to conclude that (14.6) is correct.

The second result (14.7) requires an additional duality argument. If we replace the function λ and the random processes X and Y with their negations, all of the variance calculations above remain valid. In particular, the relations (14.8), (14.10), and (14.11) permit us to apply Fact A.3 with the roles of T_J and B^m reversed. This step yields

$$\mathbb{P}\left\{\min_{\boldsymbol{u}\in \mathsf{B}^m}\max_{\boldsymbol{t}\in T_J}\big(-\lambda(\boldsymbol{t},\boldsymbol{u}) - X(\boldsymbol{t},\boldsymbol{u})\big) > -\zeta\right\} \ge \mathbb{P}\left\{\min_{\boldsymbol{u}\in \mathsf{B}^m}\max_{\boldsymbol{t}\in T_J}\big(-\lambda(\boldsymbol{t},\boldsymbol{u}) - Y(\boldsymbol{t},\boldsymbol{u})\big) > -\zeta\right\}.$$

Let us examine the saddle-point problems. Since $-\lambda - X$ is bilinear and the sets B^m and T_J are compact and convex, the Sion Minimax Theorem [Sio58] allows us to interchange the minimum and maximum. Thus

$$\begin{aligned}
\min_{\boldsymbol{u}\in \mathsf{B}^m}\max_{\boldsymbol{t}\in T_J}\big(-\lambda(\boldsymbol{t},\boldsymbol{u}) - X(\boldsymbol{t},\boldsymbol{u})\big) &= \max_{\boldsymbol{t}\in T_J}\min_{\boldsymbol{u}\in \mathsf{B}^m}\big(-\lambda(\boldsymbol{t},\boldsymbol{u}) - X(\boldsymbol{t},\boldsymbol{u})\big) \\
&= -\min_{\boldsymbol{t}\in T_J}\max_{\boldsymbol{u}\in \mathsf{B}^m}\big(\lambda(\boldsymbol{t},\boldsymbol{u}) + X(\boldsymbol{t},\boldsymbol{u})\big) \\
&= -\min_{\boldsymbol{t}\in T_J}\big(\left\|\boldsymbol{A}_J\boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\| + \|\boldsymbol{t}_{J^c}\|\,\gamma\big)_+.
\end{aligned}$$

Similarly,

$$\min_{\boldsymbol{u}\in \mathsf{B}^m}\max_{\boldsymbol{t}\in T_J}\big(-\lambda(\boldsymbol{t},\boldsymbol{u}) - Y(\boldsymbol{t},\boldsymbol{u})\big) = -\min_{\boldsymbol{t}\in T_J}\big(\left\|\boldsymbol{A}_J\boldsymbol{t}_J + \boldsymbol{h}\,\|\boldsymbol{t}_{J^c}\|\right\| + \boldsymbol{g}_{J^c}\cdot\boldsymbol{t}_{J^c}\big)_+.$$

Combining the last three displays, we reach

$$\mathbb{P}\left\{\min_{\boldsymbol{t}\in T_J}\big(\left\|\boldsymbol{A}_J\boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\| + \|\boldsymbol{t}_{J^c}\|\,\gamma\big)_+ < \zeta\right\} \ge \mathbb{P}\left\{\min_{\boldsymbol{t}\in T_J}\big(\left\|\boldsymbol{A}_J\boldsymbol{t}_J + \boldsymbol{h}\,\|\boldsymbol{t}_{J^c}\|\right\| + \boldsymbol{g}_{J^c}\cdot\boldsymbol{t}_{J^c}\big)_+ < \zeta\right\}.$$

Proceeding as before, conditioning on the sign of γ, we may remove the dependence on γ from the left-hand side to obtain

$$\mathbb{P}\left\{\min_{\boldsymbol{t}\in T_J}\left\|\boldsymbol{A}_J\boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c}\boldsymbol{t}_{J^c}\right\| \ge \zeta\right\} \ge 2\mathbb{P}\left\{\min_{\boldsymbol{t}\in T_J}\big(\left\|\boldsymbol{A}_J\boldsymbol{t}_J + \boldsymbol{h}\,\|\boldsymbol{t}_{J^c}\|\right\| + \boldsymbol{g}_{J^c}\cdot\boldsymbol{t}_{J^c}\big)_+ \ge \zeta\right\}.$$

Clearly, this inequality is useful only for $\zeta \ge 0$ in which case the right hand side is equal to

$$2\mathbb{P}\left\{\min_{\boldsymbol{t}\in T_J}\left\|\boldsymbol{A}_J\boldsymbol{t}_J + \boldsymbol{h}\,\|\boldsymbol{t}_{J^c}\|\right\| + \boldsymbol{g}_{J^c}\cdot\boldsymbol{t}_{J^c} \ge \zeta\right\}.$$

We confirm that (14.7) is true. □

14.4. **Proposition 14.1: Reducing the Gaussian Matrix to Some Gaussian Vectors.** Our next goal is to convert the probability bounds from Lemma 14.4 into expectation bounds. Lemma 14.5 gives a lower bound that is valid for every subset T_J of the unit ball, and Lemma 14.6 gives an upper bound that is valid when T_J is also convex.

Lemma 14.5 (Proposition 14.1: Reducing the Gaussian Matrix—Lower Bound)**.** *Adopt the notation and hypotheses of Proposition 14.1. Let $\boldsymbol{A}$ be a fixed $m \times n$ matrix, and let $\boldsymbol{g} \in \mathbb{R}^n$ and $\boldsymbol{h} \in \mathbb{R}^m$ be independent standard normal vectors. Then*

$$\mathbb{E} \min_{\boldsymbol{t} \in T_J} \left\| \boldsymbol{A}_J \boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c} \boldsymbol{t}_{J^c} \right\| \geq \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\left\| \boldsymbol{A}_J \boldsymbol{t}_J + \boldsymbol{h} \| \boldsymbol{t}_{J^c} \| \right\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) - 2. \tag{14.14}$$

In particular,

$$\mathbb{E} \min_{\boldsymbol{t} \in T_J} \left\| \boldsymbol{\Phi}_J \boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c} \boldsymbol{t}_{J^c} \right\| \geq \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\left\| \boldsymbol{\Phi}_J \boldsymbol{t}_J + \boldsymbol{h} \| \boldsymbol{t}_{J^c} \| \right\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) - 2.$$

Proof. We make the abbreviations

$$X := \min_{\boldsymbol{t} \in T_J} \left\| \boldsymbol{A}_J \boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c} \boldsymbol{t}_{J^c} \right\| \quad \text{and} \quad Y := \min_{\boldsymbol{t} \in T_J} \left(\left\| \boldsymbol{A}_J \boldsymbol{t}_J + \boldsymbol{h} \| \boldsymbol{t}_{J^c} \| \right\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right).$$

First, note that

$$\mathbb{E}\, Y - \mathbb{E}\, X \leq (\mathbb{E}\, Y - \mathbb{E}\, X)_+ = (\mathbb{E}\, X - \mathbb{E}\, Y)_- \leq \mathbb{E}(X - \mathbb{E}\, Y)_-.$$

The last inequality is Jensen's. To bound the right-hand side, we use integration by parts and formula (14.6) from Lemma 14.4:

$$\mathbb{E}(X - \mathbb{E}\, Y)_- = \int_0^\infty \mathbb{P}\{X - \mathbb{E}\, Y \leq -\zeta\} \, \mathrm{d}\zeta \leq \int_0^\infty 2\, \mathbb{P}\{Y - \mathbb{E}\, Y \leq -\zeta\} \, \mathrm{d}\zeta = 2\, \mathbb{E}(Y - \mathbb{E}\, Y)_-.$$

Finally, we make the estimates

$$\mathbb{E}[(Y - \mathbb{E}\, Y)_-] \leq \mathrm{Var}[Y]^{1/2} \leq 1.$$

The last inequality follows from the Gaussian variance inequality, Fact A.1. Indeed, the function $(\boldsymbol{g}, \boldsymbol{h}) \mapsto Y(\boldsymbol{g}, \boldsymbol{h})$ is 1-Lipschitz because the set T_J is contained in the unit ball. In summary,

$$\mathbb{E}\, Y - \mathbb{E}\, X \leq \mathbb{E}(X - \mathbb{E}\, Y)_- \leq 2\, \mathbb{E}(Y - \mathbb{E}\, Y)_- \leq 2.$$

We arrive the advertised bound (14.14). □

Lemma 14.6 (Proposition 14.1: Reducing the Gaussian Matrix—Upper Bound)**.** *Adopt the notation and hypotheses of Proposition 14.1. Let $\boldsymbol{A}$ be a fixed $m \times n$ matrix, and let $\boldsymbol{g} \in \mathbb{R}^n$ and $\boldsymbol{h} \in \mathbb{R}^m$ be independent standard normal vectors. Then*

$$\mathbb{E} \min_{\boldsymbol{t} \in T_J} \left\| \boldsymbol{A}_J \boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c} \boldsymbol{t}_{J^c} \right\| \leq \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\left\| \boldsymbol{A}_J \boldsymbol{t}_J + \boldsymbol{h} \| \boldsymbol{t}_{J^c} \| \right\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) + 2.$$

In particular,

$$\mathbb{E} \min_{\boldsymbol{t} \in T_J} \left\| \boldsymbol{\Phi}_J \boldsymbol{t}_J + \boldsymbol{\Gamma}_{J^c} \boldsymbol{t}_{J^c} \right\| \leq \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\left\| \boldsymbol{\Phi}_J \boldsymbol{t}_J + \boldsymbol{h} \| \boldsymbol{t}_{J^c} \| \right\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) + 2.$$

Proof. We follow the same pattern as in Lemma 14.5. Using the same notation, we calculate that

$$\mathbb{E}\, X - \mathbb{E}\, Y \leq (\mathbb{E}\, X - \mathbb{E}\, Y)_+ \leq \mathbb{E}(X - \mathbb{E}\, Y)_+ \leq 2\, \mathbb{E}(Y - \mathbb{E}\, Y)_+ \leq 2.$$

In this case, we invoke (14.7) from Lemma 14.4 to obtain the penultimate inequality. □

14.5. **Proposition 14.1: Removing the Remaining Part of the Original Random Matrix.** The next step in the proof of Proposition 14.1 is to remove the remaining section of the random matrix $\mathbf{\Phi}$ from the bounds in Lemmas 14.5 and 14.6.

Lemma 14.7 (Proposition 14.1: Removing the Original Matrix)**.** *Adopt the notation and hypotheses of Proposition 14.1. Let $\boldsymbol{X}$ be an $m \times n$ random matrix with independent, standardized entries that satisfy Assumption B.1. Then*

$$\left| \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\left\| \boldsymbol{X}_J \boldsymbol{t}_J + \boldsymbol{h} \, \| \boldsymbol{t}_{J^c} \| \right\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) - \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m} \, \| \boldsymbol{t} \| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) \right| \le \mathrm{C} B^2 \sqrt{k}.$$

In particular, this conclusion is valid when $\boldsymbol{X} = \mathbf{\Phi}$.

Proof. Abbreviate $\mathbf{\Psi} := \begin{bmatrix} \boldsymbol{X}_J & \boldsymbol{h} \end{bmatrix}$. Observe that

$$\boldsymbol{X}_J \boldsymbol{t}_J + \boldsymbol{h} \, \| \boldsymbol{t}_{J^c} \| = \mathbf{\Psi} \begin{bmatrix} \boldsymbol{t}_J \\ \| \boldsymbol{t}_{J^c} \| \end{bmatrix} \quad \text{and} \quad \left\| \begin{bmatrix} \boldsymbol{t}_J \\ \| \boldsymbol{t}_{J^c} \| \end{bmatrix} \right\| = \| \boldsymbol{t} \|.$$

Therefore, we have the deterministic bounds

$$\sigma_{\min}(\mathbf{\Psi}) \, \| \boldsymbol{t} \| \le \left\| \boldsymbol{X}_J \boldsymbol{t}_J + \boldsymbol{h} \, \| \boldsymbol{t}_{J^c} \| \right\| \le \sigma_{\max}(\mathbf{\Psi}) \, \| \boldsymbol{t} \|. \tag{14.15}$$

The $m \times (k+1)$ random matrix $\mathbf{\Psi}$ has independent, standardized entries that satisfy (B.1). For all $\zeta \ge 0$, Fact B.1 shows that its extreme singular values satisfy the probability bounds

$$\begin{aligned} \mathbb{P} \left\{ \sigma_{\max}(\mathbf{\Psi}) \ge \sqrt{m} + \mathrm{C}_0 B^2 \sqrt{k} + \zeta \right\} &\le \mathrm{e}^{-\mathrm{c}_0 \zeta^2 / B^4}, \quad \text{and} \\ \mathbb{P} \left\{ \sigma_{\min}(\mathbf{\Psi}) \le \sqrt{m} - \mathrm{C}_0 B^2 \sqrt{k} - \zeta \right\} &\le \mathrm{e}^{-\mathrm{c}_0 \zeta^2 / B^4}. \end{aligned} \tag{14.16}$$

These inequalities allow us to treat the singular values of $\mathbf{\Psi}$ as if they were equal to $\sqrt{m}$.

We may now perform the following estimates:

$$\begin{aligned} \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\left\| \boldsymbol{X}_J \boldsymbol{t}_J + \boldsymbol{h} \, \| \boldsymbol{t}_{J^c} \| \right\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) &\le \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sigma_{\max}(\mathbf{\Psi}) \, \| \boldsymbol{t} \| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) \\ &\le \mathbb{E} \left[\min_{\boldsymbol{t} \in T_J} \left(\sqrt{m} \, \| \boldsymbol{t} \| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) + \left(\sigma_{\max}(\mathbf{\Psi}) - \sqrt{m} \right)_+ \max_{\boldsymbol{t} \in T_J} \| \boldsymbol{t} \| \right] \\ &\le \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m} \, \| \boldsymbol{t} \| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) + \mathbb{E} \left(\sigma_{\max}(\mathbf{\Psi}) - \sqrt{m} \right)_+. \end{aligned}$$

The first inequality is (14.15). Then we add and subtract $\sqrt{m}$ from the maximum singular value. Last, we recall that T_J is a subset of the unit ball. Set $\alpha := \mathrm{C}_0 B^2 \sqrt{k}$, and calculate that

$$\begin{aligned} \mathbb{E} \left(\sigma_{\max}(\mathbf{\Psi}) - \sqrt{m} \right)_+ &= \int_0^\infty \mathbb{P} \left\{ \sigma_{\max}(\mathbf{\Psi}) - \sqrt{m} \ge \zeta \right\} \mathrm{d}\zeta \\ &\le \int_0^\alpha \mathrm{d}\zeta + \int_0^\infty \mathbb{P} \left\{ \sigma_{\max}(\mathbf{\Psi}) - \sqrt{m} \ge \alpha + \zeta \right\} \mathrm{d}\zeta \\ &\le \alpha + \int_0^\infty \mathrm{e}^{-\mathrm{c}_0 \zeta^2 / B^4} \, \mathrm{d}\zeta = \alpha + \mathrm{C} B^2. \end{aligned} \tag{14.17}$$

We split the integral at α, and we change variables in the second integral. For the first integral, we use the trivial bound of one on the probability. Then we invoke the probability inequality (14.16).

Combining the last two displays and collecting constants, we arrive at

$$\mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\left\| \boldsymbol{X}_J \boldsymbol{t}_J + \boldsymbol{h} \, \| \boldsymbol{t}_{J^c} \| \right\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) \le \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m} \, \| \boldsymbol{t} \| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) + \mathrm{C} B^2 \sqrt{k}.$$

An entirely similar argument delivers a matching lower bound. Together, these estimates complete the proof. □

14.6. **Proposition 14.1: Replacing the Coordinates Missing from the Excess Width.** The last step in the proof of Proposition 14.1 is to examine the excess-width-like functional from Lemma 14.7 that involves the term $\boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c}$. We must show that this term does not change very much if we reintroduce the coordinates listed in J. Lemma 14.8 gives the easy proof of the lower bound. Lemma 14.9 contains the upper bound.

Lemma 14.8 (Proposition 14.1: Missing Coordinates—Lower Bound)**.** *Adopt the notation and hypotheses of Proposition 14.1. Then*

$$\mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m}\, \|\boldsymbol{t}\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) \geq \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m}\, \|\boldsymbol{t}\| + \boldsymbol{g} \cdot \boldsymbol{t} \right).$$

Proof. This result is an immediate consequence of Jensen's inequality:

$$\begin{aligned} \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m}\, \|\boldsymbol{t}\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) &= \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m}\, \|\boldsymbol{t}\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} + \mathbb{E}_{\boldsymbol{g}_J} (\boldsymbol{g}_J \cdot \boldsymbol{t}_J) \right) \\ &\geq \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m}\, \|\boldsymbol{t}\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} + \boldsymbol{g}_J \cdot \boldsymbol{t}_J \right) \\ &= \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m}\, \|\boldsymbol{t}\| + \boldsymbol{g} \cdot \boldsymbol{t} \right) \end{aligned}$$

We rely on the fact that $\boldsymbol{g}_J$ and $\boldsymbol{g}_{J^c}$ are independent standard normal vectors. □

Lemma 14.9 (Proposition 14.1: Missing Coordinates—Upper Bound)**.** *Adopt the notation and hypotheses of Proposition 14.1. Then*

$$\mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m}\, \|\boldsymbol{t}\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) \leq \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m}\, \|\boldsymbol{t}\| + \boldsymbol{g} \cdot \boldsymbol{t} \right) + \sqrt{k}.$$

Proof. This calculation is also easy:

$$\begin{aligned} \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m}\, \|\boldsymbol{t}\| + \boldsymbol{g}_{J^c} \cdot \boldsymbol{t}_{J^c} \right) &= \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m}\, \|\boldsymbol{t}\| + \boldsymbol{g} \cdot \boldsymbol{t} - \boldsymbol{g}_J \cdot \boldsymbol{t}_J \right) \\ &\leq \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m}\, \|\boldsymbol{t}\| + \boldsymbol{g} \cdot \boldsymbol{t} \right) + \mathbb{E}\, \|\boldsymbol{g}_J\| \\ &\leq \mathbb{E} \min_{\boldsymbol{t} \in T_J} \left(\sqrt{m}\, \|\boldsymbol{t}\| + \boldsymbol{g} \cdot \boldsymbol{t} \right) + \sqrt{k}. \end{aligned}$$

In the last step, we have used the fact that $\#J = k$. □

15. Proof of Corollary 10.1 from Theorem 9.1 by Truncation

In this section, we show how to establish Corollary 10.1 as a consequence of Theorem 9.1. The proof depends on a truncation argument.

15.1. **Proof of Corollary 10.1.** Fix parameters $p > 4$ and $\nu \geq 1$. Assume that $\boldsymbol{\Phi}$ is an $m \times n$ matrix that follows Model 2.4 with parameters p and ν. Let T be a compact subset of the unit ball B^n. We will prove the lower bound for the minimum singular value of $\boldsymbol{\Phi}$ restricted to T. When T is convex, an entirely similar approach yields the corresponding upper bound.

Fix a truncation parameter R that satisfies $R^{p/2-1} \geq 2\nu^{p/2}$. Decompose the random matrix $\boldsymbol{\Phi}$ as

$$\boldsymbol{\Phi} = \boldsymbol{\Phi}_{\text{trunc}} + \boldsymbol{\Phi}_{\text{tail}}$$

by applying the truncation described below in Lemma 15.1 separately to each entry of $\boldsymbol{\Phi}$. This procedure ensures that $\boldsymbol{\Phi}_{\text{trunc}}$ contains independent, symmetric, standardized entries, each bounded by $2R$. In other words, $\boldsymbol{\Phi}_{\text{trunc}}$ follows Model 2.1 with $B = 2R$. The tail $\boldsymbol{\Phi}_{\text{tail}}$ contains independent, centered entries, each with variance bounded by $C\nu^p R^{2-p}$ and whose pth moment is bounded by $(2\nu)^p$.

We can control the restricted singular values of $\boldsymbol{\Phi}$ using the triangle inequality:

$$\sigma_{\min}(\boldsymbol{\Phi}; T) = \min_{\boldsymbol{t} \in T} \|\boldsymbol{\Phi}\boldsymbol{t}\| \geq \min_{\boldsymbol{t} \in T} \|\boldsymbol{\Phi}_{\text{trunc}}\boldsymbol{t}\| - \|\boldsymbol{\Phi}_{\text{tail}}\| = \sigma_{\min}(\boldsymbol{\Phi}_{\text{trunc}}; T) - \|\boldsymbol{\Phi}_{\text{tail}}\|. \tag{15.1}$$

We bound the restricted singular value of the bounded matrix $\boldsymbol{\Phi}_{\text{trunc}}$ using Theorem 9.1. To bound $\|\boldsymbol{\Phi}_{\text{tail}}\|$, we apply a simple norm estimate, Fact B.2, based on the matrix Rosenthal inequality [Tro15c, Thm. I].

Since $\mathbf{\Phi}_{\text{trunc}}$ follows Model 2.1 with $B = 2R$, Theorem 9.1(1) and (2) give the probability bound

$$\mathbb{P}\left\{\sigma_{\min}^2(\mathbf{\Phi}_{\text{trunc}}; T) \le \left(\mathscr{E}_m(T)\right)_+^2 - CR^2(m+n)^{0.92} - CR^2\zeta\right\} \le \mathrm{e}^{-\zeta^2/(m+n)}.$$

Select $\zeta = (m+n)^{0.92}$ to make the tail probability negligible:

$$\mathbb{P}\left\{\sigma_{\min}^2(\mathbf{\Phi}_{\text{trunc}}; T) \le \left(\mathscr{E}_m(T)\right)_+^2 - CR^2(m+n)^{0.92}\right\} \le \mathrm{e}^{-(m+n)^{0.84}}.$$

Taking square roots inside the event, we reach

$$\mathbb{P}\left\{\sigma_{\min}(\mathbf{\Phi}_{\text{trunc}}; T) \le \left(\mathscr{E}_m(T)\right)_+ - CR(m+n)^{0.46}\right\} \le \mathrm{e}^{-(m+n)^{0.84}}. \tag{15.2}$$

This step depends on the subadditivity of the square root.

Meanwhile, the entries of $\mathbf{\Phi}_{\text{tail}}$ are centered, have variances at most $C\nu^p R^{2-p}$, and have pth moments bounded by $(2\nu)^p$. Therefore, we can apply the norm bound for heavy-tailed random matrices, Fact B.2, to see that

$$\mathbb{P}\left\{\|\mathbf{\Phi}_{\text{tail}}\| \ge C\sqrt{\nu^p R^{2-p}(m+n)\log(m+n)} + \left(C\nu(m+n)^{2/p}\log(m+n)\right)\zeta\right\} \le \zeta^{-p}.$$

Define the positive quantity ε via the relation $4(1+\varepsilon) := p$. Select $\zeta = (m+n)^{\varepsilon/p}$ to obtain

$$\mathbb{P}\left\{\|\mathbf{\Phi}_{\text{tail}}\| \ge C\nu^{p/2}R^{1-p/2}\sqrt{(m+n)\log(m+n)} + C\nu(m+n)^{(2+\varepsilon)/p}\log(m+n)\right\} \le (m+n)^{-\varepsilon}. \tag{15.3}$$

The key point here is that we can arrange for $\|\mathbf{\Phi}_{\text{tail}}\|$ to have order $o(\sqrt{m+n})$ with high probability.

Combine (15.1), (15.2), and (15.3) to reach

$$\begin{aligned}\mathbb{P}\Big\{\sigma_{\min}(\mathbf{\Phi}; T) \le \left(\mathscr{E}_m(T)\right)_+ - CR(m+n)^{0.46} \\ - C\nu^{p/2}R^{1-p/2}\sqrt{(m+n)\log(m+n)} - C\nu(m+n)^{(2+\varepsilon)/p}\log(m+n)\Big\} \\ \le \mathrm{e}^{-(m+n)^{0.84}} + (m+n)^{-\varepsilon}.\end{aligned}$$

Set the truncation parameter $R = \nu(m+n)^{0.02/(1+\varepsilon)}$ to equate the exponents on $m+n$ in the two terms that depend on R. Then simplify using $p = 4(1+\varepsilon)$ to obtain

$$\begin{aligned}\mathbb{P}\left\{\sigma_{\min}(\mathbf{\Psi}; T) \le \left(\mathscr{E}_m(T)\right)_+ - C\nu\left((m+n)^{0.5-0.02(1+2\varepsilon)/(1+\varepsilon)} + (m+n)^{0.5-(\varepsilon/4)/(1+\varepsilon)}\right)\log(m+n)\right\} \\ \le \mathrm{e}^{-(m+n)^{0.84}} + (m+n)^{-\varepsilon}.\end{aligned}$$

Note that both powers in the event are bounded away from 1/2, so we can absorb the logarithm by increasing the power slightly. Furthermore, we can introduce a function $\kappa(p)$ that is strictly positive for $p > 4$ to reach the inequality

$$\mathbb{P}\left\{\sigma_{\min}(\mathbf{\Psi}; T) \le \left(\mathscr{E}_m(T)\right)_+ - C_p\nu(m+n)^{0.5-\kappa(p)}\right\} \le \mathrm{e}^{-(m+n)^{0.84}} + (m+n)^{1-p/4}.$$

The constant C_p depends only on the parameter p. The exponential vanishes faster than any polynomial, so we can combine the terms on the right-hand side to complete the proof of (10.1).

For the upper bound, we use Theorem 9.1(3) to control the expectation of the restricted singular value. In this case, the error term in the expectation bound becomes $R^4(m+n)^{0.94}$. This change presents no new difficulties, and we arrive at the result (10.2) by a slight modification of the argument.

15.2. **Corollary 10.1: Truncation of Individual Random Variables.** In this section, we describe a truncation procedure for scalar random variables. The arguments here are entirely standard, but we include details for completeness.

Lemma 15.1 (Corollary 10.1: Truncation)**.** *Let φ be a random variable that satisfies the properties listed in Model 2.4. Let R be a parameter that satisfies $R^{p/2-1} \geq 2\nu^{p/2}$. Then we have the decomposition*

$$\varphi = \varphi_{\text{trunc}} + \varphi_{\text{tail}} \tag{15.4}$$

where

$$\mathbb{E}\,\varphi_{\text{trunc}} = 0, \quad \mathbb{E}\,\varphi_{\text{trunc}}^2 = 1, \quad \textit{and} \quad |\varphi_{\text{trunc}}| \leq 2R$$

and

$$\mathbb{E}\,\varphi_{\text{tail}} = 0, \quad \mathbb{E}\,\varphi_{\text{tail}}^2 \leq \frac{6\nu^p}{R^{p-2}}, \quad \textit{and} \quad \mathbb{E}\,|\varphi_{\text{tail}}|^p \leq (2\nu)^p.$$

Proof. Define the random variable

$$\varphi_{\text{trunc}} := \frac{\varphi \mathbb{1}\{|\varphi| \leq R\}}{\alpha} \quad \text{where} \quad \alpha^2 := \mathbb{E}\left[\varphi^2 \mathbb{1}\{|\varphi| \leq R\}\right].$$

Since φ is standardized and symmetric, φ_{trunc} is also standardized and symmetric. To ensure that the decomposition (15.4) holds, we must set

$$\varphi_{\text{tail}} := \varphi \mathbb{1}\{|\varphi| > R\} - \frac{1-\alpha}{\alpha}\varphi \mathbb{1}\{|\varphi| \leq R\}. \tag{15.5}$$

The random variable φ_{tail} is also centered because of the symmetry of φ.

To establish the other properties of φ_{trunc} and φ_{tail}, we need to calculate some expectations. First, using integration by parts and Markov's inequality,

$$\begin{aligned} \mathbb{E}\left[\varphi^2 \mathbb{1}\{|\varphi| > R\}\right] &= \int_0^R 2\zeta \mathbb{P}\{|\varphi| > R\}\,\mathrm{d}\zeta + \int_R^\infty 2\zeta \mathbb{P}\{|\varphi| > \zeta\}\,\mathrm{d}\zeta \\ &\leq \int_0^R 2\zeta \frac{\mathbb{E}\,|\varphi|^p}{R^p}\,\mathrm{d}\zeta + \int_R^\infty 2\zeta \frac{\mathbb{E}\,|\varphi|^p}{\zeta^p}\,\mathrm{d}\zeta \leq \frac{2\nu^p}{R^{p-2}}. \end{aligned} \tag{15.6}$$

In the last step, we used the assumption that $p \geq 4$. A similar calculation shows that

$$\begin{aligned} \alpha^2 = \mathbb{E}\left[\varphi^2 \mathbb{1}\{|\varphi| \leq R\}\right] &= \int_0^R 2\zeta \mathbb{P}\{|\varphi| > \zeta\}\,\mathrm{d}\zeta \\ &= \int_0^\infty 2\zeta \mathbb{P}\{|\varphi| > \zeta\}\,\mathrm{d}\zeta - \int_R^\infty 2\zeta \mathbb{P}\{|\varphi| > \zeta\}\,\mathrm{d}\zeta \\ &\geq \mathbb{E}\left[\varphi^2\right] - \frac{2\nu^p}{(p-2)R^{p-2}} \geq 1 - \frac{\nu^p}{R^{p-2}}. \end{aligned}$$

The last relation holds because φ is standardized. It follows that

$$\alpha \geq 1 - \frac{\nu^{p/2}}{R^{p/2-1}} \geq \frac{1}{2}. \tag{15.7}$$

The last estimate holds because $p \geq 4$ and of the assumption $R^{p/2-1} \geq 2\nu^{p/2}$.

We are now prepared to verify the uniform bound on φ_{trunc}:

$$|\varphi_{\text{trunc}}| = \left|\frac{\varphi \mathbb{1}\{|\varphi| \leq R\}}{\alpha}\right| \leq \frac{R}{\alpha} \leq 2R.$$

The last inequality follows from (15.7).

Next, we need to bound the variance of φ_{tail}. We have

$$\mathbb{E}\left[\varphi_{\text{tail}}^2\right] = \mathbb{E}\left[\varphi^2 \mathbb{1}\{|\varphi| > R\}\right] + \left(\frac{1-\alpha}{\alpha}\right)^2 \mathbb{E}\left[\varphi^2 \mathbb{1}\{|\varphi| \leq R\}\right] \leq \frac{2\nu^p}{R^{p-2}} + \left(\frac{\nu^{p/2}/R^{p/2-1}}{1/2}\right)^2 \mathbb{E}\left[\psi^2\right] \leq \frac{6\nu^p}{R^{p-2}}.$$

The first identity holds because the two indicators are orthogonal random variables. The second relation uses the expectation calculation (15.6) and the estimate (15.7); we have dropped the indicator in the second expectation. The last estimate holds because φ is standardized.

Last, we need to check the moment inequality for φ_{tail}. This estimate follows by applying the triangle inequality to the definition (15.5):

$$\left(\mathbb{E}\,|\varphi_{\text{tail}}|^p\right)^{1/p} \le \left(\mathbb{E}\,|\varphi|^p\right)^{1/p} + \frac{1-\alpha}{\alpha}\left(\mathbb{E}\,|\varphi|^p\right)^{1/p} = \frac{1}{\alpha}\left(\mathbb{E}\,|\varphi|^p\right)^{1/p} \le 2\nu.$$

We have dropped the indicators after invoking the triangle inequality. Finally, we introduced the estimate (15.7). □

Part IV. Universality of the Embedding Dimension: Proof of Theorem I(b)

This part of the paper develops a condition under which the random projection of a set fails with high probability. This argument establishes the second part of the universality law for the embedding dimension, Theorem I(b).

Section 16 contains the main technical result, a condition under which a bounded random matrix maps a point in a set to the origin. Section 17 extends this argument to the heavy-tailed random matrix model, Model 2.4. In Section 17.2, we use the latter result to derive Theorem I(b). The remaining parts of the section lay out the supporting argument.

16. When Embedding Fails for a Bounded Random Matrix

In this section, we introduce a functional whose value determines whether a linear transformation maps a point in a set to the origin. Then we present the main technical result, which gives an estimate for this functional evaluated on a random linear map from Model 2.1. The rest of the section outlines the main steps in the proof of the result.

16.1. The RAP Functional: Dual Condition for Failure.

To study when a linear map maps a point a set to the origin, we use an approach based on polarity. Let us make the following definition.

Definition 16.1 (Range Avoids Polar (RAP))**.** Let $K \subset \mathbb{R}^m$ be a closed, convex cone. Let $\boldsymbol{A}$ be an $m \times n$ matrix. Define the quantity

$$\tau_{\min}(\boldsymbol{A}; K) := \min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in K^\circ} \|\boldsymbol{s} - \boldsymbol{A}\boldsymbol{t}\|.$$

Note that the range of the inner minimum involves the polar K° of the cone. We refer to $\tau_{\min}$ as the *RAP functional*.

To see why the RAP functional is important, consider a closed, spherically convex subset Ω of S^{m-1} for which $\operatorname{cone}(\Omega)$ is not a subspace. The second conclusion of Proposition 3.8 states that

$$\tau_{\min}(\boldsymbol{A}; \operatorname{cone}(\Omega)) > 0 \quad \text{implies} \quad \mathbf{0} \in \boldsymbol{A}(\Omega).$$

The third conclusion of Proposition 3.8 gives a similar result in the case of a subspace. Therefore, we can obtain a sufficient condition that $\boldsymbol{A}$ maps a point in Ω to zero by providing a lower bound for the RAP functional.

16.2. Theorem 16.2: Main Result for the Bounded Random Matrix Model.

The main technical result in this part of the paper is a theorem on the behavior of the RAP functional of a bounded random matrix.

Theorem 16.2 (RAP: Bounded Random Matrix Model)**.** *Place the following assumptions:*

- *Let m and n be natural numbers with $m + n \le \min\{m, n\}^{9/8}$.*
- *Let K be a closed, convex cone in $\mathbb{R}^m$, and define $\Omega := K \cap \mathsf{S}^{m-1}$.*
- *Draw an $m \times n$ random matrix $\boldsymbol{\Phi}$ from Model 2.1 with bound B.*

Then the squared RAP functional $\tau^2_{\min}(\boldsymbol{\Phi}; K)$ has the following properties.

(1) The squared RAP functional deviates below its mean on a scale of $B^2\sqrt{m+n}$:

$$\mathbb{P}\left\{\tau^2_{\min}(\boldsymbol{\Phi}; K) \le \mathbb{E}\,\tau^2_{\min}(\boldsymbol{\Phi}; K) - \mathrm{C}B^2\zeta\right\} \le \mathrm{e}^{-\zeta^2/m}$$

(2) The expected square of the RAP functional is bounded below:

$$\mathbb{E}\,\tau^2_{\min}(\boldsymbol{\Phi}; K) \ge \frac{\left(\mathscr{E}_n(\Omega)\right)_-^2}{\mathrm{C}B^2 \log(m+n)} - \mathrm{C}B^3(m+n)^{0.95}.$$

This result does *use the symmetry assumption in Model 2.1.*

The proof of Theorem 16.2 is long, even though we can borrow a lot from the proof of Theorem 9.1. This section contains an overview of the calculations that are required with forward references to the detailed arguments.

16.3. **Proof of Theorem 16.2: Lower Tail Bound.** Theorem 16.2(1) states that the quantity $\tau^2_{\min}(\boldsymbol{\Phi}; K)$ does not deviate substantially below its mean. The proof is similar to the proof of the bound (11.1) in Proposition 11.1, which shows that the squared restricted singular value $\sigma^2_{\min}(\boldsymbol{\Phi}; K)$ does not deviate substantially below its mean. We omit the repetitive details.

16.4. **Proof of Theorem 16.2: Truncation and Dissection.** Let us proceed with the proof of Theorem 16.2(2). Define the sets

$$S := K^\circ \cap R\mathsf{B}^m \quad \text{and} \quad T := \mathsf{S}^{n-1}$$

where the radius $R := \mathrm{C}_{\mathrm{rad}} B^2 \sqrt{m+n}$ for some universal constant $\mathrm{C}_{\mathrm{rad}}$. Next, we construct a family of closed, convex subsets of S and T. For each $I \subset \{1, \dots, m\}$ and each $J \subset \{1, \dots, n\}$, define

$$S_I := \{\boldsymbol{s} \in S : |s_i| \le R(\#I)^{-1/2} \text{ for all } i \in I^c\};$$
$$T_J := \{\boldsymbol{t} \in T : |t_j| \le (\#J)^{-1/2} \text{ for all } j \in J^c\}.$$

Fix the cardinality parameter $k \in \{1, \dots, \min\{m, n\}\}$. As in the proof of Theorem 9.1, we have the decompositions

$$S = \bigcup\nolimits_{\#I=k} S_I \quad \text{and} \quad T = \bigcup\nolimits_{\#J=k} T_J.$$

Furthermore,

$$\#\{S_I : \#I = k\} \times \#\{T_J : \#J = k\} \le \left(\frac{\mathrm{e}(m+n)}{k}\right)^k. \tag{16.1}$$

We maintain the heuristic that the cardinality k is much smaller than either ambient dimension m or n.

16.5. **Proof of Theorem 16.2: Lower Bound for RAP Functional.** To bound the quantity $\tau_{\min}(\boldsymbol{\Phi}; K)$, we must combine estimates from several technical results. We give an outline of the calculation here, with the details postponed to a series of propositions.

First, we must account for the error we incur when we truncate the cone K° to the wedge S. Proposition 18.1 demonstrates that

$$\mathbb{E}\,\tau^2_{\min}(\boldsymbol{\Phi}; K) = \mathbb{E} \min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in K^\circ} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 \ge \mathbb{E} \min_{\boldsymbol{t} \in T} \min_{\boldsymbol{s} \in S} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 - CB^4. \tag{16.2}$$

This inequality is based on an estimate for the norm of the point $\boldsymbol{s} \in K^\circ$ where the inner minimum is achieved, as well as a probability bound for the norm of the random matrix $\boldsymbol{\Phi}$.

Next, we pass from the minimum over the full sets S and T to minima over their subsets S_I and T_J:

$$\mathbb{E} \min_{\boldsymbol{t} \in T} \min_{\boldsymbol{s} \in S} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 \ge \min_{\#I=\#J=k} \mathbb{E} \min_{\boldsymbol{t} \in T_J} \min_{\boldsymbol{s} \in S_I} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 - CB^2 \sqrt{km \log((m+n)/k)}. \tag{16.3}$$

The proof of this inequality hews to the argument in Proposition 12.2. We just need to invoke the concentration inequality from Theorem 16.2(1) in lieu of the concentration inequality from Proposition 11.1, and we take into account the bound (16.1) on the number of subsets in the decomposition. Further details are omitted.

We are now prepared to perform the exchange argument to pass from the matrix $\boldsymbol{\Phi}$ to a matrix $\boldsymbol{\Psi}$ that contains many standard normal entries. Fix subsets $I \subset \{1, \dots, m\}$ and $J \subset \{1, \dots, n\}$, each with cardinality k. Introduce the random matrix

$$\boldsymbol{\Psi} := \boldsymbol{\Psi}(I, J) := \begin{bmatrix} \boldsymbol{\Phi}_{IJ} & \boldsymbol{\Phi}_{IJ^c} \\ \boldsymbol{\Phi}_{I^cJ} & \boldsymbol{\Gamma}_{I^cJ^c} \end{bmatrix}$$

where $\boldsymbol{\Gamma}$ is an $m \times n$ standard normal matrix. Proposition 19.1 gives the bound

$$\begin{aligned} \mathbb{E} \min_{\boldsymbol{s} \in S_I} \min_{\boldsymbol{t} \in T_J} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 &\ge \mathbb{E} \min_{\boldsymbol{t} \in T_J} \min_{\boldsymbol{s} \in S_I} \|\boldsymbol{s} - \boldsymbol{\Psi}(I, J)\boldsymbol{t}\|^2 - \frac{CB^3(m+n)^{11/6}\log(mn)}{k} \\ &\ge \mathbb{E} \min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in K^\circ} \|\boldsymbol{s} - \boldsymbol{\Psi}(I, J)\boldsymbol{t}\|^2 - \frac{CB^3(m+n)^{11/6}\log(mn)}{k} \end{aligned} \tag{16.4}$$

The proof is similar with the proof of Proposition 13.1. We discretize both sets; we smooth the minima using the soft-min function; and then we apply the Lindeberg principle. The main distinction is that we

can replace even less of the matrix $\mathbf{\Phi}$ than before. The second line in (16.4) is an immediate consequence of the facts $S_I \subset S \subset K^\circ$ and $T_J \subset T = \mathsf{S}^{n-1}$.

To continue, we must identify a geometric functional that is hiding within the expression (16.4). Write $\Omega := K \cap \mathsf{S}^{m-1}$. Proposition 20.1 demonstrates that

$$\mathbb{E} \min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in K^\circ} \|\boldsymbol{s} - \mathbf{\Psi}(I,J)\boldsymbol{t}\|^2 \geq \left(\frac{\big(\mathscr{E}_n(\Omega)\big)_-}{CB\sqrt{\log m}} - CB^2\sqrt{k \log m} \right)_+^2. \tag{16.5}$$

As in Proposition 14.1, the main tool is the Gaussian Minimax Theorem, Fact A.3, which allows us to break down the standard normal matrix $\mathbf{\Gamma}$ into simpler quantities. The proof requires some convex duality arguments, as well as some delicate considerations that did not arise before.

Last, we linearize the function $(\cdot)_+^2$ in (16.5):

$$\left(\frac{\big(\mathscr{E}_n(\Omega)\big)_-}{CB\sqrt{\log m}} - CB^2\sqrt{k \log m} \right)_+^2 \geq \frac{\big(\mathscr{E}_n(\Omega)\big)_-^2}{CB^2 \log m} - CB\sqrt{km}. \tag{16.6}$$

We have employed the observation that

$$\big(\mathscr{E}_n(\Omega)\big)_- \leq \mathbb{E} \max_{\boldsymbol{s} \in \Omega} \boldsymbol{g} \cdot \boldsymbol{s} \leq \mathbb{E}\|\boldsymbol{g}\| \leq \sqrt{m}.$$

Here, $\boldsymbol{g} \in \mathbb{R}^m$ is a standard normal vector.

Now, sequence the estimates (16.2), (16.3), (16.4), (16.5), and (16.6) to arrive at

$$\mathbb{E}\,\tau_{\min}^2(\mathbf{\Phi}; K) \geq \frac{\big(\mathscr{E}_n(\Omega)\big)_-^2}{CB^2 \log m} - CB^3 \left(\frac{(m+n)^{11/6} \log(mn)}{k} + \sqrt{km \log((m+n)/k)} \right).$$

We select $k = \lfloor (m+n)^{8/9} \rfloor$, which results in the bound

$$\mathbb{E}\,\tau_{\min}^2(\mathbf{\Phi}; K) \geq \frac{\big(\mathscr{E}_n(\Omega)\big)_-^2}{CB^2 \log(m+n)} - CB^3 (m+n)^{17/18} \log(m+n).$$

Note that $k \leq \min\{m,n\}$, as required, because we have assumed that $m+n \leq \min\{m,n\}^{9/8}$. We obtain the result quoted in Theorem 16.2(2).

17. When Embedding Fails for a Heavy-Tailed Random Matrix

In this section, we extend Theorem 16.2 to the heavy-tailed matrix model, Model 2.4. In Section 10.3, we show how to derive the second half of the universality result for the embedding dimension, Theorem I(a), as a consequence.

17.1. Corollary 17.1: Main Result for the p-Moment Random Matrix Model.

The following corollary extends Theorem 16.2 to include random matrices drawn from Model 2.4.

Corollary 17.1 (RAP: p-Moment Random Matrix Model)**.** *Fix parameters $p > 4$ and $\nu \geq 1$. Place the following assumptions:*

- *Let m and n be natural numbers with $m+n \leq \min\{m,n\}^{9/8}$.*
- *Let K be a closed, convex cone in $\mathbb{R}^m$, and define $\Omega := K \cap \mathsf{S}^{m-1}$.*
- *Draw an $m \times n$ random matrix $\mathbf{\Phi}$ that satisfies Model 2.4 with given p and ν.*

Then the RAP functional satisfies the probability bound

$$\mathbb{P}\left\{ \tau_{\min}(\mathbf{\Phi}; K) \leq \frac{\big(\mathscr{E}_n(\Omega)\big)_-}{\mathrm{C}_p \nu (m+n)^{(1-\mathrm{c}_p)\kappa(p)}} - \mathrm{C}_p \nu^{3/2} (m+n)^{1/2-\kappa(p)} \right\} \leq \mathrm{C}_p (m+n)^{1-p/4}.$$

The function $\kappa(p)$ is strictly positive for $p > 4$. The strictly positive constants c_p and C_p depend only on p.

The proof of Corollary 17.1 follows from Theorem 16.2 and the same kind of truncation argument that appears in Section 15. We omit further details.

17.2. **Proof of Theorem I(b) from Corollary 17.1.** Theorem I(b) is an easy consequence of Corollary 17.1. Let us restate the assumptions of the theorem:

- E is a compact subset of $\mathbb{R}^D$ that does not contain the origin.
- The statistical dimension of E satisfies $\delta(E) \geq \varrho D$.
- The $d \times D$ random linear map $\mathbf{\Pi}$ follows Model 2.4 with parameters $p > 4$ and $\nu \geq 1$.

We must now consider the regime where the embedding dimension $d \leq (1-\varepsilon)\,\delta(E)$. We need to demonstrate that

$$\mathbb{P}\{\mathbf{0} \in \mathbf{\Pi}(E)\} = \mathbb{P}\{E \cap \operatorname{null}(\mathbf{\Pi}) \neq \emptyset\} \geq 1 - \mathrm{C}_p D^{1-p/4}. \tag{17.1}$$

As in Section 10.3, the probability is a decreasing function of the embedding dimension, so we may as well consider the case where $d = \lfloor (1-\varepsilon)\,\delta(E) \rfloor$. It is easy to see that $d \leq D \leq \varrho^{-1} d$ because $\delta(E) \leq d$. In particular, $d + D \leq \min\{d, D\}^{9/8}$.

Introduce the spherical retraction $\Omega := \boldsymbol{\theta}(E)$. If $\operatorname{cone}(\Omega)$ is a subspace, we replace Ω by a subset Ω_0 with the property that $\operatorname{cone}(\Omega_0)$ is a subspace of one dimension fewer than $\operatorname{cone}(\Omega)$. Then we proceed with the argument.

Proposition 3.8 and relation (3.2) show that

$$\tau_{\min}(\mathbf{\Pi}^*; \operatorname{cone}(\Omega)) > 0 \quad \text{implies} \quad \mathbf{0} \in \mathbf{\Pi}(\Omega) \quad \text{implies} \quad \mathbf{0} \in \mathbf{\Pi}(E).$$

Therefore, to verify (17.1), it is enough to produce a high-probability lower bound on the RAP functional $\tau_{\min}(\mathbf{\Pi}^*; \operatorname{cone}(\Omega))$. With the choices $\mathbf{\Phi} = \mathbf{\Pi}^*$ and $K = \operatorname{cone}(\Omega)$, Corollary 17.1 yields

$$\mathbb{P}\left\{ \tau_{\min}(\mathbf{\Pi}^*; \operatorname{cone}(\Omega)) \geq \frac{\big(\mathscr{E}_d(\Omega)\big)_-}{\mathrm{C}_p \nu (d+D)^{(1-\mathrm{c}_p)\kappa(p)}} - \mathrm{C}_p \nu^{3/2} (d+D)^{1/2-\kappa(p)} \right\} \geq 1 - \mathrm{C}_p D^{1-p/4}.$$

We need to check that the lower bound on the RAP functional is positive. That is,

$$\big(\mathscr{E}_d(\Omega)\big)_- > \mathrm{C}_p \nu^{5/2} (d+D)^{1/2-\mathrm{c}_p\kappa(p)}. \tag{17.2}$$

Once again, this point follows from two relatively short calculations.

Since Ω is a subset of the unit sphere, we quickly compute the excess width using (4.3):

$$\big(\mathscr{E}_d(\Omega)\big)_- \geq \sqrt{\delta(\Omega) - 1} - \sqrt{d} = \sqrt{\delta(E) - 1} - \sqrt{d} \geq \big(1 - \sqrt{1-\varepsilon}\big)\sqrt{\delta(E) - 1}. \tag{17.3}$$

The justifications are the same as in Section 10.3.

We can easily bound the dimensional term in (17.2):

$$(d+D)^{1/2-\mathrm{c}_p\kappa(p)} \leq \mathrm{C} D^{1/2-\mathrm{c}_p\kappa(p)} \leq \mathrm{C}\varrho^{-1/2} D^{-\mathrm{c}_p\kappa(p)} \sqrt{\delta(E)} \leq \mathrm{C}\varrho^{-1/2} D^{-\mathrm{c}_p\kappa(p)} \sqrt{\delta(E) - 1}$$

The first inequality holds because $d \leq D$, and the last two relations both rely on the assumption $1 \ll D \leq \varrho^{-1}\delta(E)$. Since $\mathrm{c}_p\kappa(p)$ is positive, we can find a number $N := N(p, \nu, \varrho, \varepsilon)$ for which $D \geq N$ implies that

$$\mathrm{C}_p \nu^{5/2} (d+D)^{1/2-\mathrm{c}_p\kappa(p)} < \big(\sqrt{1+\varepsilon} - 1\big)\sqrt{\delta(E) - 1}.$$

Combine the last result with (17.3) to see that the claim (17.2) holds true.

17.3. **Proof of Proposition 6.1: Application of RAP functional to decoding with structured errors.** This section establishes the success condition in Proposition 6.1. More precisely, we show that

$$s/n < \psi_{\ell_1}^{-1}(1 - m/n) - o(1) \quad \text{implies} \quad \Omega \cap \operatorname{range}(\mathbf{\Phi}) = \emptyset \quad \text{with probability } 1 - o(1). \tag{17.4}$$

The set Ω is derived from the descent cone of the ℓ_1 norm, and the calculation (6.3) shows that

$$\delta(\Omega^\circ) = n(1 - \psi_{\ell_1}(s/n) - o(1)).$$

To begin, we assume that there are parameters $\varrho, \varepsilon > 0$ for which

- The statistical dimension $\delta(\Omega^\circ) \geq \varrho n$.
- The message length $m \leq (1-\varepsilon)\,\delta(\Omega^\circ)$.

As in Section 17.2, these conditions imply that

$$\Omega^\circ \cap \operatorname{null}(\boldsymbol{\Phi}^*) \neq \emptyset \quad \text{with probability at least } 1 - o(1).$$

Indeed, this statement corresponds to (17.1) under the change of variables $d = m$ and $D = n$ and $E = \Omega^\circ$ and $\boldsymbol{\Pi} = \boldsymbol{\Phi}^*$. By polarity, the conclusion of (17.4) holds.

It remains to show that the parameter choice in (17.4) implies the two assumptions we have made. As in the proof of Proposition 6.1, the condition $s \le (1-\xi)n$ ensures that $\delta(\Omega^\circ) \ge \varrho n$ for some $\varrho > 0$. Similarly, the condition $m \le (1-\varepsilon)\delta(\Omega^\circ)$ holds when

$$m \le (1-\varepsilon) \cdot n(1 - \psi_{\ell_1}(s/n) - o(1)).$$

Equivalently,

$$s/n \le \psi_{\ell_1}^{-1}(1 - m/((1-\varepsilon)n) - o(1)).$$

Since we can choose ε to be an arbitrarily small constant as $m \to \infty$, it suffices that

$$s/n \le \psi_{\ell_1}^{-1}(1 - m/n - o(1)).$$

This observation completes the argument.

18. Theorem 16.2: Truncation of the Cone

In this section, we argue that functional $\tau_{\min}(\boldsymbol{\Phi}; K)$ does not change very much if we truncate the cone K. Replacing the unbounded set with a compact set allows us to develop discretization arguments.

Proposition 18.1 (RAP: Truncation of Cone)**.** *Adopt the notation and hypotheses of Theorem 16.2. Let* $S := K^\circ \cap R\mathsf{B}^m$*, where* $R := \mathrm{C}_{\mathrm{rad}} B^2 \sqrt{m+n}$*. Then*

$$\mathbb{E} \min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in K^\circ} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 \ge \mathbb{E} \min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in S} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 - \mathrm{C}B^4.$$

Proof. Since $\mathbf{0} \in K^\circ$, it is easy to see that

$$\min_{\boldsymbol{s} \in K^\circ} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\| \le \|\boldsymbol{\Phi}\boldsymbol{t}\| \le \|\boldsymbol{\Phi}\| \quad \text{when } \|\boldsymbol{t}\| = 1.$$

Meanwhile, the triangle inequality gives the bound

$$\|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\| \ge \|\boldsymbol{s}\| - \|\boldsymbol{\Phi}\boldsymbol{t}\| \ge \|\boldsymbol{s}\| - \|\boldsymbol{\Phi}\| \quad \text{when } \|\boldsymbol{t}\| = 1.$$

It follows that

$$\boldsymbol{s}_\star \in \arg\min_{\boldsymbol{s} \in K^\circ} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\| \quad \text{implies} \quad \|\boldsymbol{s}_\star\| \le 2\|\boldsymbol{\Phi}\| \quad \text{when } \|\boldsymbol{t}\| = 1. \tag{18.1}$$

Since the norm of the random matrix $\boldsymbol{\Phi}$ concentrates, the bound (18.1) shows that the norm of the minimizer $\boldsymbol{s}_\star$ is unlikely to be large.

For any positive parameter R, the observation (18.1) allows us to calculate that

$$\begin{aligned}
\mathbb{E}\,\tau_{\min}^2(\boldsymbol{\Phi}; K) &= \mathbb{E} \min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in K^\circ} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 \\
&\ge \mathbb{E}\left[\min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in K^\circ} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 \mathbb{1}\{\|\boldsymbol{\Phi}\| \le R/2\} \right] \\
&= \mathbb{E}\left[\min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in K^\circ \cap R\mathsf{B}^m} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 \mathbb{1}\{\|\boldsymbol{\Phi}\| \le R/2\} \right] \\
&= \mathbb{E} \min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in K^\circ \cap R\mathsf{B}^m} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 - \mathbb{E}\left[\max_{\|\boldsymbol{t}\|=1} \max_{\boldsymbol{s} \in K^\circ \cap R\mathsf{B}^m} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 \mathbb{1}\{\|\boldsymbol{\Phi}\| > R/2\} \right].
\end{aligned} \tag{18.2}$$

To reach the last line, we write the indicator function in terms of its the complement.

To bound the second term on the right-hand side of (18.2), crude estimates suffice.

$$\begin{aligned}\mathbb{E}\left[\max_{\|\boldsymbol{t}\|=1}\max_{\boldsymbol{s}\in K^\circ\cap R\mathsf{B}^m}\|\boldsymbol{s}-\boldsymbol{\Phi}\boldsymbol{t}\|^2\,\mathbb{1}\{\|\boldsymbol{\Phi}\|>R/2\}\right] &\le \mathbb{E}\left[(R+\|\boldsymbol{\Phi}\|)^2\mathbb{1}\{\|\boldsymbol{\Phi}\|>R/2\}\right]\\ &\le 9\,\mathbb{E}\left[\|\boldsymbol{\Phi}\|^2\mathbb{1}\{\|\boldsymbol{\Phi}\|>R/2\}\right]\\ &\le 9\left(\mathbb{E}\,\|\boldsymbol{\Phi}\|^4\right)^{1/2}\left(\mathbb{P}\{\|\boldsymbol{\Phi}\|>R/2\}\right)^{1/2}.\end{aligned}$$

The last inequality is Cauchy–Schwarz. Since $\boldsymbol{\Phi}$ satisfies the condition (B.1), Fact B.1 implies that

$$\mathbb{P}\left\{\|\boldsymbol{\Phi}\|>\mathrm{C}_0B^2\sqrt{m+n}+\mathrm{C}_0B^2\zeta\right\}\le \mathrm{e}^{-\zeta^2}.$$

In particular, using integration by parts,

$$\left(\mathbb{E}\,\|\boldsymbol{\Phi}\|^4\right)^{1/4}\le \mathrm{C}B^2\sqrt{m+n}.$$

Furthermore, there is a constant $\mathrm{C}_{\mathrm{rad}}$ for which

$$\mathbb{P}\left\{\|\boldsymbol{\Phi}\|>\tfrac{1}{2}\mathrm{C}_{\mathrm{rad}}B^2\sqrt{m+n}\right\}\le (m+n)^{-2}.$$

If we set $R:=\mathrm{C}_{\mathrm{rad}}B^2\sqrt{m+n}$, then

$$\mathbb{E}\left[\max_{\|\boldsymbol{t}\|=1}\max_{\boldsymbol{s}\in K^\circ\cap R\mathsf{B}^m}\|\boldsymbol{s}-\boldsymbol{\Phi}\boldsymbol{t}\|^2\,\mathbb{1}\{\|\boldsymbol{\Phi}\|>R/2\}\right]\le \mathrm{C}B^4. \tag{18.3}$$

Introduce the estimate (18.3) into (18.2) to complete the argument. □

19. Theorem 16.2: Replacing Most Entries of the Random Matrix

In this section, we show that we can replace most of the entries of a random matrix $\boldsymbol{\Phi}$ with standard normal variables without changing the value of the functional $\mathbb{E}\,\tau_{\min}^2(\boldsymbol{\Phi};K)$ substantially

Proposition 19.1 (RAP: Partial Replacement)**.** *Let $\boldsymbol{\Phi}$ be an $m\times n$ random matrix that satisfies Model 2.1 with magnitude bound B. Fix the parameter $R:=\mathrm{C}_{\mathrm{rad}}B^2\sqrt{m+n}$. Let I be a subset of $\{1,\dots,m\}$ with cardinality k, and let S_I be a closed subset of $R\mathsf{B}^m$ for which*

$$\boldsymbol{s}\in S_I \quad \textit{implies} \quad |s_i|\le Rk^{-1/2} \quad \textit{for each index } i\in I^c.$$

Let J be a subset of $\{1,\dots,n\}$ with cardinality k, and let T_J be a closed subset of B^n for which

$$\boldsymbol{t}\in T_J \quad \textit{implies} \quad |t_j|\le k^{-1/2} \quad \textit{for each index } j\in J^c.$$

Suppose that $\boldsymbol{\Psi}$ is an $m\times n$ matrix with block form

$$\boldsymbol{\Psi}:=\boldsymbol{\Psi}(I,J):=\begin{bmatrix}\boldsymbol{\Phi}_{IJ} & \boldsymbol{\Phi}_{IJ^c}\\ \boldsymbol{\Phi}_{I^cJ} & \boldsymbol{\Gamma}_{I^cJ^c}\end{bmatrix}. \tag{19.1}$$

Then

$$\left|\mathbb{E}\min_{\boldsymbol{t}\in T_J}\min_{\boldsymbol{s}\in S_I}\|\boldsymbol{s}-\boldsymbol{\Phi}\boldsymbol{t}\|^2-\mathbb{E}\min_{\boldsymbol{t}\in T_J}\min_{\boldsymbol{s}\in S_I}\|\boldsymbol{s}-\boldsymbol{\Psi}\boldsymbol{t}\|^2\right|\le\frac{\mathrm{C}B^3(m+n)^{11/6}\log(mn)}{k}.$$

As usual, $\boldsymbol{\Gamma}$ is an $m\times n$ standard normal matrix.

19.1. **Proof of Proposition 19.1.** Fix the sets I and J. As in the proof of Proposition 13.1, the error has three components:

$$\begin{aligned}\left|\mathbb{E}\min_{\boldsymbol{t}\in T_J}\min_{\boldsymbol{s}\in S_I}\|\boldsymbol{s}-\boldsymbol{\Phi}\boldsymbol{t}\|^2-\mathbb{E}\min_{\boldsymbol{t}\in T_J}\min_{\boldsymbol{s}\in S_I}\|\boldsymbol{s}-\boldsymbol{\Psi}\boldsymbol{t}\|^2\right| &\le \mathrm{C}mn\varepsilon && \text{(Lemma 19.2)}\\ &+ \mathrm{C}\beta^{-1}(m+n)\log(1/\varepsilon) && \text{(Lemma 19.3)}\\ &+ \mathrm{C}B^3mn\left(\frac{\beta R}{k^2}+\frac{\beta^2R^3}{k^3}\right). && \text{(Lemma 19.4)}\end{aligned}$$

The first error comes from discretizing the sets S and T at a level $\varepsilon \in (0,1]$. The second error appears when we replace the minima with a soft-min function with parameter $\beta > 0$. The last error emerges from the Lindeberg exchange argument.

To complete the proof, we set $\varepsilon = (mn)^{-1}$ to make the discretization error negligible. Select the smoothing parameter so that $\beta^3 = k^3(m+n)/(B^3R^3mn)$. We arrive at

$$\left| \mathbb{E} \min_{\boldsymbol{t}\in T_J} \min_{\boldsymbol{s}\in S_I} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 \;-\; \mathbb{E} \min_{\boldsymbol{t}\in T_J} \min_{\boldsymbol{s}\in S_I} \|\boldsymbol{s} - \boldsymbol{\Psi}\boldsymbol{t}\|^2 \right| \leq \frac{\mathrm{C}B^2(m+n)^{1/3}(mn)^{2/3}}{k^2} + \frac{\mathrm{C}BR(m+n)^{2/3}(mn)^{1/3}\log(mn)}{k}.$$

Since $2\sqrt{mn} \leq m+n$ and $R = \mathrm{C}_{\mathrm{rad}}B^2\sqrt{m+n}$, the second term dominates. We reach the stated result.

19.2. **Proposition 19.1: Discretizing the Index Sets.** The first step in the proof of Proposition 19.1 is to replace the index sets by finite subsets.

Lemma 19.2 (Proposition 19.1: Discretization)**.** *Adopt the notation and hypotheses of Proposition 19.1. Fix a parameter $\varepsilon \in (0,1]$. Then S_I contains a finite subset S_I^ε and T_J contains a finite subset T_J^ε whose cardinalities satisfy*

$$\log(\#S_I^\varepsilon) + \log(\#T_J^\varepsilon) \leq (m+n)\log(3/\varepsilon).$$

Furthermore, these subsets have the property that

$$\left| \mathbb{E} \min_{\boldsymbol{t}\in T_J} \min_{\boldsymbol{s}\in S_I} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 - \mathbb{E} \min_{\boldsymbol{t}\in T_J^\varepsilon} \min_{\boldsymbol{s}\in S_I^\varepsilon} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 \right| \leq \mathrm{C}mn\varepsilon. \tag{19.2}$$

The bound (19.2) *also holds if we replace $\boldsymbol{\Phi}$ by $\boldsymbol{\Psi}$.*

Proof. We choose S_I^ε to be an $(R\varepsilon)$-covering of S_I, and T_J^ε to be an ε-covering of T_J. Since S_I is a subset of $R\mathrm{B}^m$ and T_J is a subset of B^n, we can be sure that the coverings have cardinality $\#S_I^\varepsilon \leq (3/\varepsilon)^m$ and $\#T_J^\varepsilon \leq (3/\varepsilon)^n$. See [Ver12, Lem. 5.2]. The rest of the proof is essentially the same as that of Lemma 13.2, so we omit the details. □

19.3. **Proposition 19.1: Smoothing the Minimum.** The next step in the proof of Proposition 19.1 is to pass from the minimum to the soft-min function.

Lemma 19.3 (Proposition 19.1: Smoothing)**.** *Adopt the notation and hypotheses of Proposition 19.1, and let S_I^ε and T_J^ε be the sets introduced in Lemma 19.2. Fix a parameter $\beta > 0$, and introduce the function*

$$F : \mathbb{R}^{m\times n} \to \mathbb{R} \quad \text{where} \quad F(\boldsymbol{A}) := -\frac{1}{\beta}\log \sum\nolimits_{\boldsymbol{s}\in S_I^\varepsilon} \sum\nolimits_{\boldsymbol{t}\in T_J^\varepsilon} \mathrm{e}^{-\beta\|\boldsymbol{s}-\boldsymbol{A}\boldsymbol{t}\|^2}.$$

Then

$$\left| \mathbb{E} \min_{\boldsymbol{t}\in T_J^\varepsilon} \min_{\boldsymbol{s}\in S_I^\varepsilon} \|\boldsymbol{s} - \boldsymbol{\Phi}\boldsymbol{t}\|^2 - \mathbb{E}\, F(\boldsymbol{\Phi}) \right| \leq \frac{1}{\beta}\left(\log(\#S_I^\varepsilon) + \log(\#T_J^\varepsilon)\right). \tag{19.3}$$

The estimate (19.3) *also holds if we replace $\boldsymbol{\Phi}$ by $\boldsymbol{\Psi}$.*

Proof. The proof is almost identical with that of Lemma 13.3. □

19.4. **Proposition 19.1: Exchanging the Entries of the Random Matrix.** The main challenge in the proof of Proposition 19.1 is to exchange most of the entries of the random matrix $\boldsymbol{\Phi}$ for the entries of $\boldsymbol{\Psi}$.

Lemma 19.4 (Proposition 19.1: Exchange)**.** *Adopt the notation and hypotheses of Proposition 19.1, and let F be the function defined in Lemma 19.3. Then*

$$|\mathbb{E}\, F(\boldsymbol{\Phi}) - \mathbb{E}\, F(\boldsymbol{\Psi})| \leq \mathrm{C}B^3mn\left(\frac{\beta R}{k^2} + \frac{\beta^2R^3}{k^3}\right).$$

The proof is similar with Lemma 13.4. This time, we replace only the rows of $\mathbf{\Phi}$ listed in I^c. We incur the same error for each of these $m-k$ rows, so it suffices to control the error in exchanging a single row. The following sublemma achieves this goal.

Sublemma 19.5 (Lemma 19.4: Comparison for One Row)**.** *Adopt the notation and hypotheses of Proposition 19.1, and let S_I^ϵ and T_J^ϵ be the sets defined in Lemma 19.3. For $i \in I^c$, introduce the function*

$$f : \mathbb{R}^n \to \mathbb{R} \quad \text{given by} \quad f(\boldsymbol{a}) := -\frac{1}{\beta} \log \sum\nolimits_{\boldsymbol{s} \in S_I^\epsilon} \sum\nolimits_{\boldsymbol{t} \in T_J^\epsilon} \mathrm{e}^{-\beta(s_i - \boldsymbol{a}\cdot\boldsymbol{t})^2 + q(\boldsymbol{s},\boldsymbol{t})},$$

where $q : S_I^\epsilon \times T_J^\epsilon \to \mathbb{R}$ is an arbitrary function. Suppose that $\boldsymbol{\varphi} \in \mathbb{R}^n$ is a random vector with independent, standardized entries that are bounded in magnitude by B. Suppose that $\boldsymbol{\psi} \in \mathbb{R}^n$ is a random vector with

$$\boldsymbol{\psi}_J = \boldsymbol{\varphi}_J \quad \text{and} \quad \boldsymbol{\psi}_{J^c} = \boldsymbol{\gamma}_{J^c},$$

where $\boldsymbol{\gamma} \in \mathbb{R}^n$ is a standard normal vector. Then

$$\left| \mathbb{E} f(\boldsymbol{\varphi}) - \mathbb{E} f(\boldsymbol{\psi}) \right| \le \mathrm{C} B^3 n \left(\frac{\beta R}{k^2} + \frac{\beta^2 R^3}{k^3} \right).$$

The proof of this result is much the same as the proof of Sublemma 13.5. There are only two points that require care. First, we use a slightly different result to compute the derivatives.

Sublemma 19.6 (Lemma 19.4: Derivatives)**.** *Adopt the notation and hypotheses of Proposition 19.1 and Sublemma 19.5. Let $\boldsymbol{\xi} : \mathbb{R} \to \mathbb{R}^n$ be a linear function, so its derivative $\boldsymbol{\xi}' \in \mathbb{R}^n$ is a constant vector. For $i \in I^c$, define the function*

$$r(\alpha) := f\big(\boldsymbol{\xi}(\alpha)\big) = -\frac{1}{\beta} \log \sum\nolimits_{\boldsymbol{s} \in S_I^\epsilon} \sum\nolimits_{\boldsymbol{t} \in T_J^\epsilon} \mathrm{e}^{-\beta(s_i - \boldsymbol{\xi}(\alpha)\cdot\boldsymbol{t})^2 + q(\boldsymbol{s},\boldsymbol{t})}$$

where $q : S_I^\epsilon \times T_J^\epsilon \to \mathbb{R}$ is arbitrary. The third derivative of this function satisfies

$$\left| r'''(\alpha) \right| \le 48 \left(\max_{\boldsymbol{t} \in T_J^\epsilon} \left| \boldsymbol{\xi}' \cdot \boldsymbol{t} \right|^3 \right) \left(\beta \mathbb{E}_{\boldsymbol{v}} \left| s_i - \boldsymbol{\xi}(\alpha) \cdot \boldsymbol{v} \right| + \beta^2 \mathbb{E}_{\boldsymbol{v}} \left| s_i - \boldsymbol{\xi}(\alpha) \cdot \boldsymbol{v} \right|^3 \right),$$

where $\boldsymbol{v} \in T_J^\epsilon$ is a random vector that does not depend on $\boldsymbol{\xi}(\alpha)$.

Second, when making further bounds on $\left| r'''(\alpha) \right|$, we need to exploit our control on the magnitude of $\boldsymbol{s}$ on the coordinates in I^c. Note that

$$|s_i - \boldsymbol{\xi}(\alpha) \cdot \boldsymbol{v}| \le |s_i| + |\boldsymbol{\xi}(\alpha) \cdot \boldsymbol{v}| \le R k^{-1/2} + |\boldsymbol{\xi}(\alpha) \cdot \boldsymbol{v}|$$

The second inequality holds because $|s_i| \le R k^{-1/2}$ for each $i \in I^c$. Similarly,

$$|s_i - \boldsymbol{\xi}(\alpha) \cdot \boldsymbol{v}|^3 \le \mathrm{C} |s_i|^3 + \mathrm{C} |\boldsymbol{\xi}(\alpha) \cdot \boldsymbol{v}|^3 \le \mathrm{C} R k^{-3/2} + \mathrm{C} |\boldsymbol{\xi}(\alpha) \cdot \boldsymbol{v}|^3 .$$

Repeating the arguments from Sublemma 13.5, we obtain bounds of the form

$$\mathbb{E} \left[|\varphi_j|^3 \max_{|\alpha| \le |\varphi_j|} \left| r_j'''(\alpha) \right| \right] \le \frac{\beta B^3 R}{k^2} + \frac{\mathrm{C} \beta^2 B^3 R^3}{k^3} + \frac{\mathrm{C} \beta B^4 + \mathrm{C} \beta^2 B^6}{k^{3/2}}.$$

The first two terms dominate the third because our choice $R = \mathrm{C}_{\mathrm{rad}} B^2 \sqrt{m+n}$ implies that $R \ge B k^{1/2}$. We arrive at the statement of Sublemma 19.5.

20. Theorem 16.2: Bounding the RAP Functional by the Excess Width

The most difficult part of proving Theorem 16.2 is to identify the excess width $\mathscr{E}_n(\Omega)$ after we replace the original matrix $\mathbf{\Phi}$ by the hybrid matrix $\mathbf{\Psi}$ defined in (19.1). The following result does the job.

Proposition 20.1 (Theorem 16.2: Excess Width Bound)**.** *Adopt the notation and hypotheses of Proposition 19.1. Let* $\Omega := K \cap \mathsf{S}^{m-1}$. *Then*

$$\mathbb{E} \min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in K^\circ} \|\boldsymbol{s} - \boldsymbol{\Psi}(I,J)\boldsymbol{t}\|^2 \geq \left(\frac{\big(\mathscr{E}_n(\Omega)\big)_-}{CB\sqrt{\log m}} - CB^2 \sqrt{k \log m} \right)_+^2.$$

The random matrix $\boldsymbol{\Psi}$ *is defined in* (19.1).

The proof of Proposition 20.1 occupies the rest of this section. At the highest level, the proof is similar with the argument underlying Proposition 14.1. We write the quantity of interest as a minimax, and then we apply the Gaussian Minimax Theorem to replace the Gaussian matrix with a pair of Gaussian vectors. Afterward, we analyze the resulting expression to identify the Gaussian width; the new challenges appear in this step.

20.1. **Proof of Proposition 20.1.** Here is an overview of the calculations that we will perform; the detailed justifications appear in the upcoming subsections. Let us abbreviate $U := K \cap \mathsf{B}^m$. We have the chain of inequalities

$$\begin{aligned}
&\mathbb{E} \min_{\|\boldsymbol{t}\|=1} \min_{\boldsymbol{s} \in K^\circ} \|\boldsymbol{s} - \boldsymbol{\Psi}\boldsymbol{t}\| \\
&\qquad \geq \mathbb{E} \min_{\|\boldsymbol{t}\|=1} \max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \boldsymbol{\Psi}\boldsymbol{t} && \text{(Lemma 20.2)} \\
&\qquad = \mathbb{E} \min_{\|\boldsymbol{t}\|=1} \max_{\boldsymbol{u} \in U} \begin{bmatrix} \boldsymbol{u}_I \\ \boldsymbol{u}_{I^c} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{\Phi}_{IJ} & \boldsymbol{\Phi}_{IJ^c} \\ \boldsymbol{\Phi}_{I^cJ} & \boldsymbol{\Gamma}_{I^cJ^c} \end{bmatrix} \begin{bmatrix} \boldsymbol{t}_J \\ \boldsymbol{t}_{J^c} \end{bmatrix} \\
&\qquad \geq \mathbb{E} \min_{\|\boldsymbol{t}\|=1} \max_{\boldsymbol{u} \in U} \left(\boldsymbol{u} \cdot \begin{bmatrix} \boldsymbol{\Phi}_{IJ} & \boldsymbol{\Phi}_{IJ^c} \\ \boldsymbol{\Phi}_{I^cJ} & \boldsymbol{0} \end{bmatrix} \boldsymbol{t} + (\boldsymbol{u}_{I^c} \cdot \boldsymbol{g}_{I^c}) \|\boldsymbol{t}_{J^c}\| + \|\boldsymbol{u}_{I^c}\| (\boldsymbol{h}_{J^c} \cdot \boldsymbol{t}_{J^c}) \right) - 2 && \text{(Lemma 20.3)} \\
&\qquad \geq \frac{\big(\mathscr{W}(U) - \sqrt{n}\big)_+}{CB\sqrt{\log m}} - CB^2 \sqrt{k \log m} && \text{(Lemma 20.4).}
\end{aligned}$$

Lemma 20.2 is a standard convex duality argument, and the next line follows when we write out the quantity of interest more explicitly. To reach the fourth line, we apply the Gaussian Minimax Theorem in the usual way to replace the random matrix $\boldsymbol{\Gamma}$ with two standard normal vectors $\boldsymbol{g} \in \mathbb{R}^m$ and $\boldsymbol{h} \in \mathbb{R}^n$. In a rough sense, the remaining part of the random matrix $\boldsymbol{\Phi}$ is negligible. The term $(\boldsymbol{u}_{I^c} \cdot \boldsymbol{g}_{I^c})$ generates the Gaussian width $\mathscr{W}(U)$, defined in (3.7), while the term $(\boldsymbol{h}_{J^c} \cdot \boldsymbol{t}_{J^c})$ contributes a dimensional factor $-\sqrt{n}$.

Apply the increasing convex function $(\cdot)_+^2$ to the inequality in the last display, and invoke Jensen's inequality to draw out the expectation. Notice that

$$\mathscr{W}(U) = \mathbb{E} \max_{\boldsymbol{u} \in K \cap \mathsf{B}^m} \boldsymbol{u} \cdot \boldsymbol{g} \geq \mathbb{E} \max_{\boldsymbol{u} \in K \cap \mathsf{S}^{m-1}} \boldsymbol{u} \cdot \boldsymbol{g} = \mathscr{W}(\Omega).$$

Finally, $\big(\mathscr{E}_n(\Omega)\big)_- = \big(\mathscr{W}(\Omega) - \sqrt{n}\big)_+$ because of (4.2). This point completes the proof.

20.2. **Proposition 20.1: Duality for the RAP Functional.** The first step in the argument is to apply the minimax inequality to pass to a saddle-point formulation that is amenable to analysis with the Gaussian Minimax Theorem.

Lemma 20.2 (Proposition 20.1: Duality)**.** *Adopt the notation and hypotheses of Proposition 20.1. For any point* $\boldsymbol{t} \in \mathbb{R}^n$,

$$\min_{\boldsymbol{s} \in K^\circ} \|\boldsymbol{s} - \boldsymbol{\Psi}\boldsymbol{t}\| \geq \max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \boldsymbol{\Psi}\boldsymbol{t}$$

where $U := K \cap \mathsf{B}^m$.

Proof. Write the norm as maximum:

$$\min_{\boldsymbol{s} \in K^\circ} \|\boldsymbol{s} - \boldsymbol{\Psi}\boldsymbol{t}\| = \min_{\boldsymbol{s} \in K^\circ} \max_{\boldsymbol{u} \in \mathsf{B}^m} \boldsymbol{u} \cdot (\boldsymbol{\Psi}\boldsymbol{t} - \boldsymbol{s}).$$

The minimax inequality allows us to interchange the maximum and minimum:

$$\min_{\boldsymbol{s}\in K^\circ} \|\boldsymbol{s}-\boldsymbol{\Psi}\boldsymbol{t}\| \geq \max_{\boldsymbol{u}\in \mathsf{B}^m} \min_{\boldsymbol{s}\in K^\circ} \boldsymbol{u}\cdot(\boldsymbol{\Psi}\boldsymbol{t}-\boldsymbol{s}) = \max_{\boldsymbol{u}\in \mathsf{B}^m} \left(\boldsymbol{u}\cdot\boldsymbol{\Psi}\boldsymbol{t} - \max_{\boldsymbol{s}\in K^\circ} \boldsymbol{u}\cdot\boldsymbol{s}\right)$$

The value of $\max_{\boldsymbol{s}\in K^\circ} \boldsymbol{u}\cdot\boldsymbol{s}$ equals zero when $\boldsymbol{u}\in(K^\circ)^\circ = K$; otherwise, it takes the value $+\infty$. This step uses the assumption that K is closed and convex. We conclude that

$$\min_{\boldsymbol{s}\in K^\circ} \|\boldsymbol{s}-\boldsymbol{\Psi}\boldsymbol{t}\| \geq \max_{\boldsymbol{u}\in K\cap \mathsf{B}^m} \boldsymbol{u}\cdot\boldsymbol{\Psi}\boldsymbol{t}.$$

This is the stated result. □

20.3. **Proposition 20.1: Reducing the Gaussian Matrix to Some Gaussian Vectors.** By an argument similar with the proof of Lemma 14.5, we can replace the Gaussian block of $\boldsymbol{\Psi}$ with two Gaussian vectors.

Lemma 20.3 (Proposition 20.1: Reducing the Gaussian Matrix)**.** *Adopt the notation and hypotheses of Proposition 20.1. Then*

$$\mathbb{E}\min_{\|\boldsymbol{t}\|=1}\max_{\boldsymbol{u}\in U} \boldsymbol{u}\cdot\begin{bmatrix}\boldsymbol{\Phi}_{IJ} & \boldsymbol{\Phi}_{IJ^c}\\ \boldsymbol{\Phi}_{I^cJ} & \boldsymbol{\Gamma}_{I^cJ^c}\end{bmatrix}\boldsymbol{t} \geq \mathbb{E}\min_{\|\boldsymbol{t}\|=1}\max_{\boldsymbol{u}\in U}\left(\boldsymbol{u}\cdot\begin{bmatrix}\boldsymbol{\Phi}_{IJ} & \boldsymbol{\Phi}_{IJ^c}\\ \boldsymbol{\Phi}_{I^cJ} & \boldsymbol{0}\end{bmatrix}\boldsymbol{t} + (\boldsymbol{u}_{I^c}\cdot\boldsymbol{g}_{I^c})\|\boldsymbol{t}_{J^c}\| + \|\boldsymbol{u}_{I^c}\|(\boldsymbol{h}_{J^c}\cdot\boldsymbol{t}_{J^c})\right) - 2,$$

where $\boldsymbol{g}\in\mathbb{R}^m$ and $\boldsymbol{h}\in\mathbb{R}^n$ be independent standard normal vectors.

Proof. There are no new ideas in this bound, so we refer the reader to Lemmas 14.4 and 14.5 for the pattern of argument. □

20.4. **Proposition 20.1: Finding the Gaussian Width.** To prove Proposition 20.1, most of the difficulty arises when we seek a good lower bound for the minimax problem that appears in Lemma 20.3. We have the following result.

Lemma 20.4 (Proposition 20.1: Finding the Gaussian Width)**.** *Adopt the notation and hypotheses of Proposition 20.1. Define the set $U := K\cap\mathsf{B}^m$. Then*

$$\mathbb{E}\min_{\|\boldsymbol{t}\|=1}\max_{\boldsymbol{u}\in U}\left(\boldsymbol{u}\cdot\begin{bmatrix}\boldsymbol{\Phi}_{IJ} & \boldsymbol{\Phi}_{IJ^c}\\ \boldsymbol{\Phi}_{I^cJ} & \boldsymbol{0}\end{bmatrix}\boldsymbol{t} + (\boldsymbol{u}_{I^c}\cdot\boldsymbol{g}_{I^c})\|\boldsymbol{t}_{J^c}\| + \|\boldsymbol{u}_{I^c}\|(\boldsymbol{h}_{J^c}\cdot\boldsymbol{t}_{J^c})\right) \geq \frac{\mathscr{W}(U)-\sqrt{n}}{CB\sqrt{\log m}} - CB^2\sqrt{k\log m}.$$

Proof. The proof of this bound is lengthy, so we break the argument into several steps. The overall result follows when we sequence the inequalities in Sublemmas 20.5, 20.6, 20.7, and 20.8 and consolidate the error terms. □

20.4.1. *Lemma 20.4: Simplifying the Minimax I.* The first step in the proof of Lemma 20.4 is to simplify the minimax so we can identify the key terms.

Sublemma 20.5 (Lemma 20.4: Simplifying the Minimax I)**.** *Adopt the notation and hypotheses of Lemma 20.4. Then*

$$\begin{aligned}\mathbb{E}\min_{\|\boldsymbol{t}\|=1}\max_{\boldsymbol{u}\in U}&\left(\boldsymbol{u}\cdot\begin{bmatrix}\boldsymbol{\Phi}_{IJ} & \boldsymbol{\Phi}_{IJ^c}\\ \boldsymbol{\Phi}_{I^cJ} & \boldsymbol{0}\end{bmatrix}\boldsymbol{t} + (\boldsymbol{u}_{I^c}\cdot\boldsymbol{g}_{I^c})\|\boldsymbol{t}_{J^c}\| + \|\boldsymbol{u}_{I^c}\|(\boldsymbol{h}_{J^c}\cdot\boldsymbol{t}_{J^c})\right)\\ &\geq \mathbb{E}\min_{\|\boldsymbol{t}\|=1}\max_{\boldsymbol{u}\in U}\left(\boldsymbol{u}\cdot\begin{bmatrix}\boldsymbol{\Phi}_J & \boldsymbol{g}\end{bmatrix}\begin{bmatrix}\boldsymbol{t}_J\\ \|\boldsymbol{t}_{J^c}\|\end{bmatrix} - \sqrt{n}\|\boldsymbol{t}_{J^c}\|\right)_+ - CB^2\sqrt{k}.\end{aligned}$$

Proof. Let us introduce notation for the quantity of interest:

$$Q_1 := \mathbb{E}\min_{\|\boldsymbol{t}\|=1}\max_{\boldsymbol{u}\in U}\left(\boldsymbol{u}\cdot\begin{bmatrix}\boldsymbol{\Phi}_{IJ} & \boldsymbol{\Phi}_{IJ^c}\\ \boldsymbol{\Phi}_{I^cJ} & \boldsymbol{0}\end{bmatrix}\boldsymbol{t} + (\boldsymbol{u}_{I^c}\cdot\boldsymbol{g}_{I^c})\|\boldsymbol{t}_{J^c}\| + \|\boldsymbol{u}_{I^c}\|(\boldsymbol{h}_{J^c}\cdot\boldsymbol{t}_{J^c})\right)_+. \tag{20.1}$$

We can introduce the positive-part operator because the fact that $\boldsymbol{0}\in U$ ensures that the minimax is nonnegative.

The first step in the argument is to reintroduce the missing piece of the random vector $\boldsymbol{g}$. Adding and subtracting the quantity $(\boldsymbol{u}_I \cdot \boldsymbol{g}_I)\,\|\boldsymbol{t}_{J^c}\|$ inside the positive-part operator in (20.1), we obtain the bound

$$\begin{aligned} Q_1 &\geq \mathbb{E} \min_{\|\boldsymbol{t}\|=1} \max_{\boldsymbol{u}\in U} \left(\boldsymbol{u} \cdot \begin{bmatrix} \boldsymbol{\Phi}_{IJ} & \boldsymbol{\Phi}_{IJ^c} \\ \boldsymbol{\Phi}_{I^cJ} & \mathbf{0} \end{bmatrix} \boldsymbol{t} + (\boldsymbol{u}\cdot\boldsymbol{g})\,\|\boldsymbol{t}_{J^c}\| + \|\boldsymbol{u}_{I^c}\|\,(\boldsymbol{h}_{J^c}\cdot\boldsymbol{t}_{J^c}) \right)_+ - \mathbb{E} \max_{\boldsymbol{u}\in U} \boldsymbol{u}_I \cdot \boldsymbol{g}_I \\ &\geq \mathbb{E} \min_{\|\boldsymbol{t}\|=1} \max_{\boldsymbol{u}\in U} \left(\boldsymbol{u} \cdot \begin{bmatrix} \boldsymbol{\Phi}_{IJ} & \boldsymbol{\Phi}_{IJ^c} \\ \boldsymbol{\Phi}_{I^cJ} & \mathbf{0} \end{bmatrix} \boldsymbol{t} + (\boldsymbol{u}\cdot\boldsymbol{g})\,\|\boldsymbol{t}_{J^c}\| + \|\boldsymbol{u}_{I^c}\|\,(\boldsymbol{h}_{J^c}\cdot\boldsymbol{t}_{J^c}) \right)_+ - \sqrt{k}. \end{aligned} \tag{20.2}$$

The second inequality holds because $\#I = k$ and U is a subset of the unit ball. This step is similar with the proof of Lemma 14.8.

Next, we combine the terms in (20.2) involving $\boldsymbol{\Phi}_{IJ^c}$ and the *row* vector $\boldsymbol{h}_{J^c}$. Since $\|\boldsymbol{u}\| \leq 1$,

$$\begin{aligned} \boldsymbol{u}_I \cdot \boldsymbol{\Phi}_{IJ^c}\boldsymbol{t}_{J^c} + \|\boldsymbol{u}_{I^c}\|\,(\boldsymbol{h}_{J^c}\cdot\boldsymbol{t}_{J^c}) &= \begin{bmatrix} \boldsymbol{u}_I \\ \|\boldsymbol{u}_{I^c}\| \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{\Phi}_{IJ^c} \\ \boldsymbol{h}_{J^c} \end{bmatrix} \boldsymbol{t}_{J^c} \\ &\geq -\left\| \begin{bmatrix} \boldsymbol{\Phi}_{IJ^c} \\ \boldsymbol{h}_{J^c} \end{bmatrix} \right\| \|\boldsymbol{t}_{J^c}\| \\ &\geq -\sqrt{n}\,\|\boldsymbol{t}_{J^c}\| - \left(\left\| \begin{bmatrix} \boldsymbol{\Phi}_{IJ^c} \\ \boldsymbol{h}_{J^c} \end{bmatrix} \right\| - \sqrt{n} \right)_+ . \end{aligned} \tag{20.3}$$

The $(k+1)\times(n-k)$ random matrix on the right-hand side has independent, standardized entries that satisfy the subgaussian estimate (B.1) with bound B. Repeating the calculations in (14.17), we see that

$$\mathbb{E} \left(\left\| \begin{bmatrix} \boldsymbol{\Phi}_{IJ^c} \\ \boldsymbol{h}_{J^c} \end{bmatrix} \right\| - \sqrt{n} \right)_+ \leq CB^2\sqrt{k}. \tag{20.4}$$

Apply the estimate (20.3) inside the minimax in (20.2) and then use (20.4) to arrive at the lower bound

$$\begin{aligned} Q_1 &\geq \mathbb{E} \min_{\|\boldsymbol{t}\|=1} \max_{\boldsymbol{u}\in U} \left(\boldsymbol{u} \cdot \begin{bmatrix} \boldsymbol{\Phi}_{IJ} \\ \boldsymbol{\Phi}_{I^cJ} \end{bmatrix} \boldsymbol{t}_J + (\boldsymbol{u}\cdot\boldsymbol{g})\,\|\boldsymbol{t}_{J^c}\| - \sqrt{n}\,\|\boldsymbol{t}_{J^c}\| \right)_+ - \mathbb{E} \left(\left\| \begin{bmatrix} \boldsymbol{\Phi}_{IJ^c} \\ \boldsymbol{h}_{J^c} \end{bmatrix} \right\| - \sqrt{n} \right)_+ \\ &\geq \mathbb{E} \min_{\|\boldsymbol{t}\|=1} \max_{\boldsymbol{u}\in U} \left(\boldsymbol{u} \cdot \begin{bmatrix} \boldsymbol{\Phi}_J & \boldsymbol{g} \end{bmatrix} \begin{bmatrix} \boldsymbol{t}_J \\ \|\boldsymbol{t}_{J^c}\| \end{bmatrix} - \sqrt{n}\,\|\boldsymbol{t}_{J^c}\| \right)_+ - CB^2\sqrt{k}. \end{aligned}$$

In the second line, we have simply consolidated terms. □

20.4.2. *Lemma 20.4: Simplifying the Minimax II.* The next step in the proof of Lemma 20.4 is to reduce the minimax problem in Sublemma 20.5 to a scalar optimization problem.

Sublemma 20.6 (Lemma 20.4: Simplifying the Minimax II)**.** *Adopt the notation and hypotheses of Lemma 20.4. Then*

$$\begin{aligned} &\mathbb{E} \min_{\|\boldsymbol{t}\|=1} \max_{\boldsymbol{u}\in U} \left(\boldsymbol{u} \cdot \begin{bmatrix} \boldsymbol{\Phi}_J & \boldsymbol{g} \end{bmatrix} \begin{bmatrix} \boldsymbol{t}_j \\ \|\boldsymbol{t}_{J^c}\| \end{bmatrix} - \sqrt{n}\,\|\boldsymbol{t}_{J^c}\| \right)_+ \\ &\qquad \geq \mathbb{E} \min_{\alpha\in[0,1]} \max \left\{ 0, \left(\max_{\boldsymbol{u}\in U} \boldsymbol{u}\cdot\boldsymbol{g} - \sqrt{n} \right)\alpha, \min_{\|\boldsymbol{s}\|=1} \max_{\boldsymbol{u}\in U} \boldsymbol{u} \cdot \begin{bmatrix} \boldsymbol{\Phi}_J & \boldsymbol{g} \end{bmatrix} \boldsymbol{s} - \sqrt{n}\alpha \right\} - \sqrt{k}. \end{aligned}$$

Proof. Introduce the notation

$$Q_2 := \max_{\boldsymbol{u}\in U} \left(\boldsymbol{u} \cdot \begin{bmatrix} \boldsymbol{\Phi}_J & \boldsymbol{g} \end{bmatrix} \begin{bmatrix} \boldsymbol{t}_J \\ \|\boldsymbol{t}_{J^c}\| \end{bmatrix} - \sqrt{n}\,\|\boldsymbol{t}_{J^c}\| \right)_+ . \tag{20.5}$$

We will develop two lower bounds on the maximum by coupling $\boldsymbol{u}$ to the random matrix in different ways. Afterward, we combine these results into a single bound.

In the first place, we can choose the *row* vector $\boldsymbol{u}$ so that it depends only on the remaining Gaussian vector:

$$\boldsymbol{u}(\boldsymbol{g}) \in \arg\max_{\boldsymbol{u}\in U} \boldsymbol{u}\cdot\boldsymbol{g}.$$

Since $\|\boldsymbol{t}\| = 1$, we obtain the bound

$$\begin{aligned}\max_{\boldsymbol{u}\in U}\left(\boldsymbol{u}\cdot\begin{bmatrix}\boldsymbol{\Phi}_J & \boldsymbol{g}\end{bmatrix}\begin{bmatrix}\boldsymbol{t}_J\\ \|\boldsymbol{t}_{J^c}\|\end{bmatrix} - \sqrt{n}\,\|\boldsymbol{t}_{J^c}\|\right) &\geq \left(\boldsymbol{u}(\boldsymbol{g})\cdot\boldsymbol{g} - \sqrt{n}\right)\|\boldsymbol{t}_{J^c}\| + \boldsymbol{u}(\boldsymbol{g})\cdot\boldsymbol{\Phi}_J\boldsymbol{t}_J\\ &\geq \left(\max_{\boldsymbol{u}\in U}\boldsymbol{u}\cdot\boldsymbol{g} - \sqrt{n}\right)\|\boldsymbol{t}_{J^c}\| - \left\|\boldsymbol{u}(\boldsymbol{g})\boldsymbol{\Phi}_J\right\|.\end{aligned} \tag{20.6}$$

The second term on the right-hand side of (20.6) satisfies

$$\mathbb{E}\left\|\boldsymbol{u}(\boldsymbol{g})\boldsymbol{\Phi}_J\right\| \leq \left(\mathbb{E}\left\|\boldsymbol{u}(\boldsymbol{g})\boldsymbol{\Phi}_J\right\|^2\right)^{1/2} \leq \sqrt{k}. \tag{20.7}$$

Indeed, $\boldsymbol{u}(\boldsymbol{g})$ is a random vector with $\|\boldsymbol{u}(\boldsymbol{g})\| \leq 1$ that is stochastically independent from $\boldsymbol{\Phi}_J$, and the $m \times k$ random matrix $\boldsymbol{\Phi}_J$ has independent, standardized entries.

The second bound is even simpler. Since $\|\boldsymbol{t}\| = 1$,

$$\max_{\boldsymbol{u}\in U}\left(\boldsymbol{u}\cdot\begin{bmatrix}\boldsymbol{\Phi}_J & \boldsymbol{g}\end{bmatrix}\begin{bmatrix}\boldsymbol{t}_J\\ \|\boldsymbol{t}_{J^c}\|\end{bmatrix} - \sqrt{n}\,\|\boldsymbol{t}_{J^c}\|\right) \geq \min_{\|\boldsymbol{s}\|=1}\max_{\boldsymbol{u}\in U}\boldsymbol{u}\cdot\begin{bmatrix}\boldsymbol{\Phi}_J & \boldsymbol{g}\end{bmatrix}\boldsymbol{s} - \sqrt{n}\,\|\boldsymbol{t}_{J^c}\|. \tag{20.8}$$

In this expression, the variable $\boldsymbol{s} \in \mathbb{R}^{k+1}$.

Introducing (20.6) and (20.8) into (20.5), we arrive at

$$\begin{aligned}Q_2 &\geq \mathbb{E}\min_{\|\boldsymbol{t}\|=1}\max\left\{0, \left(\max_{\boldsymbol{u}\in U}\boldsymbol{u}\cdot\boldsymbol{g} - \sqrt{n}\right)\|\boldsymbol{t}_{J^c}\| - \left\|\boldsymbol{u}(\boldsymbol{g})\boldsymbol{\Phi}_J\right\|, \min_{\|\boldsymbol{s}\|=1}\max_{\boldsymbol{u}\in U}\boldsymbol{u}\cdot\begin{bmatrix}\boldsymbol{\Phi}_J & \boldsymbol{g}\end{bmatrix}\boldsymbol{s} - \sqrt{n}\,\|\boldsymbol{t}_{J^c}\|\right\}\\ &\geq \mathbb{E}\min_{\alpha\in[0,1]}\max\left\{0, \left(\max_{\boldsymbol{u}\in U}\boldsymbol{u}\cdot\boldsymbol{g} - \sqrt{n}\right)\alpha, \min_{\|\boldsymbol{s}\|=1}\max_{\boldsymbol{u}\in U}\boldsymbol{u}\cdot\begin{bmatrix}\boldsymbol{\Phi}_J & \boldsymbol{g}\end{bmatrix}\boldsymbol{s} - \sqrt{n}\alpha\right\} - \sqrt{k}.\end{aligned} \tag{20.9}$$

The zero branch in the maximum accounts for the positive-part operator in (20.5). The second line follows from (20.7). We have also introduced a new parameter α to stand in for $\|\boldsymbol{t}_{J^c}\|$. □

20.4.3. *Lemma 20.4: Probabilistic Bounds.* The last major step in Lemma 20.4 is to develop probabilistic bounds for the terms that arise in Sublemma 20.6.

Sublemma 20.7 (Lemma 20.4: Probabilistic Bounds)**.** *Adopt the notation and hypotheses of Lemma 20.4. Then*

$$\begin{aligned}&\mathbb{E}\min_{\alpha\in[0,1]}\max\left\{0, \left(\max_{\boldsymbol{u}\in U}\boldsymbol{u}\cdot\boldsymbol{g} - \sqrt{n}\right)\alpha, \min_{\|\boldsymbol{s}\|=1}\max_{\boldsymbol{u}\in U}\boldsymbol{u}\cdot\begin{bmatrix}\boldsymbol{\Phi}_J & \boldsymbol{g}\end{bmatrix}\boldsymbol{s} - \sqrt{n}\alpha\right\}\\ &\qquad\geq \frac{1}{2}\min_{\alpha\in[0,1]}\max\left\{(\mathscr{W}(U) - \sqrt{n} - 2)\alpha, \frac{\mathscr{W}(U) - 2}{2B\sqrt{\log m}} - \sqrt{n}\alpha\right\} - CB^2\sqrt{k\log m}.\end{aligned}$$

Proof. Introduce the notation

$$Q_3 := \mathbb{E}\min_{\alpha\in[0,1]}\max\left\{0, \left(\max_{\boldsymbol{u}\in U}\boldsymbol{u}\cdot\boldsymbol{g} - \sqrt{n}\right)\alpha, \min_{\|\boldsymbol{s}\|=1}\max_{\boldsymbol{u}\in U}\boldsymbol{u}\cdot\begin{bmatrix}\boldsymbol{\Phi}_J & \boldsymbol{g}\end{bmatrix}\boldsymbol{s} - \sqrt{n}\alpha\right\}.$$

We assume that $\mathscr{W}(U) \geq 2 + \sqrt{n}$, which is permitted because the final result would otherwise become vacuous.

The next stage in the argument is to introduce probabilistic bounds for the branches of the maximum and use these to control the expectation. It is convenient to abbreviate

$$X := \max_{\boldsymbol{u}\in U}\boldsymbol{u}\cdot\boldsymbol{g} - \sqrt{n} \quad\text{and}\quad Y := \min_{\|\boldsymbol{s}\|=1}\max_{\boldsymbol{u}\in U}\boldsymbol{u}\cdot\begin{bmatrix}\boldsymbol{\Phi}_J & \boldsymbol{g}\end{bmatrix}\boldsymbol{s}.$$

Note that $\mathbb{E}\,X = \mathscr{W}(U) - \sqrt{n}$. Since $\boldsymbol{g} \mapsto X(\boldsymbol{g})$ is 1-Lipschitz, the Gaussian concentration inequality, Fact A.2, implies that

$$\mathbb{P}\left\{X \geq \mathscr{W}(U) - \sqrt{n} - 2\right\} \geq 3/4.$$

On the other hand, Sublemma 20.10 will demonstrate that

$$\mathbb{P}\left\{Y \geq \frac{\mathscr{W}(U)}{CB\sqrt{\log m}} - CB\sqrt{k\log m}\right\} \geq 3/4.$$

Therefore, taking complements and a union bound,

$$\mathbb{P}\left\{ X > \mathscr{W}(U) - \sqrt{n} - 2 \quad \text{and} \quad Y > \frac{\mathscr{W}(U)}{CB\sqrt{\log m}} - CB\sqrt{k \log m} \right\} \geq 1/2. \tag{20.10}$$

For each nonnegative random variable Z and each number $L > 0$, it holds that $\mathbb{E}\, Z \geq L\mathbb{P}\{Z > L\}$. Using the estimate (20.9) and the probability bound (20.10), we find that

$$\begin{aligned} Q_3 &\geq \frac{1}{2} \min_{\alpha \in [0,1]} \max\left\{ 0, (\mathscr{W}(U) - \sqrt{n} - 2)\alpha, \left(\frac{\mathscr{W}(U)}{CB\sqrt{\log m}} - CB\sqrt{k \log m} \right) - \sqrt{n}\alpha \right\} - \sqrt{k} \\ &\geq \frac{1}{2} \min_{\alpha \in [0,1]} \max\left\{ (\mathscr{W}(U) - \sqrt{n} - 2)\alpha, \frac{\mathscr{W}(U) - 2}{CB\sqrt{\log m}} - \sqrt{n}\alpha \right\} - CB^2\sqrt{k \log m}. \end{aligned}$$

Once again, we have used shift-invariance of the maximum to combine the error terms. For convenience, we have also dropped the zero branch of the maximum and introduced the number two into the numerator in the second branch. □

20.4.4. *Lemma 20.4: Solving the Scalar Minimax Problem.* The final step in the proof of Lemma 20.4 is to solve the scalar minimax problem that emerges in Sublemma 20.7.

Sublemma 20.8 (Lemma 20.4: Solving the Minimax Problem)**.** *Adopt the notation and hypotheses of Lemma 20.4. Then*

$$\min_{\alpha \in [0,1]} \max\left\{ (\mathscr{W}(U) - \sqrt{n} - 2)\alpha, \frac{\mathscr{W}(U) - 2}{CB\sqrt{\log m}} - \sqrt{n}\alpha \right\} \geq \frac{\mathscr{W}(U) - \sqrt{n} - 2}{CB\sqrt{\log m}}.$$

Proof. The first branch of the maximum is increasing in α while the second branch is decreasing in α, so the minimum occurs when the two branches are equal, provided that this situation occurs when $\alpha \in [0,1]$. Setting the branches equal, we identify the point $\alpha_\star$ where the saddle value is achieved.

$$\alpha_\star := \frac{b}{a+c} \quad \text{where} \quad a := \mathscr{W}(U) - \sqrt{n} - 2 \quad \text{and} \quad b := \frac{\mathscr{W}(U) - 2}{CB\sqrt{\log m}} \quad \text{and} \quad c := \sqrt{n}.$$

We quickly verify that $\alpha_\star \in [0,1]$, so the minimax takes the value

$$\frac{ab}{a+c} = \frac{\mathscr{W}(U) - \sqrt{n} - 2}{\mathscr{W}(U) - 2} \times \frac{\mathscr{W}(U) - 2}{2B\sqrt{\log m}} = \frac{\mathscr{W}(U) - \sqrt{n} - 2}{CB\sqrt{\log m}}.$$

This is the required estimate. □

20.4.5. *Sublemma 20.7: Probabilistic Lower Bound for Bilinear Minimax.* In this section, we explain how to obtain a lower bound for the minimax problem in (20.8) in terms of the Gaussian width of the set U.

Sublemma 20.9 (Sublemma 20.7: Probability Bound for Bilinear Minimax)**.** *Assume that $k < m$, and let $\boldsymbol{X}$ be an $m \times k$ random matrix that satisfies Model 2.1 with bound B. Let $\boldsymbol{g} \in \mathbb{R}^m$ be standard normal. Let U be a subset of the unit ball in $\mathbb{R}^m$. Then*

$$\mathbb{P}\left\{ \min_{\|\boldsymbol{s}\|=1} \max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \begin{bmatrix} \boldsymbol{X} & \boldsymbol{g} \end{bmatrix} \boldsymbol{s} > \frac{\mathscr{W}(U)}{CB\sqrt{\log m}} - CB\sqrt{k \log m} \right\} \geq 3/4.$$

Proof. Fix $\varepsilon = m^{-1}$. Let $\mathscr{N}$ be an ε-net for the unit sphere in $\mathbb{R}^{k+1}$. The cardinality of the net satisfies $\log(\mathscr{N}) \leq (k+1)\log(3m)$ by the standard volumetric argument [Ver12, Lem. 5.2]

We can estimate the quantity of interest below by discretizing the parameter $\boldsymbol{s}$. Since $\mathscr{N}$ and U are subsets of the unit ball, we have the bound

$$\min_{\|\boldsymbol{s}\|=1} \max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \begin{bmatrix} \boldsymbol{X} & \boldsymbol{g} \end{bmatrix} \boldsymbol{s} \geq \min_{\boldsymbol{s} \in \mathscr{N}} \max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \begin{bmatrix} \boldsymbol{X} & \boldsymbol{g} \end{bmatrix} \boldsymbol{s} - \left\| \begin{bmatrix} \boldsymbol{X} & \boldsymbol{g} \end{bmatrix} \right\| \varepsilon \tag{20.11}$$

We will establish a probabilistic lower bound for the right-hand side of (20.11).

First, we develop a probability bound for the second term on the right-hand side of (20.11). A simple spectral norm estimate suffices. The $m \times k$ random matrix $\boldsymbol{X}$ has standardized entries and $\boldsymbol{g} \in \mathbb{R}^m$ is standard normal, so

$$\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{X} & \boldsymbol{g}\end{bmatrix}\right\| \le \mathbb{E}\left\|\begin{bmatrix}\boldsymbol{X} & \boldsymbol{g}\end{bmatrix}\right\|_{\mathrm{F}} \le \sqrt{(k+1)m}.$$

As usual, $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. Markov's inequality now implies that

$$\mathbb{P}\left\{\left\|\begin{bmatrix}\boldsymbol{X} & \boldsymbol{g}\end{bmatrix}\right\| \ge 6\sqrt{(k+1)m}\right\} \ge 5/6.$$

It follows that

$$\mathbb{P}\left\{\left\|\begin{bmatrix}\boldsymbol{X} & \boldsymbol{g}\end{bmatrix}\right\| \varepsilon \ge 6\right\} \ge 5/6. \tag{20.12}$$

We have used the facts that $\varepsilon = m^{-1}$ and $k < m$.

Let us turn to the second quantity on the right-hand side of (20.11). We develop a strong probability bound for each fixed point $\boldsymbol{s} \in \mathcal{N}$, and we extend it to the entire net using the union bound. For technical reasons, it is easier to treat the random matrix $\boldsymbol{X}$ and the random vector $\boldsymbol{g}$ separately.

Fix a point $\boldsymbol{s} \in \mathcal{N}$, and decompose it as $\boldsymbol{s} = \begin{bmatrix}\boldsymbol{s}_1 & s_{k+1}\end{bmatrix}$ where $\boldsymbol{s}_1 \in \mathbb{R}^k$. Construct a random vector $\boldsymbol{u} \in U$ that satisfies

$$\boldsymbol{u}(\boldsymbol{X}) \in \arg\max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \boldsymbol{X}\boldsymbol{s}_1.$$

We may calculate that

$$\begin{aligned}\max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \begin{bmatrix}\boldsymbol{X} & \boldsymbol{g}\end{bmatrix} \boldsymbol{s} &\ge \boldsymbol{u}(\boldsymbol{X}) \cdot \begin{bmatrix}\boldsymbol{X} & \boldsymbol{g}\end{bmatrix} \begin{bmatrix}\boldsymbol{s}_1 \\ s_{k+1}\end{bmatrix} \\ &= \boldsymbol{u}(\boldsymbol{X}) \cdot \boldsymbol{X}\boldsymbol{s}_1 + (\boldsymbol{u}(\boldsymbol{X}) \cdot \boldsymbol{g}) s_{k+1} \\ &\ge \max_{\boldsymbol{u} \in S} \boldsymbol{u} \cdot \boldsymbol{X}\boldsymbol{s}_1 - \left|\boldsymbol{u}(\boldsymbol{X}) \cdot \boldsymbol{g}\right|.\end{aligned} \tag{20.13}$$

The last estimate relies on the fact that $|s_{k+1}| \le 1$ because $\|\boldsymbol{s}\| = 1$.

The second term on the right-hand side of (20.13) is easy to handle using the Gaussian concentration inequality, Fact A.2:

$$\mathbb{P}\left\{\left|\boldsymbol{u}(\boldsymbol{X}) \cdot \boldsymbol{g}\right| \ge \zeta\right\} \le 2\mathrm{e}^{-\zeta^2/2} \tag{20.14}$$

Indeed, $\boldsymbol{g} \mapsto \boldsymbol{u}(\boldsymbol{X}) \cdot \boldsymbol{g}$ is a 1-Lipschitz function with mean zero because the random vector $\boldsymbol{u}(\boldsymbol{X})$ is stochastically independent from $\boldsymbol{g}$ and has norm bounded by one.

We can interpret the first term on the right-hand side of (20.13) as an "empirical width." It takes some work to compare this quantity with the Gaussian width. Sublemma 20.10 contains a bound for the expectation, and Sublemma 20.11 contains a tail bound. Together, they deliver the probability inequality

$$\mathbb{P}\left\{\max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \boldsymbol{X}\boldsymbol{s}_1 \le \frac{\mathscr{W}(U)}{CB\sqrt{\log m}} - \zeta\right\} \le \mathrm{e}^{-\zeta^2/(8B^2)} \quad \text{for each } \boldsymbol{s} \in \mathcal{N}. \tag{20.15}$$

In other words, the empirical width of U is comparable with the Gaussian width, modulo a logarithmic factor.

Introduce the two probability bounds (20.14) and (20.15) into the deterministic estimate (20.13). We arrive at

$$\mathbb{P}\left\{\max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \begin{bmatrix}\boldsymbol{X} & \boldsymbol{g}\end{bmatrix} \boldsymbol{s} \le \frac{\mathscr{W}(U)}{CB\sqrt{\log m}} - 2\zeta\right\} \le 3\mathrm{e}^{-\zeta^2/(8B^2)} \quad \text{for each } \boldsymbol{s} \in \mathcal{N}.$$

Finally, we take a union bound over $\boldsymbol{s} \in \mathcal{N}$ to obtain an estimate that is uniform over the net. Recall that $\log(\#\mathcal{N}) \le (k+1)\log(3m)$ and select $\zeta = 4B\sqrt{(k+1)\log(3m)}$ to reach

$$\begin{aligned}\mathbb{P}\left\{\min_{\boldsymbol{s} \in \mathcal{N}} \max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \boldsymbol{X}\boldsymbol{s}_1 \le \frac{\mathscr{W}(U)}{CB\sqrt{\log m}} - CB\sqrt{(k+1)\log(3m)}\right\} \\ \le 3(\#\mathcal{N})\mathrm{e}^{-\zeta^2/(8B^2)} \le 3\mathrm{e}^{-(k+1)\log(3m)} = 3(3m)^{-(k+1)} \le 1/18.\end{aligned} \tag{20.16}$$

The numerical estimate holds because $1 \le k < m$.

The two probability bounds (20.12) and (20.16) hold simultaneously with probability at least 3/4. Therefore, we can substitute these results into (20.11) and adjust constants to obtain the stated bound. □

20.4.6. *Sublemma 20.7: Lower Estimate for the Empirical Width.* The next sublemma demonstrates that the Gaussian width of a set is not more than a logarithmic factor larger than the empirical width of the set as computed with bounded random variables. This is the only step in the argument for bounded random matrices that requires the symmetry assumption.

Sublemma 20.10 (Sublemma 20.7: Empirical Width Bound)**.** *Adopt the notation and hypotheses of Sublemma 20.9, and let $\boldsymbol{s}$ be a fixed unit-norm vector. Then*

$$\mathbb{E} \max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \boldsymbol{X}\boldsymbol{s} \geq \frac{\mathscr{W}(U)}{CB\sqrt{\log m}}.$$

Proof of Sublemma 20.10. Define the random vector $\boldsymbol{v} := (V_1, \dots, V_n) := \boldsymbol{X}\boldsymbol{s}$. Our goal is to compare the empirical width of the set U computed using the vector $\boldsymbol{v}$ with the Gaussian width of the set.

First, we develop a lower bound on the first moment of the entries of $\boldsymbol{v}$. Fix an index i. Since the entries of $\boldsymbol{X}$ are independent and symmetric

$$\mathbb{E}|V_i| = \mathbb{E}\left|\sum\nolimits_{j=1}^{k} X_{ij} s_j\right| = \mathbb{E}\left|\sum\nolimits_{j=1}^{k} \eta_j X_{ij} s_j\right|$$

where $\{\eta_j\}$ is an independent family of Rademacher random variables, independent from $\boldsymbol{X}$. The Khintchine inequality [LO94] allows us to compare the first moment with the second moment:

$$\mathbb{E}|V_i| \geq \frac{1}{\sqrt{2}} \mathbb{E}_{\boldsymbol{X}} \left(\mathbb{E}_{\boldsymbol{\eta}} \left| \sum\nolimits_{j=1}^{k} \eta_j X_{ij} s_j \right|^2 \right)^{1/2} = \frac{1}{\sqrt{2}} \mathbb{E} \left(\sum\nolimits_{j=1}^{k} |X_{ij}|^2 |s_j|^2 \right)^{1/2}.$$

Since $\boldsymbol{s}$ has unit norm, we can regard the sum as a weighted average, and we can invoke Jensen's inequality to draw the average out of the square root:

$$\mathbb{E}|V_i| \geq \frac{1}{\sqrt{2}} \sum\nolimits_{j=1}^{k} \big(\mathbb{E}|X_{ij}|\big) |s_j|^2.$$

Last, note that $1 = \mathbb{E}|X_{ij}|^2 \leq B\,\mathbb{E}|X_{ij}|$ because the entries of $\boldsymbol{X}$ are standardized and bounded by B. Thus,

$$\mathbb{E}|V_i| \geq \frac{1}{\sqrt{2}} \sum\nolimits_{j=1}^{k} \frac{1}{B} |s_j|^2 = \frac{1}{B\sqrt{2}}.$$

Let us bound the width-like functional below. Since $\boldsymbol{v}$ has independent, symmetric coordinates,

$$\mathbb{E} \max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \boldsymbol{X}\boldsymbol{s} = \mathbb{E} \max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \boldsymbol{v} = \mathbb{E} \max_{\boldsymbol{u} \in U} \sum\nolimits_{i=1}^{n} V_i u_i = \mathbb{E} \max_{\boldsymbol{u} \in U} \sum\nolimits_{i=1}^{n} \eta_i |V_i| u_i$$

where, again, $\{\eta_i\}$ is an independent family of Rademacher random variables, independent from $\boldsymbol{u}$. Using a corollary of the contraction principle [LT11, Lem. 4.5],

$$\mathbb{E} \max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \boldsymbol{X}\boldsymbol{s} \geq \min\nolimits_i (\mathbb{E}|V_i|)\, \mathbb{E}_{\boldsymbol{\eta}} \max_{\boldsymbol{u} \in U} \sum\nolimits_{i=1}^{n} \eta_i u_i \geq \frac{1}{B\sqrt{2}} \mathbb{E} \max_{\boldsymbol{u} \in U} \sum\nolimits_{i=1}^{n} \eta_i u_i.$$

Applying the contraction principle again [LT11, Eqn. (4.9)], we can randomize the sum with independent Gaussian variables:

$$\mathbb{E} \max_{\boldsymbol{u} \in U} \boldsymbol{u} \cdot \boldsymbol{X}\boldsymbol{s} \geq \frac{1}{2B\sqrt{\log n}} \mathbb{E} \max_{\boldsymbol{u} \in U} \sum\nolimits_{i=1}^{n} g_i u_i$$

Here, $\boldsymbol{g} := (g_1, \dots, g_n)$ is a standard normal vector. Identify the Gaussian width to complete the proof. □

20.4.7. *Sublemma 20.7: Lower Tail of the Empirical Width.* Last, we present a concentration inequality for the empirical width.

Sublemma 20.11 (Sublemma 20.7: Lower Tail of Empirical Width)**.** *Adopt the notation and hypotheses of Sublemma 20.9. Let $\boldsymbol{s}$ be a fixed unit-norm vector. Then*

$$\mathbb{P}\left\{\max_{\boldsymbol{u}\in U} \boldsymbol{u}\cdot\boldsymbol{X}\boldsymbol{s} \le \mathbb{E}\max_{\boldsymbol{u}\in U} \boldsymbol{u}\cdot\boldsymbol{X}\boldsymbol{s} - \zeta\right\} \le \mathrm{e}^{-\zeta^2/8B^2}.$$

The most direct proof of this result relies on a version of Talagrand's inequality [BLM13, Thm. 8.6] obtained from the transportation cost method.

Fact 20.12 (Talagrand's Inequality)**.** *Let $\mathscr{X}$ be a metric space. Suppose that $f : \mathscr{X}^p \to \mathbb{R}$ fulfills the one-sided Lipschitz bound*

$$f(\boldsymbol{a}) - f(\boldsymbol{z}) \le \sum_{i=1}^p c_i(\boldsymbol{a})\mathbb{1}_{a_i\neq z_i} \quad \text{for all } \boldsymbol{a},\boldsymbol{z}\in\mathscr{X}^p,$$

where $c_i : \mathbb{R}^p \to \mathbb{R}$ are auxiliary functions that satisfy

$$\sum_{i=1}^p c_i^2(\boldsymbol{a}) \le v \quad \text{for all } \boldsymbol{a}\in\mathscr{X}^p.$$

Let $(X_1,\dots,X_p)$ be an independent sequence of random variables taking values in $\mathscr{X}$, and define

$$Y := f(X_1,\dots,X_p).$$

Then, for all $\zeta \ge 0$,

$$\mathbb{P}\{Y - \mathbb{E}\,Y \ge \zeta\} \le \mathrm{e}^{-\zeta^2/(2v)}, \quad \text{and}$$
$$\mathbb{P}\{Y - \mathbb{E}\,Y \le -\zeta\} \le \mathrm{e}^{-\zeta^2/(2v)}.$$

Proof of Sublemma 20.11. Let $\mathscr{X}$ be the interval $[-B,B]$ equipped with the Euclidean distance. For a matrix $\boldsymbol{A}\in\mathscr{X}^{m\times k}$, introduce the function

$$f(\boldsymbol{A}) := \max_{\boldsymbol{u}\in U} \boldsymbol{u}\cdot\boldsymbol{A}\boldsymbol{s}.$$

Select a point $\boldsymbol{t}\in\arg\max_{\boldsymbol{u}\in U}\boldsymbol{u}\cdot\boldsymbol{A}\boldsymbol{s}$. Then

$$\begin{aligned} f(\boldsymbol{A}) - f(\boldsymbol{Z}) &\le \boldsymbol{t}\cdot\boldsymbol{A}\boldsymbol{s} - \boldsymbol{t}\cdot\boldsymbol{Z}\boldsymbol{s} \\ &= \sum_{i=1}^m\sum_{j=1}^k t_i s_j (a_{ij} - z_{ij}) \\ &\le \sum_{i=1}^m\sum_{j=1}^k 2B\left|t_i s_j\right| \mathbb{1}_{a_{ij}\neq z_{ij}}. \end{aligned}$$

We used the fact that the entries of $\boldsymbol{A}$ and $\boldsymbol{Z}$ are bounded in magnitude by B. With the choice $c_{ij}(\boldsymbol{A}) = 2B\left|t_i s_j\right|$, we see that

$$\sum_{i=1}^m\sum_{j=1}^k c_{ij}^2(\boldsymbol{A}) \le 4B^2\sum_{i,j}\left|t_i s_j\right|^2 \le 4B^2$$

because $\boldsymbol{t},\boldsymbol{s}$ both belong to the Euclidean unit ball. Invoke Fact 20.12 to complete the argument. □

Part V. Back Matter

Two appendices contain statements of some results that we use throughout the paper. Appendix A presents some facts about Gaussian analysis, while Appendix B describes some spectral bounds for random matrices with independent entries. We conclude with acknowledgments and a list of works cited.

Appendix A. Tools from Gaussian Analysis

We make extensive use of methods from Gaussian analysis to provide precise information about the behavior of functions of Gaussian random variables. These results come up in many places in the paper, so we have collected them here.

A.1. Concentration for Gaussian Lipschitz Functions.

We begin with two concentration results that apply to a Lipschitz function of independent Gaussian variables. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz constant L when

$$\big|f(\boldsymbol{a}) - f(\boldsymbol{b})\big| \le L\,\|\boldsymbol{a} - \boldsymbol{b}\| \quad \text{for all } \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n.$$

We also say, more briefly, that f is L-Lipschitz. The first result [Bog98, Thm. 1.6.4] gives a bound on the variance of a Lipschitz function. The second result [Bog98, Thm. 1.7.6] provides a normal concentration inequality for Lipschitz functions.

Fact A.1 (Gaussian Variance Inequality)**.** *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz constant L. Let $\boldsymbol{\gamma} \in \mathbb{R}^n$ be a standard normal random vector. Then*

$$\operatorname{Var}[f(\boldsymbol{\gamma})] \le L.$$

Equivalently,

$$\mathbb{E}\, f(\boldsymbol{\gamma})^2 \le \big(\mathbb{E}\, f(\boldsymbol{\gamma})\big)^2 + L.$$

Fact A.2 (Gaussian Concentration Inequality)**.** *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz constant L. Let $\boldsymbol{\gamma} \in \mathbb{R}^n$ be a standard normal random vector. Then, for all $\zeta \ge 0$,*

$$\mathbb{P}\big\{f(\boldsymbol{\gamma}) \ge \mathbb{E}\, f(\boldsymbol{\gamma}) + \zeta\big\} \le \mathrm{e}^{-\zeta^2/2}, \quad \text{and}$$
$$\mathbb{P}\big\{f(\boldsymbol{\gamma}) \le \mathbb{E}\, f(\boldsymbol{\gamma}) - \zeta\big\} \le \mathrm{e}^{-\zeta^2/2}.$$

A.2. The Gaussian Minimax Theorem.

To compute the expectations of certain functions of Gaussian random variables, we depend on a comparison principle due to Gordon [Gor85, Thm. 1.1].

Let S be an abstract set. A family $\{Z_s : s \in S\}$ of real random variables is called a *centered Gaussian process* when each element Z_s has mean zero and each finite subcollection $\{Z_{s_1}, \dots, Z_{s_n}\}$ has a jointly Gaussian distribution.

Fact A.3 (Gaussian Minimax Theorem)**.** *Let T and U be finite sets. Consider two centered Gaussian processes $\{X_{tu}\}$ and $\{Y_{tu}\}$, indexed over $T \times U$. For all choices of indices, suppose that*

$$\begin{cases} \mathbb{E}\, X_{tu}^2 = \mathbb{E}\, Y_{tu}^2 \\ \mathbb{E}\, X_{tu} X_{tu'} \le \mathbb{E}\, Y_{tu} Y_{tu'} \\ \mathbb{E}\, X_{tu} X_{t'u'} \ge \mathbb{E}\, Y_{tu} Y_{t'u'} \quad \text{when } t \ne t'. \end{cases}$$

Then, for all real numbers λ_{tu} and ζ,

$$\mathbb{P}\left\{\min_{t\in T} \max_{u\in U} \big(\lambda_{tu} + X_{tu}\big) \ge \zeta\right\} \ge \mathbb{P}\left\{\min_{t\in T} \max_{u\in U} \big(\lambda_{tu} + Y_{tu}\big) \ge \zeta\right\}.$$

Fact A.3 extends to infinite index sets T, U by approximation.

APPENDIX B. SPECTRAL BOUNDS FOR RANDOM MATRICES

Our argument also depends heavily on some non-asymptotic bounds for the spectrum of a random matrix with independent entries. These results only give rough estimates, but they are adequate for our purposes.

The first result gives tail bounds for the extreme singular values of a rectangular matrix with independent, subgaussian entries.

Fact B.1 (Subgaussian Matrix: Tail Bounds)**.** *Let $\boldsymbol{X}$ be an $d_1 \times d_2$ random matrix with independent, standardized entries that are uniformly subgaussian:*

$$\sup_{p \geq 1} \frac{1}{\sqrt{p}} \left(\mathbb{E}\,|X_{ij}|^p\right)^{1/p} \leq B. \tag{B.1}$$

Then the largest singular value $\sigma_{\max}(\boldsymbol{X})$ and the d_2-th largest singular value $\sigma_{\min}(\boldsymbol{X})$ satisfy the bounds

$$\mathbb{P}\left\{\sigma_{\max}(\boldsymbol{X}) > \sqrt{d_1} + \mathrm{C}B^2\sqrt{d_2} + \mathrm{C}B^2\zeta\right\} \leq \mathrm{e}^{-\zeta^2}$$
$$\mathbb{P}\left\{\sigma_{\min}(\boldsymbol{X}) < \sqrt{d_1} - \mathrm{C}B^2\sqrt{d_2} - \mathrm{C}B^2\zeta\right\} \leq \mathrm{e}^{-\zeta^2}.$$

This result follows from [Ver12, Thm. 5.39] when we track the role of the subgaussian constant through the proof.

The second result gives a tail bound for the norm of a matrix with independent entries that may only have two moments; it is based on the matrix Rosenthal inequality [Tro15c, Thm. 1.1] and a standard concentration inequality [BLM13, Thm. 15.5].

Fact B.2 (Heavy-Tailed Matrix: Norm Bound)**.** *Fix a parameter $p \in [2, \log(d_1 + d_2)]$. Let $\boldsymbol{X}$ be a $d_1 \times d_2$ random matrix with independent entries that have the following properties.*

- *The entries are centered:* $\mathbb{E}\,X_{ij} = 0$.
- *The variances of the entries are uniformly bounded:* $\mathrm{Var}(X_{ij}) \leq \alpha$.
- *The entries have uniformly bounded pth moments:* $\mathbb{E}\,|X_{ij}|^p \leq \nu^p$.

Then

$$\mathbb{P}\left\{\|\boldsymbol{X}\| \geq \mathrm{C}\sqrt{\alpha(d_1+d_2)\log(d_1+d_2)} + \left(\mathrm{C}\nu(d_1+d_2)^{2/p}\log(d_1+d_2)\right)\zeta\right\} \leq \zeta^{-p}.$$

Proof Sketch. Write the random matrix as a sum of independent random matrices:

$$\boldsymbol{X} = \sum\nolimits_{i=1}^{d_1}\sum\nolimits_{j=1}^{d_2} X_{ij}\mathbf{E}_{ij},$$

where $\mathbf{E}_{ij}$ is the $d_1 \times d_2$ matrix with a one in the (i,j) position and zeros elsewhere. A straightforward application of the matrix Rosenthal inequality [Tro15c, Thm. I] yields

$$\begin{aligned}
\mathbb{E}\,\|\boldsymbol{X}\| &\leq \mathrm{C}\sqrt{\alpha(d_1+d_2)\log(d_1+d_2)} + \mathrm{C}\left(\mathbb{E}\max\nolimits_{ij}|X_{ij}|^p\right)^{1/p}\log(d_1+d_2) \\
&\leq \mathrm{C}\sqrt{\alpha(d_1+d_2)\log(d_1+d_2)} + \mathrm{C}\nu(d_1d_2)^{1/p}\log(d_1+d_2) \\
&\leq \mathrm{C}\sqrt{\alpha(d_1+d_2)\log(d_1+d_2)} + \mathrm{C}\nu(d_1+d_2)^{2/p}\log(d_1+d_2).
\end{aligned}$$

The second line follows when we replace the maximum by a sum and exploit the uniform moment estimate. The third line is just the inequality between the geometric and arithmetic means.

A standard concentration inequality for moments [BLM13, Thm. 15.5] gives

$$\left[\mathbb{E}(\|\boldsymbol{X}\| - \mathbb{E}\,\|\boldsymbol{X}\|)_+^p\right]^{1/p} \leq \mathrm{C}\sqrt{p}\left(\mathbb{E}\,V_+^{p/2}\right)^{1/p}$$

In this expression, the variance parameter

$$V_+ := \sum\nolimits_{ij}\mathbb{E}\left[\left(\|\boldsymbol{X}\| - \|\boldsymbol{X}^{(ij)}\|\right)_+^2 \,|\, \boldsymbol{X}\right].$$

The (i, j) entry of $\boldsymbol{X}^{(ij)}$ is an independent copy of the corresponding entry of $\boldsymbol{X}$; the remaining entries of the two matrices are the same. Applying the usual method [BLM13, Ex. 3.14], we see that

$$V_+ \le \mathrm{C} \max_{ij} \mathbb{E}\left[\left(X_{ij} - X'_{ij}\right)^2 \mid \boldsymbol{X}\right].$$

Applying the same considerations as in the last paragraph, we obtain

$$\left(\mathbb{E}\, V_+^{p/2}\right)^{1/p} \le \mathrm{C}\nu(d_1 + d_2)^{2/p}.$$

Combine these results and apply Markov's inequality to obtain the tail bound

$$\mathbb{P}\left\{\|\boldsymbol{X}\| \ge \mathbb{E}\,\|\boldsymbol{X}\| + \mathrm{C}\nu\sqrt{p}(d_1 + d_2)^{2/p}\zeta\right\} \le \zeta^{-p}.$$

Introduce the estimate for the expected norm to complete the argument. □

ACKNOWLEDGMENTS

The authors would like to thank David Donoho, Surya Ganguli, Babak Hassibi, Michael McCoy, Andrea Montanari, Ivan Nourdin, Giovanni Peccati, Adrian Röllin, Jared Tanner, Christos Thrampoulidis, and Madeleine Udell for helpful conversations. We also thank the anonymous reviewers and the editors for their careful reading and suggestions. SO was generously supported by the Simons Institute for the Theory of Computing and NSF award CCF-1217058. JAT gratefully acknowledges support from ONR award N00014-11-1002 and the Gordon & Betty Moore Foundation.

REFERENCES

[ALMT14] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Inf. Inference*, 3(3):224–294, 2014.

[BDN15] J. Bourgain, S. Dirksen, and J. Nelson. Toward a unified theory of sparse dimensionality reduction in Euclidean space. *Geom. Funct. Anal.*, 25(4):1009–1088, 2015.

[BGVV14] S. Brazitikos, A. Giannopoulos, P. Valettas, and B.-H. Vritsiou. *Geometry of isotropic convex bodies*, volume 196 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2014.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[BLM15] M. Bayati, M. Lelarge, and A. Montanari. Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.*, 25(2):753–822, 2015.

[BM12] M. Bayati and A. Montanari. The LASSO risk for Gaussian matrices. *IEEE Trans. Inform. Theory*, 58(4):1997–2017, 2012.

[Bog98] V. I. Bogachev. *Gaussian measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1998.

[BS10] Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, second edition, 2010.

[BY93] Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.*, 21(3):1275–1294, 1993.

[CDS98] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[Cha06] S. Chatterjee. A generalization of the Lindeberg principle. *Ann. Probab.*, 34(6):2061–2076, 2006.

[CRPW12] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

[CRT06] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[CRTV05] E. Candès, M. Rudelson, T. Tao, and R. Vershynin. Error correction via linear programming. In *Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on*, pages 668–681, Oct 2005.

[CSPW11] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.*, 21(2):572–596, 2011.

[CW13] K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. In *STOC'13—Proceedings of the 2013 ACM Symposium on Theory of Computing*, pages 81–90. ACM, New York, 2013.

[DDW+07] M. A. Davenport, M. F. Duarte, M. B. Wakin, J. N. Laska, D. Takhar, K. F. Kelly, and R. G. Baraniuk. The smashed filter for compressive classification and target recognition. In *Proc. SPIE*, volume 6498, pages 64980H–64980H–12, 2007.

[DGM13] D. L. Donoho, M. Gavish, and A. Montanari. The phase transition of matrix recovery from Gaussian measurements matches the minimax MSE of matrix denoising. *Proc. Natl. Acad. Sci. USA*, 110(21):8405–8410, 2013.

[DH01] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory*, 47(7):2845–2862, 2001.

[DJM13] D. L. Donoho, I. Johnstone, and A. Montanari. Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE Trans. Inform. Theory*, 59(6):3396–3433, 2013.

[Don06a] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[Don06b] D. L. Donoho. For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59(6):797–829, 2006.

[Don06c] D. L. Donoho. High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete Comput. Geom.*, 35(4):617–652, 2006.

[DS01] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*, pages 317–366. North-Holland, Amsterdam, 2001.

[DT06] D. Donoho and J. Tanner. Thresholds for the recovery of sparse solutions via l1 minimization. In *Information Sciences and Systems, 2006 40th Annual Conference on*, pages 202–206, March 2006.

[DT09a] D. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 367(1906):4273–4293, 2009. With electronic supplementary materials available online.

[DT09b] D. L. Donoho and J. Tanner. Counting faces of randomly projected polytopes when the projection radically lowers dimension. *J. Amer. Math. Soc.*, 22(1):1–53, 2009.

[DT10] D. L. Donoho and J. Tanner. Counting the faces of randomly-projected hypercubes and orthants, with applications. *Discrete Comput. Geom.*, 43(3):522–541, 2010.

[EK12] Y. C. Eldar and G. Kutyniok, editors. *Compressed sensing*. Cambridge University Press, Cambridge, 2012. Theory and applications.

[Faz02] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford, 2002.

[FM14] R. Foygel and L. Mackey. Corrupted sensing: novel guarantees for separating structured signals. *IEEE Trans. Inform. Theory*, 60(2):1223–1247, 2014.

[FR13] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.

[GG15] P. Gao and S. Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Curr. Opinion Neurobiology*, 32:148–155, 2015.

[GNP14] L. Goldstein, I. Nourdin, and G. Peccati. Gaussian phase transitions and conic intrinsic volumes: Steining the Steiner formula. Available at http://arXiv.org/abs/1411.6265, Nov. 2014.

[Gor85] Y. Gordon. Some inequalities for Gaussian processes and applications. *Israel J. Math.*, 50(4):265–289, 1985.

[Gor88] Y. Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbf{R}^n$. In *Geometric aspects of functional analysis (1986/87)*, volume 1317 of *Lecture Notes in Math.*, pages 84–106. Springer, Berlin, 1988.

[Gru07] P. M. Gruber. *Convex and discrete geometry*, volume 336 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Berlin, 2007.

[HMT11] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, 2011.

[JM14] A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional regression under the Gaussian random design model: asymptotic theory. *IEEE Trans. Inform. Theory*, 60(10):6522–6554, 2014.

[Kle55] V. L. Klee, Jr. Separation properties of convex cones. *Proc. Amer. Math. Soc.*, 6:313–318, 1955.

[KM11] S. B. Korada and A. Montanari. Applications of the Lindeberg principle in communications and statistical learning. *IEEE Trans. Inform. Theory*, 57(4):2440–2450, 2011.

[KN14] D. M. Kane and J. Nelson. Sparser Johnson-Lindenstrauss transforms. *J. ACM*, 61(1):Art. 4, 23, 2014.

[KY14] A. Knowles and J. Yin. Anisotropic local laws for random matrices. Available at http://arXiv.org/abs/1410.3516, Nov. 2014.

[Lin22] J. W. Lindeberg. Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Math. Z.*, 15(1):211–225, 1922.

[LO94] R. Latała and K. Oleszkiewicz. On the best constant in the Khinchin-Kahane inequality. *Studia Math.*, 109(1):101–104, 1994.

[LT11] M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition.

[Mah11] M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

[Mas00] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, 28(2):863–884, 2000.

[Men10] S. Mendelson. Empirical processes with a bounded ψ_1 diameter. *Geom. Funct. Anal.*, 20(4):988–1027, 2010.

[Men14] S. Mendelson. A remark on the diameter of random sections of convex bodies. In B. Klartag and E. Milman, editors, *Geometric Aspects of Functional Analysis*, volume 2116 of *LNM*, pages 395–404. Springer, 2014.

[MOO10] E. Mossel, R. O'Donnell, and K. Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. *Ann. of Math. (2)*, 171(1):295–341, 2010.

[MPTJ07] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, 17(4):1248–1282, 2007.

[MT13] M. B. McCoy and J. A. Tropp. The achievable performance of convex demixing. Available at http://arXiv.org/abs/arXiv:1309.7478, Sep. 2013.

[MT14a] M. B. McCoy and J. A. Tropp. From Steiner formulas for cones to concentration of intrinsic volumes. *Discrete Comput. Geom.*, 51(4):926–963, 2014.

[MT14b] M. B. McCoy and J. A. Tropp. Sharp recovery bounds for convex demixing, with applications. *Found. Comput. Math.*, 14(3):503–567, 2014.

[NN13] J. Nelson and H. L. Nguyen. OSNAP: faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science—FOCS 2013*, pages 117–126. IEEE Computer Soc., Los Alamitos, CA, 2013.

[OH10] S. Oymak and B. Hassibi. New null space results and recovery thresholds for matrix rank minimization. Available at http://arXiv.org/abs/1011.6326, Nov. 2010.

[OH13] S. Oymak and B. Hassibi. Asymptotically exact denoising in relation to compressed sensing. Available at http://arXiv.org/abs/1305.2714, May 2013.

[OTH13] S. Oymak, C. Thrampoulidis, and B. Hassibi. The squared-error of generalized lasso: A precise analysis. In *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, pages 1002–1009, Oct 2013. Available at http://arXiv.org/abs/1311.0830.

[PW15] M. Pilanci and M. Wainwright. Randomized sketches of convex programs with sharp guarantees. *Information Theory, IEEE Transactions on*, 61(9):5096–5115, Sept 2015.

[Rot73] V. I. Rotar'. Certain limit theorems for polynomials of degree two. *Teor. Verojatnost. i Primenen.*, 18:527–534, 1973.

[RV08] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, 2008.

[San52] L. A. Santaló. Integral geometry in spaces of constant curvature. *Repub. Argentina. Publ. Comision Nac. Energia Atomica. Ser. Mat.*, 1(1):68, 1952.

[San76] L. A. Santaló. *Integral geometry and geometric probability*. Addison-Wesley Publishing Co., Reading, Mass.-London-Amsterdam, 1976. With a foreword by Mark Kac, Encyclopedia of Mathematics and its Applications, Vol. 1.

[Sar06] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS '06. 47th Annual IEEE Symposium on*, pages 143–152, Oct 2006.

[Sch50] L. Schläfli. *Gesammelte mathematische Abhandlungen. Band I*. Verlag Birkhäuser, Basel, 1950.

[Sio58] M. Sion. On general minimax theorems. *Pacific J. Math.*, 8:171–176, 1958.

[Sto09] M. Stojnic. Various thresholds for ℓ_1-optimization in compressed sensing. Available at http://arXiv.org/abs/0907.3666, July 2009.

[Sto13] M. Stojnic. Regularly random duality. Available at http://arXiv.org/abs/1303.7295, Mar. 2013.

[SW08] R. Schneider and W. Weil. *Stochastic and integral geometry*. Probability and its Applications (New York). Springer-Verlag, Berlin, 2008.

[TAH15] C. Thrampoulidis, E. Abbasi, and B. Hassibi. High-dimensional error anlaysis of regularized m-estimators. Forthcoming, Nov. 2015.

[Tao12] T. Tao. *Topics in random matrix theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[TH15] C. Thrampoulidis and B. Hassibi. Isotropically random orthogonal matrices: Performance of lasso and minimum conic singular values. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 556–560, June 2015. Available at http://arXiv.org/abs/1503.07236.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[Tik15] K. Tikhomirov. The limit of the smallest singular value of random matrices with i.i.d. entries. *Adv. Math.*, 284:1–20, 2015.

[TOH15] C. Thrampoulidis, S. Oymak, and B. Hassibi. The Gaussian min–max theorem in the presence of convexity. In *Proceedings of The 28th Conference on Learning Theory (COLT)*, Jul. 2015. Available at http://arXiv.org/abs/1408.4837.

[Tro59] H. F. Trotter. An elementary proof of the central limit theorem. *Arch. Math.*, 10:226–234, 1959.

[Tro06] J. A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory*, 52(3):1030–1051, 2006.

[Tro15a] J. A. Tropp. Code for reproducing figures from Oymak & Tropp, *Universality Laws for Randomized Dimension Reduction, with Applications*, 2015.ropp, *Universality Laws for Randomized Dimension Reduction, with Applications*, 2015. Available at http://users.cms.caltech.edu/~jtropp, Nov. 2015.

[Tro15b] J. A. Tropp. Convex recovery of a structured signal from independent random measurements. In G. Pfander, editor, *Sampling Theory: A Rennaissance*. Birkhäuser Verlag, 2015. Available at http://arXiv.org/abs/1405.1102.

[Tro15c] J. A. Tropp. The expected norm of a sum of independent random matrices: An elementary approach. In *High-Dimensional Probability VII*, Cargèse, June 2015. To appear. Available at http://arXiv.org/abs/1506.04711.

[TV15] T. Tao and V. Vu. Random matrices: universality of local spectral statistics of non-Hermitian matrices. *Ann. Probab.*, 43(2):782–874, 2015.

[Ver12] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

[Ver15] R. Vershynin. Estimation in high dimensions: A geometric perspective. In G. Pfander, editor, *Sampling Theory: A Rennaissance*. Birkhäuser Verlag, 2015.

[Wen62] J. G. Wendel. A problem in geometric probability. *Math. Scand.*, 11:109–111, 1962.

[Woo14] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[Yin86] Y. Q. Yin. Limiting spectral distribution for a class of random matrices. *J. Multivariate Anal.*, 20(1):50–68, 1986.