Tail Bounds on the Smallest Singular Value of a Random Rectangular Matrix

Shyam Narayanan

1 Introduction

There has been significant research into the singular values of random rectangular matrices. Perhaps the most famous result in this area is the Marchenko-Pastur law, which characterizes the distribution of singular values of an $N \times \lambda N$ matrix (for some $0 < \lambda < 1$), where each entry is drawn i.i.d. from some variance-1 distribution. The Marchenko-Pastur law gives us a simple and explicit distribution for the set of all singular values as $N \to \infty$, after an appropriate rescaling. Moreover, the Marchenko-Pastur distribution also predicts the largest and smallest singular values $(\sqrt{N} \cdot (1 \pm \sqrt{\lambda}))$, in the $N \to \infty$ limit, though this was formally proven later by [6]. (We remark that Marchenko-Pastur does not immediately imply anything about the top and bottom singular values, since a single very large or very small singular value would not affect the law as $N \to \infty$.) One can ask for more fine-grained guarantees on the largest or smallest singular value. More recently, it was proven that both the largest and smallest singular values follow a Tracy-Widom law around $\sqrt{N} \cdot (1 \pm \sqrt{\lambda})$ [1].

However, these laws do not characterize extreme tail events, such as what is the probability that the smallest singular value being less than some $c < (1 - \sqrt{\lambda}) \cdot \sqrt{N}$? For instance, can you show that the probability of the smallest singular value being $\frac{(1-\sqrt{\lambda}) \cdot \sqrt{N}}{2}$ decays exponentially with N? More generally, one can consider general $N \times n$ matrices, where perhaps the ratio n/N isn't a fixed constant $\lambda < 1$. In this case, the matrix could be square or close to square, which could potentially make it more difficult to bound the smallest eigenvalue with very high probability.

Main Theorem: In a work of Rudelson and Vershynin [5], the authors show a very strong tail bound on the smallest singular value of a random rectangular matrix, as long as the entries of the rectangular matrix are i.i.d. drawn from any subgaussian distribution. Specifically, they prove the following theorem.

Theorem 1.1. *Let B be an $N \times n$ random matrix ($N \geq n$), where each element of B is an i.i.d. copy of a mean zero subgaussian random variable with unit variance. Let $s_n(B)$ represent the smallest singular value of B. Then, for every $\varepsilon > 0$, we have*

$$\mathbb{P}\left(s_n(A) \leq \varepsilon \cdot (\sqrt{N} - \sqrt{n-1})\right) \leq (C\varepsilon)^{N-n+1} + e^{-c \cdot N} \tag{1}$$

for some appropriate constants C, c.

Theorem 1.1 implies that for tall rectangular matrices (for instance, $N \geq 1.01n$), the smallest singular value is at least a constant fraction of what is expected with exponential failure probability. Even for more squarish matrices (for instance, $N = n + \sqrt{n}$), we still get that the smallest

singular value is within a constant factor of what is expected with $e^{-\sqrt{n}}$ failure probability, and this probability decays for smaller ε.

We remark that the bound of Equation (1) was well-known when $N \geq Kn$ for a sufficiently large constant K [2], and was later proved for $N \geq (1+\delta)n$ for any fixed constant δ [3] (in the latter case, the constants C, c may have depended on n). So, the main contribution of [5] was to prove this bound holds for all $N \geq n$, not just for N a reasonable factor larger than n.

For this project, the main goal I had was to understand the technical proof of Theorem 1.1 and to provide a simpler summary of the main ideas. In Sections 3 and 4, I provide a proof sketch of each of the pieces required to prove Theorem 1.1, and how the pieces combine together.

Experiments: The second goal I had was to investigate empirically whether Theorem 1.1 holds in practice, even for reasonably small values of N and n (so that enough matrices can be simulated to observe the tail bounds). Moreover, I was also interested in knowing whether changing the distributions on the entries of the random matrix affects the smallest singular value. For instance, are the tail bounds on the smallest singular value different, if the entries are i.i.d. drawn from $\mathcal{N}(0,1)$, versus if each entry is random from ± 1?

In addition, perhaps these strong concentration bounds on smallest singular value still hold even if the entries are *not* drawn i.i.d. from a subgaussian distribution. Hence, a natural question is whether we can still empirically obtain good bounds even the entries of the matrix are heavy-tailed. Theoretically, it is known that the smallest singular value of an $N \times \lambda N$ dimensional matrix converges to $\sqrt{N}(1-\sqrt{\lambda})$ with probability 1 as $N \to \infty$, as long as the entries have bounded variance [7]. However, whether we can get results similar to Theorem 1.1 if the entries only have bounded variance, as far as I know, is unknown.

In Section 5, we investigate this question empirically, for various distributions on the entries, and see how the tail of the smallest singular value distribution changes.

2 Preliminaries

2.1 Notation

For a vector v, we use $\|v\|_2$ to denote the Euclidean norm. For a matrix M, we use $\|M\|_{op}$ to denote the operator norm of M.

We use $\text{dist}(x,y)$ to denote the Euclidean distance $\|x-y\|_2$. For a subset S of Euclidean space, we will write $\text{dist}(x,S) = \min_{y \in S} \text{dist}(x,y)$.

We will use the expression $\ll$ to denote "less than up to a tiny constant", i.e., $m \ll n$ if there is some tiny constant c such that $m \leq cn$. We will also use standard Big-O notation, such as $O(\cdot), \Omega(\cdot), o(\cdot)$, and $\Theta(\cdot)$.

Assumptions on N and n: Due to the work of [2] and [3], Rudelson and Vershynin assume that $N - n \ll n$ in the proof. (We also make this assumption in Sections 3 and 4.) Otherwise, the exponentially small failure probability bound has already been proved.

Constants: We use C or K to denote large constants and c or κ to denote a small constant. (We may also use $C_1, C_2, \ldots$ or $c_1, c_2, \ldots$ similarly.) For simplicity, I am lazy with the notation of

constants and may reuse C or c in multiple places (in these cases, it only matters that there exists some constant C or c).

2.2 Definitions

We first recall the definition of subgaussisan random variables.

Definition 2.1. A *subgaussian* random variable is a univariate random variable ξ such that for some fixed constant $L > 0$, $\mathbb{P}(|\xi| > t) \leq 2e^{-t^2/L^2}$ for all $t > 0$. The parameter L is often called the *subgaussian* moment.

When dealing with subgaussian variables, we may implicitly assume the variable has mean 0 and variance 1.

We remark the following basic property of sums of subgaussian variables (e.g., Fact 2.1 in [3]).

Lemma 2.1. *Suppose that $\xi_1, \ldots, \xi_n$ are subgaussian random variables (each with subgaussian moment at most L). Then, for any fixed $z_1, \ldots, z_n$ with $\sum z_i^2 = 1$, the random variable $\zeta = \sum_{i=1}^n z_i \cdot \xi_i$ is also subgaussian with subgaussian moment $O(L)$.*

Next, we define a property called *compressibility* of a vector. This property in a sense tells us how evenly spread out the vector is among its coordinates.

Definition 2.2. Fix some constants $0 < \delta, \rho < 1$. A vector $x \in \mathbb{R}^n$ is δ-*sparse* if the number of nonzero entries in x is at most $\delta \cdot n$. A unit vector $x \in \mathbb{R}^n$ is (δ, ρ)-*compressible* if x is within distance ρ of some δ-sparse vector. Finally, a unit vector $x \in \mathbb{R}^n$ is (δ, ρ)-*incompressible* if it is *not* (δ, ρ)-compressible.

The following simple fact is known about incompressible vectors (e.g., Lemma 3.4 in [4]).

Lemma 2.2 (Informal)**.** *Let $x \in \mathbb{R}^n$ be (δ, ρ)-incompressible. Then, there exist $\Omega(n)$ coordinates i such that $|x_i|$ is $\Theta(1/\sqrt{n})$.*

2.3 Starting point: ε-nets and the largest singular value

Before understanding how the proof in [5] works, it is useful to understand the method of using ε-nets and how we can bound the largest singular value of a random matrix. This is also useful in bounding the smallest singular value of sufficiently tall matrices with exponentialy small failure probability.

The intuition is as follows. Let $A \in \mathbb{R}^{n \times n}$ be a random matrix where each entry is i.i.d. drawn from some subgaussian distribution (with mean 0 and variance 1). The maximum singular value of A equals $\max v^\top A w$ over unit vectors v, w. So, it suffices to bound $v^\top A w$ for every v, w.

For any fixed unit vectors v, w, we can write $v^\top A w = \sum A_{ij} v_i w_j$. Since $\sum_{i,j} v_i^2 w_j^2 = (\sum_i v_i^2) \cdot (\sum_j w_j^2) = 1$, we can apply Lemma 2.1 to say that $v^\top A w$ is also subgaussian. Thus, $\mathbb{P}(v^\top A w \geq t) \leq 2e^{-\Omega(t^2)}$.

This gets us really good bounds for a single choice of v and w, but the problem is that there are an infinite number of choices for v and w. To avoid this issue of having to use a union bound over an infinite number of v and w, the main tool we need is what is called an ε-net, which we now define.

Definition 2.3. Given a subset $S \subset \mathbb{R}^n$, a set of points $\mathcal{N} \subset S$ is an *ε-net* of S if for every point $x \in S$, there is some $y \in \mathcal{N}$ such that $\|x - y\|_2 \leq \varepsilon$.

The main tool is to use an ε-net $\mathcal{N}$ over the n-dimensional sphere, so that we can bound $v^\top Aw$ as "close" to $(v')^\top A(w')$ for some $v', w' \in \mathcal{N}$. Intuitively, each point in the net should cover a radius ε ball in the sphere, and thus cover an ε^n fraction of the volume of the sphere. So, a well-constructed net should have roughly $(1/\varepsilon)^n$ points. Indeed, the following is true.

Lemma 2.3 (Folklore, or see Proposition 2.1 in [5])**.** *There exists an ε-net of the unit sphere S^{n-1} in $\mathbb{R}^n$, of size $2n \cdot \left(1 + \frac{2}{\varepsilon}\right)^{n-1}$.*

Now, let $\mathcal{N}$ be a 0.1-net of the unit sphere, of size at most 30^n (which exists by Lemma 2.3). For any $v, w \in \mathcal{N}$, we have already established that $\mathbb{P}(v^\top Aw \geq t) \leq 2e^{-t^2}$[1]. Therefore, we have that the probability of there even existing $v, w \in \mathcal{N}$ such that $\mathbb{P}(v^\top Aw \geq t)$ is at most $O(30^{2n} \cdot e^{-t^2})$, since each of v, w have 30^n options. So, for $t = 5\sqrt{n}$, this is at most e^{-n}, meaning the probability of there even existing $v, w \in \mathcal{N}$ such that $\mathbb{P}(v^\top Aw \geq 5\sqrt{n})$ is at most e^{-n}.

How can we use the net property to generalize this to arbitrary v, w, rather than just $v, w \in \mathcal{N}$? Assume that $(v')^\top A(w') < 5\sqrt{n}$ for all $v', w' \in \mathcal{N}$. The trick is to realize that for any v, w in the unit sphere S^{n-1}, we can write $\|v - v'\|_2 \leq 0.1$ for some $v' \in \mathcal{N}$, so $v = v' + a \cdot y$ for $a = \|v - v'\|_2 \leq 0.1$ and $y = \frac{v - v'}{\|v - v'\|_2}$ is on the unit sphere. Similarly, we can write $w = w' + b \cdot z$ for $b \leq 0.1$ and z on the unit sphere.

Thus,

$$v^\top Aw = (v' + ay)^\top A(w' + bz) = (v')^\top Aw' + a \cdot y^\top Aw + b \cdot (v')^\top Az + ab \cdot y^\top Az.$$

We know that $(v')^\top Aw' \leq 5\sqrt{n}$. Moreover, we know that $y^\top Aw, (v')^\top Az, y^\top Az$ are all at most the operator norm $\|A\|_{op}$. Therefore, we have that for *any* unit vectors v, w, $v^\top Aw \leq 10\sqrt{n} + (a + b + ab) \cdot \|A\|_{op} = 5\sqrt{n} + 0.21 \cdot \|A\|_{op}$. By taking the supremum over all v, w, we obtain $\|A\|_{op} \leq 5\sqrt{n} + 0.21 \cdot \|A\|_{op}$, which means that $\|A\|_{op} \leq 7\sqrt{n}$. This holds as long as the bound holds for all $v', w' \in \mathcal{N}$, so the failure probability is *exponentially small* in n.

Implications: We remark that this bound also has implications for the smallest singular value. Specifically, it can imply that for an $N \times n$ matrix B, if N is much larger than n (say $N \geq 30n$), the smallest singular value must be $\Omega(\sqrt{N})$ with exponential failure probability. The idea is that the smallest singular value of B is $\min \|Bv\|_2$. If we again write $v = v' + ay$, where $v' \in \mathcal{N}$, $a \leq 0.1$ and y is a unit vector. We can write $Bv = Bv' + aBy$, so by the Triangle inequality, $\|Bv\|_2 \geq \|Bv'\|_2 - a \cdot \|By\|_2 \geq \|Bv'\|_2 - 0.1 \cdot \|B\|_{op}$. Note that v' lies in $\mathbb{R}^n$ (so the net has size at most 30^n), and it turns out that a simple union bound and some Chernoff-type inequalities will establish that $\|Bv'\|_2 \in [0.9\sqrt{N}, 1.1\sqrt{N}]$ for all v' in the net. (This will crucially assume that N is much larger than n.) Thus, assuming that $\|B\|_{op} \leq 7\sqrt{N}$ (which will hold based on our arguments about the largest singular value), we get that for all unit vectors v, $\|Bv\|_2 \geq 0.9\sqrt{N} - 0.1 \cdot 7\sqrt{N} \geq 0.2\sqrt{N}$. Thus, the smallest singular value is at least $\Omega(\sqrt{N})$, and this holds with exponentially small failure probability.

We remark that this type of observation can also be used to bound $\|Bx\|_2$ for *compressible* vectors x, since almost all of the mass of x is supported on a small fraction of coordinates. As a result, the following can be shown to hold (see Lemma 3.3 in [4] or Lemma 2.6 in [5]).

[1] It will really be $2e^{-\Omega(t^2)}$, but for simplicity we will remove this Ω factor.

Lemma 2.4. *Let B be an $N \times n$ random matrix ($N \geq n/2$), whose elements are independent copies of a subgaussian random variable. Then, there exist constants δ, ρ, c such that*

$$\mathbb{P}\left(\inf_{x \text{ is } (\delta,\rho)-compressible} \|Bx\|_2 \leq c\sqrt{N} \right) \leq e^{-cN}.$$

3 Bounding the Least Singular Value

In this section, we provide an overview for the proof of Theorem 1.1, modulo an important technical lemma (Theorem 3.1) that we will defer the proof of to Section 4. Recall that our goal is to bound the smallest singular value of the matrix $B \in \mathbb{R}^{N \times n}$, which has each entry i.i.d. drawn from a subgaussian distribution. Equivalently, we wish to provide a uniform lower bound on $\|Bv\|_2$ for all unit vectors $v \in \mathbb{R}^n$. We will think of $X_i \in \mathbb{R}^N$ as the ith column of B.

The first step is to consider the case when v is (δ, ρ)-compressible for some small constants δ, ρ. In this case, Lemma 2.4 already solves this case, showing that $\|Bv\|_2 > \Omega(\sqrt{N})$ with exponential failure probability in N. Hence, in the remainder of this section, we assume that v is an incompressible vector. So, our goal is to bound the minimum of $\|Bv\|_2$ over (δ, ρ)-incompressible vectors $v \in \mathbb{R}^n$.

3.1 Bound in terms of spread

Any ε-net over v (even for incompressible v) has size roughly $(1/\varepsilon)^n$, but we would like to deal with a smaller-sized net. The idea behind reducing the net size is that we can write $v = (v_1, v_2)$, where $v_1 \in \mathbb{R}^d$ and $v_2 \in \mathbb{R}^{n-d}$, for some parameter d that we will set later. Let's also think of $B = \begin{pmatrix} B_1 & B_2 \end{pmatrix}$, where B_1 is the matrix of the first d columns, and B_2 is the rest. Now, if we fix v_1, the minimum over v_2 of $\|Bv\|_2 = \|B_1 v_1 + B_2 v_2\|_2$ is simply the distance from $B_1 v_1$ to the subspace spanned by the columns of B_2. The hope is that, if we set d properly, we can union bound over a net for v_1 (which should have size $(1/\varepsilon)^d$ instead of $(1/\varepsilon)^n$), and that the distance from $B_1 v_1$ to the subspace spanned by B_2 should be small.

Another useful observation is that any incompressible vector $v \in \mathbb{R}^n$ should be thought of as a vector that has most coordinates v_i roughly $\frac{1}{\sqrt{n}}$ in absolute value. If every $v \in v_1$ (for v_1 the first d entries of v) has this property, then $\|v_1\|_2 = \Theta(\sqrt{d/n})$, which is intuitively the correct scaling for v_1. If $\|v_1\|_2$ were much smaller, then the distance from $B_1 v_1$ to the subspace spanned by B_2 may be too small, but we can hopefully avoid this using the fact that v is incompressible.

To make this formal, Rudelson and Vershynin define, for any subset $J \subset [n]$,

$$H_J := \text{span}(X_k)_{k \in J},$$

where we recall that X_k is the kth column of B. Moreover, for a subset $J \subset [n]$ of size d, they define $S(\mathbb{R}^J)$ to be the set of unit vectors in $\mathbb{R}^n$ entirely supported on the indices in J, and

$$\text{Spread}_J := \left\{ y \in S(\mathbb{R}^J) : \frac{\kappa}{\sqrt{d}} \leq |y_k| \leq \frac{K}{\sqrt{d}} \text{ for all } k \in J \right\},$$

where $0 < \kappa < 1 < K$ are some appropriate constants. Based on the intuition given above, one would like to say that for $J = \{1, 2, \ldots, d\}$, $\|Bv\|_2$ is at least the distance from Bv_1 to $H_{[n]\setminus J}$, where v_1 is in Spread_J, up to a scaling factor of roughly $\sqrt{d/n}$. This turns out to be a bit stronger than what we can prove, but they show that if J were chosen randomly, this claim holds with reasonable probability.

Lemma 3.1. *Fix constants $0 < \delta, \rho < 1$. Then, there are some constants $C_1, c_1 > 0$ such that for any $J \subset [n]$ of size d, and for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\inf_{v \text{ is } (\delta,\rho)-\text{incompressible}} \|Bv\|_2 \leq c_1 \varepsilon \cdot \frac{d}{\sqrt{n}}\right) \leq C_1^d \cdot \mathbb{P}\left(\inf_{z \in \text{Spread}_J} \text{dist}(Bz, H_{[n]\setminus J}) < \varepsilon\sqrt{d}\right).$$

Proof Sketch. First, note that the choice of J is not relevant for computing probabilities, since every row of B is chosen i.i.d. So, we will think of J as being a uniformly random subset of size d.

By Lemma 2.2, for any incompressible v, at least a constant fraction of the $|v_i|$'s are $\Theta(1/\sqrt{n})$. So, if we choose a random subset J of size d which is much smaller than n, the probability that every v_i for $i \in J$ is $\Theta(1/\sqrt{n})$ is at least some constant to the power of d.

In the setting where we picked such a J, we can think of z as v_J normalized to have unit norm, where v_J is the vector where we take v and zero out all coordinates of v that are not in J. This will scale v_J by a factor of $\Theta(\sqrt{n/d})$. If we ignore the scaling factor, $Bv = Bv_J + Bv_{[n]\setminus J}$, which means $\|Bv\|_2$ is at least the distance from Bv_J to the subspace $H_{[n]\setminus J}$. This is exactly what the right hand side is bounding, up to the $\sqrt{n/d}$ scaling factor. (This is why the right-hand side of the equation has $\varepsilon\sqrt{d}$ whereas the left hand side has $\varepsilon \cdot d/\sqrt{n}$.) The extra C_1^d factor on the right-hand side is coming from the fact that we only picked an appropriate J with some constant to the power of d probability. □

3.2 Bounding the spread

The final part requires bounding the actual spread. In other words, we wish to show that

$$\mathbb{P}\left(\inf_{z \in \text{Spread}_J} \text{dist}(Bz, H_{[n]\setminus J}) < \varepsilon\sqrt{d}\right)$$

is very small, for a random matrix $B \in \mathbb{R}^{N \times n}$.

The first step is to bound the quantity $\mathbb{P}(\text{dist}(Bz, H_{[n]\setminus J}) < \varepsilon\sqrt{d})$ for a fixed $z \in \text{Spread}_J$. To do this, we need the following crucial lemma.

Theorem 3.1 (Distance to a random subspace)**.** *Let ξ be a subgaussian distribution, and suppose $X \in \mathbb{R}^N$ has every entry i.i.d. drawn from ξ. Let H be a random subspace in $\mathbb{R}^N$ spanned by $N-m$ vectors $X_1, X_2, \ldots, X_{N-m}$, where $m \ll n$, and every coordinate of each X_i is i.i.d. drawn from some subgaussian distribution (possibly different from ξ).*

Then, for every $v \in \mathbb{R}^N$ and every $\varepsilon > 0$, we have that

$$\mathbb{P}\left(\text{dist}(X, H + v) < \varepsilon\sqrt{m}\right) \leq (C\varepsilon)^m + e^{-cN},$$

for some appropriate constants $C, c > 0$.

We remark that if the distribution ξ is in fact Gaussian, the above lemma is in fact quite simple to prove. Indeed, if $X \in \mathbb{R}^N$ were a standard Gaussian vector, then the distance from X to $H + v$ is just the distance from $\Pi_{H^\perp} X$ to $\Pi_{H^\perp} v$. Regardless of what $\Pi_{H^\perp}$ is, it has dimension $N - (N - m) = m$, so $\Pi_{H^\perp} X$ has a m-dimensional Gaussian over the subspace $H^\perp$. Finally, it is well-known that the PDF of a Gaussian is always at most $(1/\sqrt{2\pi})^m$, so the probability that $\Pi_{H^\perp} X$ lies in the ball of radius $\varepsilon\sqrt{m}$ around any $\Pi_{H^\perp} v$ is at most $(1/\sqrt{2\pi})^m$ times the volume of the ball, which is at most $(C\varepsilon)^m$ for some fixed C.

The difficult part is really in showing this holds (allowing an additional e^{-cN} error) for arbitrary subgaussian distributions. In Section 4, we focus on proving Theorem 3.1.

Why is Theorem 3.1 useful? Note that for any fixed unit vector z, if and every entry of B is i.i.d. drawn from some subgaussian distribution, it is well-known that this implies each entry of Bz is also subgaussian (and for fixed z, each entry is also independent). Moreover, $H_{[n]\setminus J}$ is just the span of the columns of B not in J, which means that since z is supported only on J, Bz and $H_{[n]\setminus J}$ are independent. Also, if $|J| = d$, then $H_{[n]\setminus J}$ is spanned by $n-d = N-(N-n+d)$ vectors. We will set $d = N - n + 1$, so this equals $N - (2d-1)$ vectors. Hence, Theorem 3.1 implies that $\mathbb{P}(\text{dist}(Bz, H_{[n]\setminus J}) \le \varepsilon\sqrt{d}) \le (C\varepsilon)^{2d-1} + e^{-cN}$ (since $\sqrt{d}$ and $\sqrt{2d-1}$ are the same up to a constant factor).

Finally, we need to union bound over all $z \in \text{Spread}_J$. While the size of an ε-net of unit vectors supported on J is at most $2d \cdot (1 + \frac{2}{\varepsilon})^{d-1} \le (3/\varepsilon)^{d-1}$ (so a union bound may seem straightforward), we need to make sure that when we change z by a distance ε, the distance between $\text{dist}(Bz, H_{[n]\setminus J})$ doesn't change by much. Indeed, Bz can change by up to $\|B\|_{op} \cdot \varepsilon$, so we have to be careful. We can actually be more careful with how much the distance between Bz and $H_{[n]\setminus J}$ changes if we move z by distance ε. This distance equals $P \cdot Bz$, where P is the projection onto the orthogonal complement of $H_{[n]\setminus J}$, which has dimension $N - (n-d) = 2d - 1$.

It will turn out that the matrix $P \cdot B$ behaves more like a random d-dimensional matrix (at least on z, which is supported on a d-dimensional subspace $\mathbb{R}^J$). Indeed, Rudelson and Vershynin show that the operator norm of $P \cdot B$ (restricted to act on the subspace $\mathbb{R}^J$) usually is bounded by $O(\sqrt{d})$, and in fact the probability that it exceeds some $t\sqrt{d}$ is at most $e^{-\Omega(t^2 d)}$. As a result, they can show the following.

Lemma 3.2. *Let W be the matrix $P \cdot B|_{\mathbb{R}^J}$, i.e., the matrix $P \cdot B$ restricted to act on $\mathbb{R}^J \subset \mathbb{R}^n$. Recall that $\|Wz\|_2 = \text{dist}(Bz, H_{[n]\setminus J})$. Then, for any constant C_1, for sufficiently small ε we have*

$$\mathbb{P}\left(\inf_{z \in \text{Spread}_J} \|Wz\|_2 < \varepsilon\sqrt{d}\right) \le 2C_1^{-2d}$$

Proof Sketch. We will let K be some large enough constant. First, we may assume $\|W\|_{op} \le K\sqrt{d}$, because for large enough K, the failure probability of this is C_1^{-2d}. Next, we have that since z lies in d dimensions, there is an $\frac{\varepsilon}{10K}$ net of size $(30K/\varepsilon)^{d-1}$ for z, where K is going to be some large constant. Each such point satisfies the desired bound $\text{dist}(Bz, H_{[n]\setminus J}) = \|Wz\|_2 \ge \varepsilon\sqrt{d}$ with at least $(C\varepsilon)^{2d-1}$ failure probability (there will be an extra e^{-cN} term, but for simplicity we'll ignore it). So, across the net, the failure probability is at most $(C\varepsilon)^{2d-1} \cdot (30K/\varepsilon)^{d-1} \le (30C^2K\varepsilon)^d$.

Next, any z is within $\frac{\varepsilon}{10K}$ of some point z' in the net, and we are assuming $\|W\|_{op} \le K\sqrt{d}$. So, $\|Wz\|_2 - \|Wz'\|_2 \le \|W\|_{op} \cdot \frac{\varepsilon}{10K} \le \frac{\varepsilon}{10}\sqrt{d}$. So, if the operator norm bound holds, and the bound holds for all points in the net, then for any $z \in \text{Spread}_J$, we have $\|Wz\|_2 \ge 0.9\sqrt{d}$. By rescaling ε, the bound is complete.

The overall failure probability is $C_1^{-2d} + (30C^2K\varepsilon)^d$, so for small enough ε, this is at most $2C_1^{-2d}$. □

We have therefore established that $\mathbb{P}\left(\inf_{z \in \text{Spread}_J} \text{dist}(Bz, H_{[n]\setminus J}) < \varepsilon\sqrt{d}\right) \le 2C_1^{-2d}$. By combining this with Lemma 3.1, we thus have that

$$\mathbb{P}\left(\inf_{v \text{ is } (\delta,\rho)-\text{incompressible}} \|Bv\|_2 \le c_1\varepsilon \cdot \frac{d}{\sqrt{n}}\right) \le 2C_1^{-d}.$$

Since $d = N - n + 1$, this gives us a bound exponentially small in $N - n + 1$ for the probability that $\mathbb{P}(\|B\|_{op} \leq \varepsilon \cdot (\sqrt{N} - \sqrt{n}))$ is exponentially small in $N - n + 1$, since $\frac{d}{\sqrt{n}} = \Theta(\sqrt{N} - \sqrt{n})$ is true whenever $n \geq N/2$. This proves a slightly weaker version of Theorem 1.1, where we have the probability bound of $e^{-\Omega(d)}$.

The full bound that is achievable is slighly better, i.e., at most $(C\varepsilon)^d + e^{-\Omega(N)}$. This is an improvement for small ε and when $d = N - n + 1 = o(N)$. For the purpose of this exposition, I will only briefly explain how this can be achieved.

In the proof sketch of Lemma 3.2, the C_1^{-2d} term comes from the operator norm bound on W, whereas the union bound over the net gives us something like ε^d which is what we want. To avoid this C_1^{-2d} term that doesn't decay with ε, the idea is to consider various scalings of $\|W\|_{op}$. Indeed, the probability that $\|W\|_{op}$ is way bigger than $\sqrt{d}$ (for instance $\sqrt{d}/\varepsilon$) behaves like e^{-d/ε^2} which indeed decays very quickly with ε. There is a caveat which is that the event that W has large operator norm and some $\|Wz\|_2$ being small for z in the net may be correlated. However, to fix this, Rudelson and Vershynin show a "decoupling" argument which allows us to bound the probability that both $\|Wz\|_2 \approx K\sqrt{d}$ and $\|Wz\|_2 \geq \varepsilon\sqrt{d}$ for all z in the (ε/K)-net, even for K that may be super-constant. (See Subsection 7.2 in [5] for more details of this decoupling argument.) This will essentially allow ut so replace the C_1^{-2d} term with $O(\varepsilon)^d + e^{-cN}$, which will be sufficient to prove Theorem 1.1.

4 Bounding the distance to a random subspace

In this section, we give a rough proof of Theorem 3.1. We first recall the statement.

Theorem 3.1 (Distance to a random subspace)**.** *Let ξ be a subgaussian distribution, and suppose $X \in \mathbb{R}^N$ has every entry i.i.d. drawn from ξ. Let H be a random subspace in $\mathbb{R}^N$ spanned by $N - m$ vectors $X_1, X_2, \ldots, X_{N-m}$, where $m \ll n$, and every coordinate of each X_i is i.i.d. drawn from some subgaussian distribution (possibly different from ξ).*

Then, for every $v \in \mathbb{R}^N$ and every $\varepsilon > 0$, we have that

$$\mathbb{P}\left(\operatorname{dist}(X, H + v) < \varepsilon\sqrt{m}\right) \leq (C\varepsilon)^m + e^{-cN},$$

for some appropriate constants $C, c > 0$.

The high-level approach is the following. We can think of the distance of X to $H + v$ as the distance from X to v, after we project out the subspace H (equivalently, project onto the orthogonal complement $H^\perp$). Using some Fourier analytic methods, we will bound the desired probability $\mathbb{P}(\operatorname{dist}(X, H + v) < \varepsilon\sqrt{m})$ in terms of something called the least common denominator of H. This definition will appear somewhat naturally from the quantities we are trying to bound, but perhaps will seem unusual. Then, we will need to bound this least common denominator.

4.1 Least Common Denominator

We first define the least common denomator of a unit vector a, and then for a set of orthogonal unit vectors (or equivalently, the spanned subspace). Intuitively, the least common denominator represents the smallest multiple of the unit vector a (or the shortest vector in the subspace) that is close to a non-trivial integer lattice point. We give a slightly different from the one in [5], but it is functionally the same.

Definition 4.1. Fix $a \in \mathbb{R}^N$ as a unit vector, and some parameters $\alpha, \gamma \in (0,1)$. We define the *least common denominator* of a as

$$\mathrm{LCD}_{\alpha,\gamma}(a) := \inf\left\{\theta > 0 : \mathrm{dist}(\theta \cdot a, \mathbb{Z}^N) < \min(\gamma \cdot \theta, \alpha\right\} \tag{2}$$

Definition 4.2. For a subspace $H \subset \mathbb{R}^N$, we define the *least common denominator* of H as

$$\mathrm{LCD}_{\alpha,\gamma}(H) := \inf\left\{\|v\|_2 : v \in H, \mathrm{dist}(v, \mathbb{Z}^N) < \min(\gamma \cdot \|v\|_2, \alpha\right\}.$$

If $a^{(1)}, \ldots, a^{(m)}$ is an orthonormal basis for the subspace H, we may also define $\mathrm{LCD}_{\alpha,\gamma}(a^{(1)}, \ldots, a^{(m)})$ as the same as $\mathrm{LCD}_{\alpha,\gamma}(H)$.

Whether we are talking about the least common denominator of a single vector or a subspace/set of vectors is clear from context. Note that for any subset H, $\mathrm{LCD}_{\alpha,\gamma}(H) = \inf_{a \in H, \|a\|_2 = 1} \mathrm{LCD}_{\alpha,\gamma}(v)$.

4.2 Small ball probability

Our goal in this subsection is to bound the left-hand side of Equation (2), using the least common denominator.

Consider a projection matrix $P_{H^\perp}$ which projects $\mathbb{R}^N$ to the orthogonal complement $H^\perp$ of the subspace $H \subset \mathbb{R}^N$. If H has dimension $n - m$, then $H^\perp$ has dimension m, spanned by some $a^{(1)}, \ldots, a^{(m)} \in \mathbb{R}^N$. If $A \in \mathbb{R}^{N \times m}$ is the matrix with columns $a^{(1)}, \ldots, a^{(m)}$, let $a_1, \ldots, a_N \in \mathbb{R}^m$ represent the rows of A. Now, note that $\mathrm{dist}(X, H + v) = \mathrm{dist}(X - v, H) = \|\Pi_{H^\perp}(X - v)\|_2$. We can write $\Pi_{H^\perp} = AA^\top$, so $\|\Pi_{H^\perp}(X - v)\|_2 = \|AA^\top(X - v)\|_2 = \|A^\top(X - v)\|_2 = \|A^\top X - A^\top v\|_2$, using the fact that A's columns are orthonormal. Since every entry of $X \in \mathbb{R}^N$ is i.i.d. from ξ, if we write $X = (\xi_1, \ldots, \xi_N)$, we can define $S := A^\top X = \sum_{i=1}^N a_i \xi_i$, and $w := A^\top v$. Thus, it is equivalent to bound the probability that $\|S - w\|_2 < \varepsilon\sqrt{m}$.

Rudelson and Vershynin indeed prove a probability bound on this quantity in terms of the least common denominator. Specifically, they show the following result.

Theorem 4.1. *Let $\xi_1, \ldots, \xi_N$ be drawn from the univariate subgaussian distribution ξ. Let $a_1, \ldots, a_N \in \mathbb{R}^m$ be the rows of a matrix $A \in \mathbb{R}^{N \times m}$ with orthonormal columns. Let $S = \sum \xi_i a^{(i)}$. Then, for every $\alpha > 0$ and $\gamma \in (0,1)$, for every $w \in \mathbb{R}^m$, and any $\varepsilon \geq \frac{\sqrt{m}}{\mathrm{LCD}_{\alpha,\gamma}(a^{(1)}, \ldots, a^{(m)})}$, we have*

$$\mathbb{P}_{\xi_1, \ldots, \xi_N \sim \xi}\left(\|S - w\|_2 > \varepsilon\sqrt{m}\right) \leq \left(\frac{C\varepsilon}{\gamma}\right)^m + C^m \cdot e^{-2b\alpha^2},$$

for some constants C and b (which may depend on the subgaussian norm of ξ).

This result can be thought of as an *anti-concentration* result. In other words, it shows that the random variable S is *not* concentrated in any ball of radius $\varepsilon\sqrt{m}$.

Intuition behind proof: The very high-level intuition behind proving this is as follows. First, note that for fixed $a^{(i)}$, S can be thought of as a sum of N independent random variables. Bounding the distribution of this appears daunting, but the nice thing is that the Fourier transform of S is easy to bound. This is because the Fourier transform of a sum of random variables is the product of the Fourier transforms of the individual random variables. Also, each $a^{(i)}$ is a fixed vector and each ξ_i is a one-dimensional random variable, which will make things easier for us.

There is in fact a known result (the Esseen Lemma), which can bound the probability that $\|S - w\|_2 > \varepsilon\sqrt{m}$ in terms of the Fourier transform. The behavior of the Fourier transform will cause some terms of the form $1 - \cos(2\pi x)$ to show up, which in a sense is an indicator to how close x is to an integer. This is the intuition for why this "least common denominator", which captures distances to integer points, ends up appearing naturally in the bound for Theorem 4.1.

We now give a slightly more detailed proof sketch of Theorem 4.1.

Proof Sketch of Theorem 4.1. The first step in the proof is to use the Esseen lemma, which tells us that for any random variable $Y \in \mathbb{R}^m$, and any vector w,

$$\mathbb{P}(\|Y - w\|_2 \leq \sqrt{m}) \leq C^m \cdot \int_{B(0,\sqrt{m})} |\phi_Y(\theta)|, \tag{3}$$

where $\phi_Y(\theta) = \mathbb{E}[e^{2\pi i \langle \theta, Y \rangle}]$ is the characteristic function (i.e., Fourier Transform) of Y.

One can compute the squared Fourier transform $|\phi_S(\theta)|^2$, by decomposing S into the individual sums $a_i x_i$, as $\prod_{k=1}^{N} \mathbb{E}[\cos(2\pi \cdot \frac{\langle \theta, a_k \rangle}{\varepsilon} \cdot \bar{\xi})]$, where $\bar{\xi}$ is the random variable generated by taking the difference between two independent random copies of ξ. This ends up being a simple calculation (each term in the product comes from the Fourier transform of one $\xi_k \cdot a_k$).

Now, let's suppose $\bar{\xi}$ equals some fixed value z with some fixed constant probability. In this case, the term $\cos(2\pi \cdot \frac{\langle \theta, a_k \rangle}{\varepsilon} \cdot \bar{\xi})$ equals $\cos(2\pi \cdot \frac{\langle \theta, a_k \rangle}{\varepsilon} \cdot z)$ with constant probability, and otherwise is still bounded by 1. We can always write $\cos(2\pi x) = 1 - \Theta(\text{dist}(x, \mathbb{Z})^2)$, where $\text{dist}(x, \mathbb{Z})$ is the distance from x to the closest integer. Thus, $\mathbb{E}\left[\cos(2\pi \cdot \frac{\langle \theta, a_k \rangle}{\varepsilon} \cdot \bar{\xi})\right] \leq 1 - \Theta(\text{dist}(\frac{z}{\varepsilon} \cdot \langle \theta, a_k \rangle, \mathbb{Z})^2) \leq \exp\left(-\Theta(\text{dist}(\frac{z}{\varepsilon} \cdot \langle \theta, a_k \rangle, \mathbb{Z})^2)\right)$. Multiplying this across all N, we can move the product inside the exponential to obtain

$$|\phi_S(\theta)| \leq \exp\left(-\Theta\left(\sum_{k=1}^{N} \text{dist}\left(\frac{z}{\varepsilon} \cdot \langle \theta, a_k \rangle, \mathbb{Z}\right)^2\right)\right) = \exp\left(-\Theta\left(\text{dist}\left(\frac{z}{\varepsilon} \cdot A \cdot \theta, \mathbb{Z}^N\right)^2\right)\right).$$

While we don't have a guarantee that there is any value z such that $\bar{\xi} = z$ with such high probability, it is known that $\bar{\xi}$, which is the difference of two copies of ξ, is a sub-gaussian random variable with variance 2. This is enough to ensure that $|\xi| \geq 1$ with at least some constant probability (which may depend on the subgaussian norm of ξ). Rudelson and Vershynin show that one can a sort of "average" over the choices of $\bar{\xi}$ that are at least 1 in absolute value, and use Jensen's inequality to prove that

$$\int_{B(0,\sqrt{m})} |\phi_S(\theta)| \leq \sup_{z \geq 1} \int_{B(0,\sqrt{m})} \exp\left(-\Theta\left(\text{dist}\left(\frac{z}{\varepsilon} \cdot A \cdot \theta, \mathbb{Z}^N\right)^2\right)\right) d\theta. \tag{4}$$

By combining with Equation (3), it suffices to bound the integral on the right-hand side above for any fixed choice of $z \geq 1$.

To bound this integral, one can bound the volume of points $\theta \in B(0, \sqrt{m})$ such that $\text{dist}\left(\frac{z}{\varepsilon} \cdot A \cdot \theta, \mathbb{Z}^N\right) \leq t$, for any t. Recall the bounds on $\alpha, \gamma, \varepsilon$ in the statement of Theorem 4.1. Rudelson and Vershynin prove that, under these assumptions on $\alpha, \gamma, \varepsilon$, and for any $t \leq \alpha/2$, the volume of this set $I(t) := \left\{\theta \in B(0, \sqrt{m}) : \text{dist}\left(\frac{z}{\varepsilon} \cdot A \cdot \theta, \mathbb{Z}^N\right) \leq t\right\}$ is at most $\left(\frac{Ct\varepsilon}{\gamma\sqrt{m}}\right)^m$ for some constant C.

The proof of this bound roughly works as follows. If we consider two points θ, θ' in this set $I(t)$, consider the value $\tau = \frac{z}{\varepsilon} \cdot A \cdot (\theta - \theta')$. One can prove that the Euclidean norm of τ must either be quite small or quite large. This is because the definition of the least common denominator tells us that, since τ is in the subspace generated by the columns of A, either $\|\tau\|_2 \geq \text{LCD}_{\alpha,\gamma}(H)$, or $\text{dist}(\tau, \mathbb{Z}^N) \geq \alpha$, or $\text{dist}(\tau, \mathbb{Z}^N) \geq \gamma \cdot \|\tau\|_2$. The first option is the case where the norm of τ is large. We also know that $\text{dist}(\tau, \mathbb{Z}^N) \leq 2t < \alpha$, since we assume $t < \alpha/2$. So the second case is actually impossible. Finally, in the third case, since $\text{dist}(\tau, \mathbb{Z}^n) \leq 2t$, we get an upper bound on τ. Finally, since A has orthonormal columns and since $\frac{z}{\varepsilon}$ is a scalar, this bound on τ implies that for any $\theta, \theta' \in I(t)$, either θ, θ' are close in distance, or are far in distance. This will in fact imply that the set $I(t)$ of possible θ can be covered by a series of small but well-separated balls, which will be sufficient to prove that the volume of $I(t)$ is at most $\left(\frac{Ct\varepsilon}{\gamma\sqrt{m}}\right)^m$.

The final step of the proof is to combine the Equations (3) and (4) with the volume bound on $I(t)$. We have a good bound for $t = \text{dist}\left(\frac{z}{\varepsilon} \cdot A \cdot \theta, \mathbb{Z}^N\right) < \frac{\alpha}{2}$, but beyond this, we are integrating a quantity that is at most $e^{-\Omega(\alpha^2)}$. Overall, this will be sufficient to prove the theorem. □

4.3 Bounding the least common denominator

To complete the proof of Theorem 3.1, we wish to bound $\text{LCD}_{\alpha,\gamma}(H^\perp) = \text{LCD}_{\alpha,\gamma}(a^{(1)}, \ldots, a^{(m)})$, for some appropriate choices for α, γ. This way, we can set $\varepsilon = \frac{\sqrt{m}}{\text{LCD}_{\alpha,\gamma}(a^{(1)},\ldots,a^{(m)})}$ in Theorem 4.1 to prove Theorem 3.1. This bound will use the fact that H is not just an arbitrary subspace, but instead the subspace generated by the random vectors $X_1, \ldots, X_{N-m}$. Let B be the matrix with columns $X_1, \ldots, X_{N-m}$, so $H = Span(B)$ and $H^\perp = Ker(B^\top) = \{v : B^\top v = 0\}$.

To show that $\text{LCD}_{\alpha,\gamma}(H^\perp)$ is at least some quantity T, we need to show that for any vector $v \in H^\perp$ with $\|v\|_2 \leq T$, either $\text{dist}(v, \mathbb{Z}^N) \geq \alpha$ or $\text{dist}(v, \mathbb{Z}^N) \geq \gamma \cdot \|v\|_2$. The proof of the result is based on the following outline.

1. With very high probability, any unit vector in $H^\perp$ must be *incompressible*.
2. Every incompressible vector must have $\text{LCD}_{\alpha,\gamma}(v) \geq \Omega(\sqrt{N})$ (for appropriately chosen parameters). Thus, $\text{LCD}_{\alpha,\gamma}(H^\perp) \geq \Omega(\sqrt{N})$.
3. The above bound on $\text{LCD}_{\alpha,\gamma}(H^\perp)$ is not sufficient. The next step is to consider the set of incompressible vectors with LCD between $\Omega(\sqrt{N})$ and $O(\sqrt{N} \cdot e^{cN/m})$, and show that for any fixed v in this set, it is very unlikely to be close to the kernel of $B^\top$ (and thus in $H^\perp$).
4. Finally, one needs to show the above set has a small ε-net, and thus we can apply a union bound based on this ε net, for each D up to roughtly $N \cdot e^{cN/m}$ for some constant c. Thus will allow us to bound $\text{LCD}_{\alpha,\gamma}(H^\perp) \geq \Omega(\sqrt{N} \cdot e^{cN/m})$, which will in fact be sufficient.

We now give a brief proof overview of each of these steps.

Lemma 4.1 (Random subspaces are incompressible)**.** *There exist some constants $0 < \delta, \rho < 1$ such that with at most e^{-cN} failure probability, every unit vector in $H^\perp$ is (δ, ρ)-incompressible.*

Proof. Suppose the opposite holds, so there is some vector $v \in H^\perp$ that is (δ, ρ)-compressible. Then, $B^\top v = 0$, so $\|B^\top v\|_2 = 0$. But $B^\top$ is an $(N - m) \times N$ dimensional matrix, which means the number of rows is at least half the number of columns. Thus, we can apply Lemma 2.4 which tells us the probability of any (δ, ρ)-compressible vector v having $\|B^\top v\|_2 = 0$ is at most e^{-cN}. □

Lemma 4.2 (LCD is somewhat large)**.** *Fix $0 < \delta, \rho < 1$. Then, there exist constants c_1, c_2, depending on δ, ρ, such that for every $0 < \gamma < c_1$ and $\alpha > 0$, and for every (δ, ρ)-incompressible vector $a \in \mathbb{R}^N$, $\mathrm{LCD}_{\alpha,\gamma}(a) > c_2 \cdot \sqrt{N}$.*

Proof Sketch. Let's start by assuming that $a = (a_1, a_2, \ldots, a_N)$, where $|a_i| \cdot \sqrt{N}$ is always between two constants (for instance, always between $1/2$ and 2). Now, suppose there is a scaling parameter $0 < \theta \leq c_2\sqrt{N}$ such that $\mathrm{dist}(\theta \cdot a, \mathbb{Z}^N) \leq \gamma \cdot \theta$. If c_1 is chosen as less than 1, then the closest point to $\theta \cdot a$ can't be the all 0's vector, because then $\mathrm{dist}(\theta \cdot a, \mathbb{Z}^N) = \theta > \gamma \cdot \theta$. So, some $\theta \cdot a_i$ must have absolute value at least $1/2$, so $\theta \geq \Omega(\sqrt{N})$.

For general incompressible a, we still know that there is a subset $\mathcal{S}$ of the coordinates $[N]$ such that $|\mathcal{S}| = \Omega(N)$ and $|a_i| = \Theta(1/\sqrt{N})$ for all $i \in \mathcal{S}$, by Lemma 2.2. If θ is much smaller than $\Omega(\sqrt{N})$, then $|\theta \cdot a_i| < \frac{1}{2}$ for all $i \in \mathcal{S}$. So, the closest integer lattice point p to $\theta \cdot a$ still has $p_i = 0$ for all $i \in S$, so $\|\theta \cdot a_i - p\|_2 \geq \sqrt{\sum_{i \in \mathcal{S}} \theta^2 \cdot a_i^2} \geq \theta \cdot \Omega(1/\sqrt{N} \cdot \sqrt{|\mathcal{S}|}) \geq \Omega(\theta)$. So, we can choose c_1 small enough so that the condition of $\mathrm{dist}(\theta \cdot a, \mathbb{Z}^N) \leq \gamma \cdot \theta$ is violated. Thus, we must have $|\theta \cdot a_i| \geq \frac{1}{2}$ for some $i \in \mathcal{S}$, which implies $\theta \geq \Omega(\sqrt{N})$. □

This completes the first two steps. We fix some α to be a small multiple of $\sqrt{N}$ and γ to be some small constant. Define S_D to be the subset of (δ, ρ)-incompressible unit vectors x such that $D \leq \mathrm{LCD}_{\alpha,\gamma}(x) < 2D$. Note that S_D is a deterministic set. We already know that any unit vector in $H^\perp$ has $\mathrm{LCD}_{\alpha,\gamma}(x) \geq c_2\sqrt{N}$ from Lemmas 4.1 and 4.2. So, the next step is to show that for each of $D = c_2\sqrt{N}, 2c_2\sqrt{N}, 4c_2\sqrt{N}, \ldots, \sqrt{N} \cdot e^{cN/m}$, S_D does not intersect $H^\perp$.

The third step is to show this holds for a fixed D and $v \in S_D$: in fact, that for any fixed $v \in S_D$, $\|B^\top v\|_2$ is not only nonzero but also reasonably large.

Lemma 4.3 (Single vector bound)**.** *Fix $v \in S_D$. Then, for any $t > 0$, we have*

$$\mathbb{P}_B(\|B^\top v\|_2 < t\sqrt{N}) \leq \left(C\left(t + \frac{1}{D} + e^{-c\alpha^2}\right)\right)^{N-m}.$$

Proof Sketch. The idea is to bound a single coordinate of $B^\top v$. Indeed, $(B^\top v)_j$ is simply $\sum B_{ij} v_j$, where each B_{ij} is drawn from the same subgaussian distribution ξ. This, in fact, is bounded by Theorem 4.1, by setting $m = 1$, $A \in \mathbb{R}^{N \times 1}$ as the vector v, and $B_{ij} = \xi_i$. For $\varepsilon = \frac{1}{\mathrm{LCD}_{\alpha,\gamma}(x)}$, this will end up showing that

$$\mathbb{P}\left(|(B^\top v)_j| < t\right) \leq Ct + \frac{C}{\mathrm{LCD}_{\alpha,\gamma}(x)} + C^m \cdot e^{-c\alpha^2}$$

for some constants C, c.

This is enough to bound each entry of $B^\top v$. Since each entry is i.i.d., one can also bound the full norm of $B^\top v$, using something called a Tensorization lemma (see, e.g., Lemma 2.2 in [4]). □

Finally, one needs to prove such a bound uniformly across all $v \in S_D$. There are only about N/m choices for D, so union bounding across all D is not an issue. This again uses a net argument, but now one needs to bound the size of a net of S_D, which is the final ingredient.

Lemma 4.4 (Net for S_D)**.** *There exists a $(4\alpha/D)$-net of S_D, with size at most $(C_0 D/\sqrt{N})^N$, where C_0 is a (reasonably small) constant.*

Proof Sketch. The trick is to realize that any unit vector v in S_D satisfies $\text{dist}(\theta v, p) \leq \alpha$ for some $D \leq \theta < 2D$ and some integer lattice vector p, by definition of S_D. Thus, we can write $v = \frac{p}{\theta} + v'$ for some v' with norm at most $\frac{\alpha}{\theta} \leq \frac{\alpha}{D}$. So, the set of points $\frac{p}{\theta}$ is a suitable net.

The number of choices of p is boundable, since p has to be an integer point with ℓ_2 norm at most $2D$, and for $D \geq \Omega(\sqrt{N})$, it is well known that the number of integer points is at most $(C_0 D/\sqrt{N})^N$. Finally, while θ has an infinite number of choices, we can round θ to the nearest multiple of $1/D$, and this will only marginally affect v'. So, there are only about $D^2 \cdot (C_0 D/\sqrt{N})^N$ choices for p and θ, and the D^2 is insignificant comparatively. □

We now can combine this together and explain (at a high level) how this bounds the least common denominator of $H^\perp$, and thus completes the proof of Theorem 3.1.

Proof outline of Theorem 3.1. The idea is to set $t = c_2\sqrt{N}/D$ in Lemma 4.3, so that as long as $D \leq e^{O(N)}$, Lemma 4.3 can be simplified as $\mathbb{P}_B(\|B^\top v\|_2 < c_2 N/D) \leq \left(C \cdot \frac{\sqrt{N}}{D}\right)^{N-m}$. Conversely, the size of the net is $(C_0 D/\sqrt{N})^N$. It will turn out that C_0 sill be a reasonable constant (much smaller than C, and so a union bound will tell us the probability of this holding for any v in the net is at most $(D/\sqrt{N})^m \cdot C_0^N / C^{N-m}$. Because m is much smaller than N and C_0 is much smaller than C, it will turn out that $C_0^N / C^{N-m} = e^{-\Omega(N)}$. So, as long as $D \leq \sqrt{N} \cdot e^{cN/m}$ for a small enough constant c, a union bound still gets us a probability of $e^{-\Omega(N)}$.

While this gives us a bound for the net, for any other point v' in S_D, Lemma 4.4 tells us that v' is of distance at most $4\alpha/D$ from some v in the net. Moreover, with exponential failure probability (as explained in Subsection 2.3), we know that $B^\top$ has operator norm at most $O(\sqrt{N})$, so $\|B^\top v'\|_2 \geq \|B^\top v\|_2 - \|B\|_{op} \cdot 4\alpha/D = c_2 N/D - O(\alpha\sqrt{N}/D)$. We can set $\alpha = \mu\sqrt{N}$ where μ is a sufficiently small constant, so that this quantity is strictly greater than 0. Thus, no $v' \in S_D$ is in the kernel of $B^\top$, at least for $D \leq \sqrt{N} \cdot e^{cN/m}$, which thus implies that $\text{LCD}_{\alpha,\gamma}(H^\perp) \geq \Omega(\sqrt{N} \cdot e^{cN/m})$ by the definition of LCD.

This bound can now be directly plugged into Theorem 4.1. As in the discussion above the statement of Theorem 4.1, the term $\|S - w\|_2$ is the same as $\text{dist}(X, H + v)$, and some basic calculations can be used to show that we exactly obtain Theorem 3.1. □

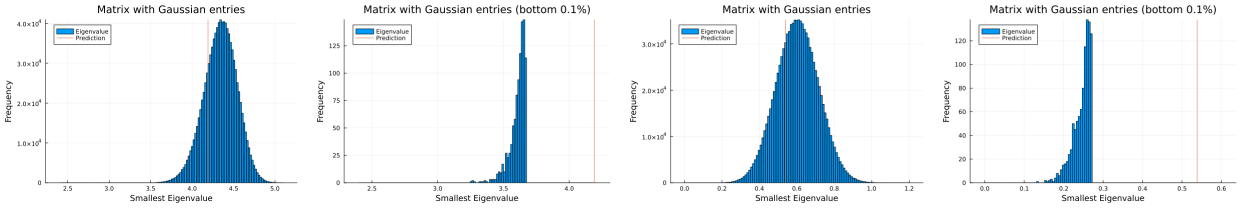
5 Experiments

In this section, we describe some experiments to determine how well the results of [5] hold in practice for reasonably small matrices, and to what extent the distribution of the entries changes the outcome. In each experiment, we draw 1 million random $N \times n$ matrices, where each entry is drawn i.i.d. from some distribution $\mathcal{D}$ with mean 0 and variance 1 (to ensure proper normalization). For each generated random matrix, we keep track of the smallest singular value. We consider two settings of (N, n): we try both $N = 200, n = 100$ (which is more rectangular) and $N = 110, n = 100$ (which is more squarish). Finally, we plot both the full histogram of the singular values, and the histogram of the bottom 1000 singular values (i.e., bottom 0.1%), to understand whether the tail bounds hold empirically. We also compare the singular value distribution to the predicted value, which is $\sqrt{N} - \sqrt{n-1}$ (which is drawn as a red line in each figure).

First, we consider three sub-gaussian distributions for the entries, for which we have strong tail bounds for (Theorem 1.1). We use the standard Normal distribution (Figure 1), the random sign

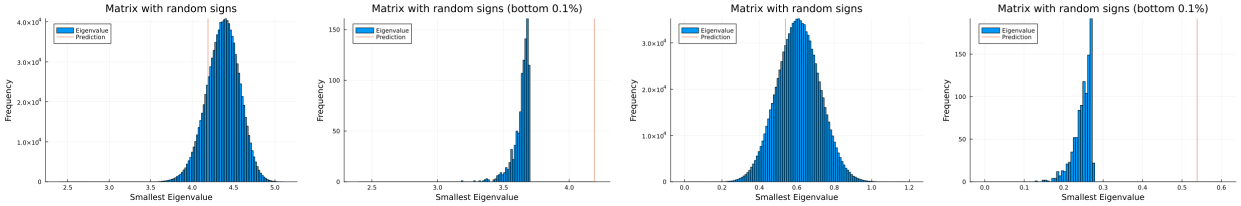
distribution, where each entry is a random ± 1 variable (Figure 2), and a properly scaled Uniform distribution over the interval $[-\sqrt{3}, \sqrt{3}]$ (Figure 3).

The distributions of the smallest singular values seem to be approximately the same, regardless of which subgaussian distribution was chosen. In both the the case where $N = 200, n = 100$ and where $N = 110, n = 100$, the smallest singular value on average is actually slightly larger than the prediction of $\sqrt{N} - \sqrt{n-1}$. This suggests that the subgaussian entries are nice enough that in the nonasymptotic regime, the smallest singular value might even be a bit larger due to the nice properties of the entries. Moreover, in the $N = 200, n = 100$ case for all three subgaussian distributions, the smallest singular value in all 1 million runs was always at least 3, which is in fact at least $2/3$ of the predicted value $\sqrt{N} - \sqrt{n-1} \approx 4.19$. In the $N = 110, n = 100$ case, the smallest singular value in all 1 million runs was always at least 0.13, which is about $1/4$ of the predicted value $\sqrt{N} - \sqrt{n-1} \approx 0.538$. The 1000th smallest entry was slightly below 0.3, or about $1/2$ of the predicted value. These predictions seem to coincide with the theory in [5], because we are always within a constant factor, but when $N - n$ decreases from 100 to 10, the tail is significantly worse.



(a) Full histogram ($N = 200, n = 100$)
(b) Bottom 0.1% ($N = 200, n = 100$)
(c) Full histogram ($N = 110, n = 100$)
(d) Bottom 0.1% ($N = 110, n = 100$)

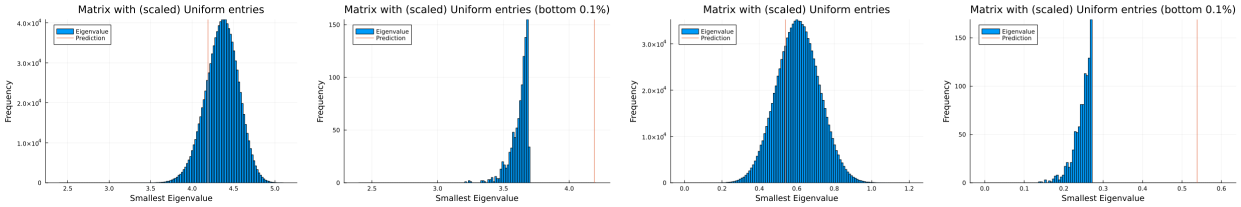
Figure 1: Histograms of the smallest singular value of 1 million random $N \times n$ matrices, where each entry is i.i.d. $\mathcal{N}(0, 1)$.



(a) Full histogram ($N = 200, n = 100$)
(b) Bottom 0.1% ($N = 200, n = 100$)
(c) Full histogram ($N = 110, n = 100$)
(d) Bottom 0.1% ($N = 110, n = 100$)

Figure 2: Histograms of the smallest singular value of 1 million random $N \times n$ matrices, where each entry is an i.i.d. random sign (i.e., 1 with probability $1/2$ and -1 with probability $1/2$).

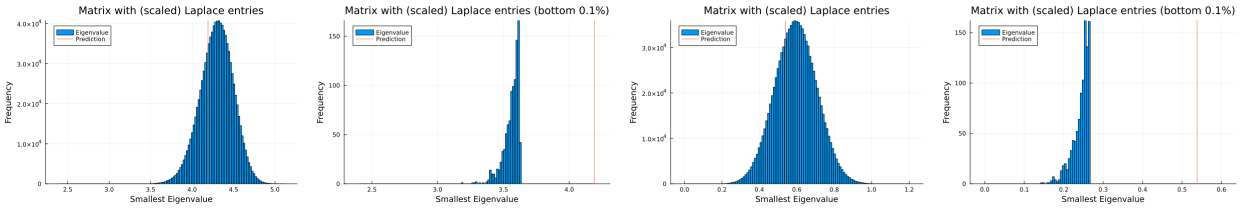
Next, we consider the case where the entries are not subgaussian but are subexponential (i.e., $\mathbb{P}(|\xi| > t) \leq e^{-\Omega(t)}$). In this case, while the results of [5] do not hold, the results seem almost



(a) Full histogram ($N = 200, n = 100$)
(b) Bottom 0.1% ($N = 200, n = 100$)
(c) Full histogram ($N = 110, n = 100$)
(d) Bottom 0.1% ($N = 110, n = 100$)

Figure 3: Histograms of the smallest singular value of 1 million random $N \times n$ matrices, where each entry is i.i.d. Unif$[-\sqrt{3}, \sqrt{3}]$. (This scaling ensures the variance of each entry is 1.)

identical in practice. (In the $N = 200, n = 100$ case, there seems to be a very small shift to the left, which may suggest that the concentration in the subexponential case could be marginally weaker.) We consider both a (scaled) Laplace distribution in Figure 4 (which is symmetric) and a (shifted) Poisson distribution in Figure 5 (which is highly asymmetric), but the overall distribution and bottom tail both look very similar.

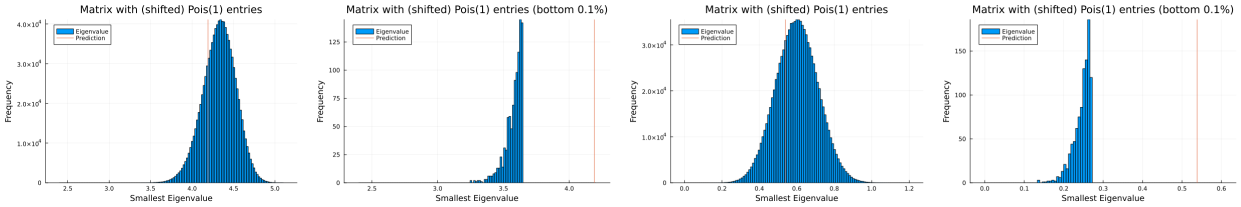


(a) Full histogram ($N = 200, n = 100$)
(b) Bottom 0.1% ($N = 200, n = 100$)
(c) Full histogram ($N = 110, n = 100$)
(d) Bottom 0.1% ($N = 110, n = 100$)

Figure 4: Histograms of the smallest singular value of 1 million random $N \times n$ matrices, where each entry is i.i.d. Laplace$(1/\sqrt{2})$. (This scaling ensures the variance of each entry is 1.)

Finally, we consider the case where the entries are from a heavy-tailed distribution. In Figure 6, we consider a student t-distribution with 3 degrees of freedom, scaled to have variance 1. This distribution is symmetric and has bounded second moments, but doesn't even have bounded third or fourth moments. We scale the distribution to have variance 1. In Figure 7, we consider the following distribution, which is the distribution with PDF $p(x)$ proportional to $\frac{1}{x^4}$ for $x \geq 1$, and $p(x) = 0$ for $x < 1$. We shift and scale the distribution to have mean 0 and variance 1. Like the t-distribution with 3 degrees of freedom, this distribution also has bounded second moments, but not bounded third or fourth moments. In addition, it is highly asymmetric, unlike the t distribution.

In the t distribution case (Figure 6), the overall distribution of the smallest singular value is reasonably close to the prediction (though on average it is slightly smaller, at least when $N = 200, n = 100$). The overall distribution and tail bounds seem to be slightly smaller than in the subgaussian case, but the tail still seems to be within a constant fraction of the predicted smallest

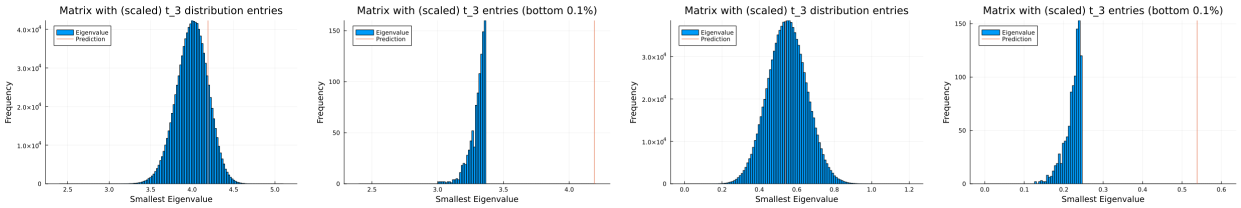


(a) Full histogram ($N = 200, n = 100$)
(b) Bottom 0.1% ($N = 200, n = 100$)
(c) Full histogram ($N = 110, n = 100$)
(d) Bottom 0.1% ($N = 110, n = 100$)

Figure 5: Histograms of the smallest singular value of 1 million random $N \times n$ matrices, where each entry is i.i.d. Pois(1) $-$ 1. (This scaling ensures the mean and variance of each entry are 0 and 1, respectively.)

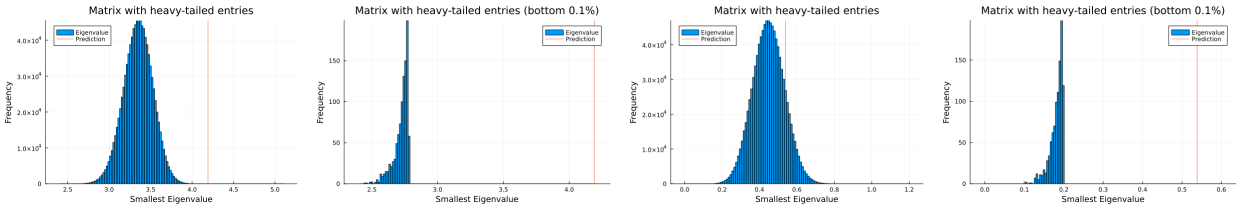
singular value. In the asymmetric heavy-tailed case (Figure 7), the overall distribution is shifted significantly to the left, and seems to be almost always less than the predicted value for $N = 200, n = 100$, though for $N = 110, n = 100$, the difference is a bit less signicant. Moreover, the tail is also significantly smaller than in the subgaussian case (especially in the $N = 110, n = 100$ case) but still seems to be within a constant factor of the predicted value.

One surprising aspect in the final case is that the full histogram is not centered anywhere near the predicted value, for $N = 200, n = 100$. (The ratio of the empirical average to the prediction is approximately 0.8.) However, even for heavy-tailed distributions, there is a known asymptotic convergence result to $(1+o(1))\cdot(\sqrt{N}-\sqrt{n})$ [7]. So, this convergence may be much slower for heavy-tailed distributions. Indeed, we tried repeating this for a 1000×500 matrix (to keep the ratio N/n the same) for 10 thousand iterations (see Figure 8). In this case, the histogram of smallest singular values was still consistently smaller than the predicted value, but the ratio of the empirical average to the prediction was approximately 0.9. So, this suggests a slower rate of convergence.



(a) Full histogram ($N = 200, n = 100$)
(b) Bottom 0.1% ($N = 200, n = 100$)
(c) Full histogram ($N = 110, n = 100$)
(d) Bottom 0.1% ($N = 110, n = 100$)

Figure 6: Histograms of the smallest singular value of 1 million random $N \times n$ matrices, where each entry is i.i.d. from a t distribution with 3 degrees of freedom, scaled by a factor of $1/\sqrt{3}$. (This scaling ensures the variance of each entry is 1.)



(a) Full histogram ($N = 200, n = 100$)
(b) Bottom 0.1% ($N = 200, n = 100$)
(c) Full histogram ($N = 110, n = 100$)
(d) Bottom 0.1% ($N = 110, n = 100$)

Figure 7: Histograms of the smallest singular value of 1 million random $N \times n$ matrices, where each entry is drawn i.i.d. from the distribution with PDF proportional to $1/x^4$ for $x \geq 1$, and then shifted/scaled to ensure the mean and variance are 0 and 1, respectively.)

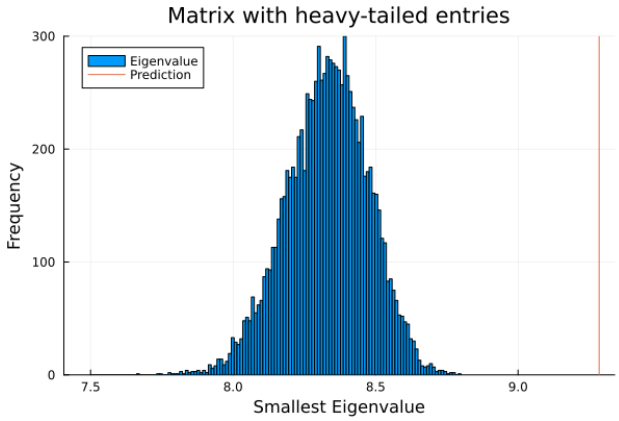


Figure 8: Histogram of the smallest singular value of 10 thousand random 1000×500 matrices, where each entry is drawn i.i.d. from the distribution with PDF proportional to $1/x^4$ for $x \geq 1$, and then shifted/scaled to ensure the mean and variance are 0 and 1, respectively.)

References

[1] Z. Bao, G. Pan, and W. Zhou. Tracy-widom law for the extreme eigenvalues of sample correlation matrices. *Electron. J. Probab*, 17:1–32, 2012.

[2] G. Bennett, L. E. Dor, V. Goodman, W. B. Johnson, and C.M. Newman. On uncomplemented subspaces of l_p, $1 < p < 2$. *Isr. J. Math.*, 26:178–187, 1977.

[3] A.E. Litvaka, A. Pajorb, M. Rudelson, and N. Tomczak-Jaegermann. Smallest singular value of random matrices and geometry of random polytopes. *Isr. J. Math.*, 26:178–187, 1977.

[4] M. Rudelson and R. Vershynin. The littlewood-offord problem and invertibility of random matrices. *Adv. Math.*, 218:600–633, 2008.

[5] M. Rudelson and R. Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009.

[6] J. W. Silverstein. The smallest eigenvalue of a large dimensional wishart matrix. *Ann. Probab.*, 13(4):1364–1368, 1985.

[7] K. Tikhomirov. The limit of the smallest singular value of random matrices with i.i.d. entries. *Adv. Math.*, 284:1–20, 2015.