Lecture Notes on Random Matrix Theory in Data Science and Statistics

Dmitriy (Tim) Kunisky

Fall 2025 (Last Updated: September 12, 2025)

Contents

1 | Random Vector Theory

Before we continue to random matrix theory, it will be useful to establish some preliminaries about random *vectors*, which will also lead us to a few important intuitions about high-dimensional probability and geometry.

1.1 Natural Random Models and Orthogonal Invariance

As in the case of matrices, vectors have at least two natural interpretations: they are a **list of numbers**, as well as a geometric object with a **magnitude and direction** (related notions about vector and Banach spaces are abstractions of this latter viewpoint) with respect to a basis of the space to which they belong. Each interpretation gives rise to a natural model[1] of random vector.

In the list-of-numbers interpretation, it is most logical to model a random vector where the entries are as statistically decoupled as possible. Thus, we consider

$$\boldsymbol{x} \sim \mu^{\otimes d}, \tag{1.1.1}$$

a vector with entries i.i.d. from some probability measure μ on $\mathbb{R}$.

In the magnitude-and-direction interpretation, it is most logical to model a random vector where the magnitude and direction are as statistically decoupled as possible. Thus, we take these two properties to be independent (that is, a random $\boldsymbol{x}$ where $\|\boldsymbol{x}\|$ and $\boldsymbol{x}/\|\boldsymbol{x}\|$ are independent). Since the magnitude can be any random non-negative scalar, the more interesting new information of such a vector distribution is in its direction. Thus, we consider

$$\boldsymbol{x} \sim \mathsf{Unif}(\mathbb{S}^{d-1}), \tag{1.1.2}$$

a vector drawn from the uniform surface measure on the unit sphere $\mathbb{S}^{d-1} := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$ (so denoted because it is a $(d-1)$-dimensional manifold).

More generally, the following property of a vector distribution is natural to consider in the context of the magnitude-and-direction interpretation.

Definition 1.1.1 (Orthogonal invariance)**.** *We say that a random vector $\boldsymbol{x} \in \mathbb{R}^d$ (or its law $\mathsf{Law}(\boldsymbol{x})$) is* orthogonally invariant *if, for each (deterministic) orthogonal matrix $\boldsymbol{Q} \in \mathcal{O}(d)$, we have $\mathsf{Law}(\boldsymbol{Q}\boldsymbol{x}) = \mathsf{Law}(\boldsymbol{x})$.*

[1] We will sometimes use this word "model" instead of "law" or "distribution" to emphasize that such a choice includes some judgement of what distributional assumption is natural or reasonable—we make such a choice to try to *model* typical situations we might encounter in applications.

Without going into more technical details of actually defining this probability measure explicitly, the following result should clarify the meaning of the measure $\mathsf{Unif}(\mathbb{S}^{d-1})$.

Proposition 1.1.2. *There exists a unique probability measure supported on $\mathbb{S}^{d-1}$ that is orthogonally invariant, which we denote $\mathsf{Unif}(\mathbb{S}^{d-1})$.*

The following also shows that such distributions are nothing more than vectors distributed as $\mathsf{Unif}(\mathbb{S}^{d-1})$ rescaled by some independent random scalar, and thus are precisely the more general class of models natural to the magnitude-and-direction interpretation that we indicated above.

Proposition 1.1.3. *Suppose that $\boldsymbol{x}$ is orthogonally invariant. Then, $\|\boldsymbol{x}\|$ is independent of $\boldsymbol{x}/\|\boldsymbol{x}\|$, and $\mathsf{Law}(\boldsymbol{x}/\|\boldsymbol{x}\|) = \mathsf{Unif}(\mathbb{S}^{d-1})$.*

Each of these distributions makes the random vector easy to reason about in its "native" interpretation: the entries of an i.i.d. vector and the direction and magnitude of an orthogonally invariant vector are both described directly in the definitions. On the other hand, the information "across" interpretations is harder to access. We will build up some tools to address these questions now: what do the entries of $\boldsymbol{x} \sim \mathsf{Unif}(\mathbb{S}^{d-1})$ look like? And, what are the magnitude and direction of an i.i.d. random vector?

1.2 Gaussian Random Vectors

As we will see, in many ways the most natural of the i.i.d. random vectors is the standard multivariate Gaussian $\boldsymbol{x} \sim \mathcal{N}(0,1)^{\otimes d}$. We also denote this law in the more conventional multivariate Gaussian notation $\mathcal{N}(0,1)^{\otimes d} = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. The general meaning of this notation, which you should be familiar with, is as follows.

Definition 1.2.1. *$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for parameters $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}_{\succeq \mathbf{0}}$ is the probability measure with density*

$$\frac{1}{\sqrt{\det^*(2\pi\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^+ (\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{1.2.1}$$

relative to the Lebesgue measure on the row space of $\boldsymbol{\Sigma}$. Here, $\boldsymbol{\Sigma}^+$ is the Moore-Penrose pseudoinverse, and $\det^*$ is the product of all non-zero eigenvalues. When $\boldsymbol{\Sigma}$ is invertible, then $\det^*$ above is the ordinary determinant, $\boldsymbol{\Sigma}^+ = \boldsymbol{\Sigma}^{-1}$, and the above density is with respect to the Lebesgue measure on all of $\mathbb{R}^d$. We call $\boldsymbol{x}$ a Gaussian random vector *if $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for some $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and a* standard Gaussian random vector *if $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_d$.*

The following are the two fundamental properties of Gaussian random vectors.

Proposition 1.2.2. *The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and covariance matrix of the random vector $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:*

$$\boldsymbol{\mu} = \mathbb{E}\boldsymbol{x}, \tag{1.2.2}$$

$$\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top = \mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top \tag{1.2.3}$$

Thus, the law of a Gaussian random vector is determined by its mean and covariance, or equivalently by its linear and quadratic moments, $\mathbb{E}\boldsymbol{x}$ and $\mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top$.

Proposition 1.2.3. *If $\boldsymbol{x} \in \mathbb{R}^d$ is a Gaussian random vector and $\boldsymbol{A} \in \mathbb{R}^{d' \times d}$ and $\boldsymbol{b} \in \mathbb{R}^{d'}$, then $\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}$ is also a Gaussian random vector. If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the corresponding parameters after this linear transformation are*

$$\mathsf{Law}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top). \tag{1.2.4}$$

These two facts taken together make it easy to do certain computations with Gaussian random vectors: provided we only make linear transformations, we both remain within the class of Gaussian random vectors and can keep track of all distributions involved merely by linear algebra.[2]

The following remarkable facts about standard Gaussian random vectors then follow by simple calculations using the above. These show that, in fact, a standard Gaussian random vector is *both* an i.i.d. random vector *and* an orthogonally invariant one, thus bridging the two natural classes of models of vectors we were discussing before.

Proposition 1.2.4. *Suppose $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. For any $\boldsymbol{a} \in \mathbb{S}^{d-1}$, we have $\mathsf{Law}(\langle \boldsymbol{a}, \boldsymbol{x} \rangle) = \mathcal{N}(0, 1)$. In particular, this law does not depend on $\boldsymbol{a}$; the distribution of $\boldsymbol{x}$ is "the same in any direction." Further, $\boldsymbol{x}$ is orthogonally invariant.*

Proof. Using the previous results, for $\boldsymbol{Q} \in \mathcal{O}(d)$,

$$\mathsf{Law}(\boldsymbol{Q}\boldsymbol{x}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^\top \boldsymbol{I}_d \boldsymbol{Q}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d) = \mathsf{Law}(\boldsymbol{x}), \tag{1.2.5}$$

giving the second result. The first follows from the second by considering a $\boldsymbol{Q}$ whose first row is $\boldsymbol{a}$. □

Corollary 1.2.5. *If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, then $\mathsf{Law}(\boldsymbol{x}/\|\boldsymbol{x}\|) = \mathsf{Unif}(\mathbb{S}^{d-1})$.*

Proof. The result follows directly from the above together with Proposition 1.1.3. □

In fact, the following elegant result, which we will not prove here, shows that the Gaussian scalar measure is the *only* one having this property.

Theorem 1.2.6 (Maxwell)**.** *Suppose that $\boldsymbol{x} \sim \mu^{\otimes d}$ for some μ a probability measure on $\mathbb{R}$. If $d \geq 2$ and $\boldsymbol{x}$ is orthogonally invariant, then $\mu = \mathcal{N}(0, \sigma^2)$ for some $\sigma^2 \geq 0$.*

1.3 Concentration of Gaussian Vector Norms

The above tells us that the direction of a Gaussian random vector is uniformly distributed; in the exercises, you will see some similar more quantitative statements for more general classes of i.i.d. random vectors. What about the magnitudes of such vectors? Again, let us focus here on the Gaussian case.

A simple calculation gives the expectation of the norm (squared): if $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, then

$$\mathbb{E}\|\boldsymbol{x}\|^2 = d \cdot \mathbb{E}x_1^2 = d. \tag{1.3.1}$$

[2] We will not use it in this class, but another important fact to be aware of is that the more complicated operation of *conditioning* a Gaussian random vector $\boldsymbol{x}$ on the values of some linear forms $\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$ also yields another Gaussian random vector.

And, since $\|\boldsymbol{x}\|^2 = \sum_{i=1}^d x_i^2$ is a sum of i.i.d. random variables, the basic theory of such sums from scalar probability (law of large numbers, central limit theorem, etc.) tells us to expect that we should have $\|\boldsymbol{x}\|^2 \approx d$ with high probability; more precisely, we should have $\|\boldsymbol{x}\|^2 = d + O(\sqrt{d})$.

Let us first try to establish this using some basic concentration inequalities. Markov's inequality gives

$$\mathbb{P}[\|\boldsymbol{x}\|^2 \geq d + t] \leq \frac{d}{d+t} = \frac{1}{1+\frac{t}{d}}, \tag{1.3.2}$$

which only tells us that $\|\boldsymbol{x}\|^2 = O(d)$ with high probability (say, $\mathbb{P}[\|\boldsymbol{x}\|^2 \leq 100d] \geq 0.99$ and similar statements); the inequality "kicks in" and gives useful information only when $t \gtrsim d$. Chebyshev's inequality gives

$$\mathbb{P}[\|\boldsymbol{x}\|^2 \geq d + t] = \mathbb{P}[\|\boldsymbol{x}\|^2 \geq \mathbb{E}\|\boldsymbol{x}\|^2 + t] \leq \frac{\mathsf{Var}[\|\boldsymbol{x}\|^2]}{t^2} = \frac{2d}{t^2}, \tag{1.3.3}$$

where we calculate $\mathsf{Var}[\|\boldsymbol{x}\|^2] = d\,\mathsf{Var}[x_1^2] = 2d$. This gives the correct scaling, kicking in when $t \gtrsim \sqrt{d}$. However, as the following result shows, the actual dependence on t above is far from optimal.

Lemma 1.3.1. *For $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and any $t \geq 0$,*

$$\begin{aligned}
\mathbb{P}\left[\left|\|\boldsymbol{x}\|^2 - d\right| \geq t\right] &= \mathbb{P}\left[\left|\sum_{i=1}^d x_i^2 - d\right| \geq t\right] \\
&\leq 2\left\{\begin{array}{ll} \exp\left(-\frac{t^2}{8d}\right) & \text{if } t \leq d, \\ \exp\left(-\frac{t}{8}\right) & \text{if } t \geq d \end{array}\right\} \qquad (1.3.4) \\
&= 2\exp\left(-\frac{1}{8}\min\left\{\frac{t^2}{d}, t\right\}\right). \qquad (1.3.5)
\end{aligned}$$

Proof. The proof uses the "Chernoff method," a proof technique behind concentration inequalities like the Hoeffding, Bernstein, Azuma-Hoeffding, and McDiarmid inequalities. The method is a general instance of something you might call a "nonlinear Markov" inequality that also appears in the proofs of other bounds like Chebyshev's inequality. This whole family of approaches is important to know about.

We will only deal with one of the tails; the other follows similarly and the factor of 2 appears from taking a union bound. Note that $\mathbb{E}x_i^2 = 1$, so, defining $s_i := x_i^2 - 1$, we have

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^d x_i^2 - d \geq t\right] &= \mathbb{P}\left[\sum_{i=1}^d s_i \geq t\right] \\
&= \mathbb{P}\left[\exp\left(\lambda \sum_{i=1}^d s_i\right) \geq \exp(\lambda t)\right] \\
&\leq \frac{\mathbb{E}\exp\left(\lambda \sum_{i=1}^d s_i\right)}{\exp(\lambda t)} && \text{(Markov inequality)} \\
&= \frac{(\mathbb{E}\exp(\lambda s_1))^d}{\exp(\lambda t)} && \text{(independence)} \\
&= \exp\left(-\lambda t + d\log \mathbb{E}\exp(\lambda s_1)\right). \qquad (1.3.6)
\end{aligned}$$

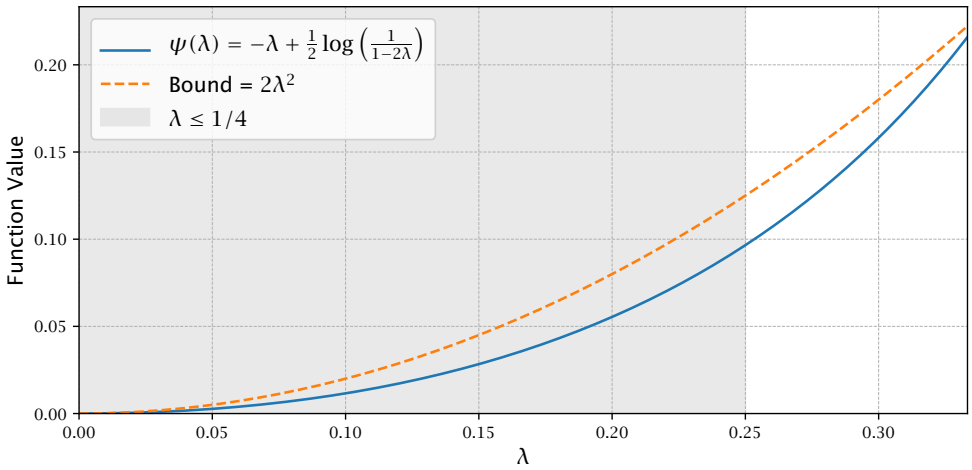

Figure 1.1: The cumulant generating function $\psi(\lambda)$ appearing in Lemma 1.3.1, our bound on it, and the region where we apply this bound.

Let us write

$$\psi(\lambda) := \log \mathbb{E} \exp(\lambda s_1) = \log \mathbb{E} \exp(\lambda(x_1^2 - 1)), \tag{1.3.7}$$

the *cumulant generating function* of the random variable $x_1^2 - 1$ (where we retain the -1 so that the expectation is zero. You may check as a calculus exercise that the remaining expectation is finite if and only if $\lambda < \frac{1}{2}$, and in this case

$$\mathbb{E} \exp(\lambda x_1^2) = \frac{1}{\sqrt{1-2\lambda}}, \tag{1.3.8}$$

whereby

$$\psi(\lambda) = -\lambda + \frac{1}{2}\log\left(\frac{1}{1-2\lambda}\right). \tag{1.3.9}$$

Taylor expansion shows that $\psi(\lambda) = \lambda^2 + O(\lambda)$. Turning this into a concrete bound, you may convince yourself by plotting (see Figure 1.1) and check by a slightly tedious algebra exercise that $\psi(\lambda) \le 2\lambda^2$ for all $\lambda \le 1/4$.

Thus we have that, provided $\lambda \le 1/4$,

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^d x_i^2 - d \ge t\right] &\le \exp\left(-\lambda t + d\psi(\lambda)\right) \\ &\le \exp\left(-\lambda t + 2d\lambda^2\right).\end{aligned}$$

Inside the exponential now remains a convex quadratic function of λ, which has a unique minimizer that we compute to be at $\lambda^* := \frac{t}{4d}$. We would like to set $\lambda := \lambda^*$, but we must be careful to respect that we have used a bound only holding for $\lambda \le \frac{1}{4}$. So, we consider two cases:

Case 1: $t \le d$. In this case, we may set $\lambda := \lambda^* \le \frac{1}{4}$. This gives

$$\mathbb{P}\left[\sum_{i=1}^d x_i^2 - d \ge t\right] \le \exp\left(-\frac{t^2}{8d}\right),$$

as claimed.

Case 2: $t \geq d$. In this case, we set $\lambda := \frac{1}{4}$. This gives

$$\mathbb{P}\left[\sum_{i=1}^{d} x_i^2 - d \geq t\right] \leq \exp\left(-\frac{t}{4} + \frac{d}{8}\right) \leq \exp\left(-\frac{t}{8}\right),$$

using that $t \geq d$. □

The heart of the calculation, you can convince yourself, is that $\psi(\lambda) = O(\lambda^2)$ for λ smaller than some constant. Note that the bound on λ is required, since $\psi(\lambda)$ becomes infinite for $\lambda \geq \frac{1}{2}$ in our case. This property, with the bound on λ, is called a random variable's being *subexponential*, a weaker version of the property of being *subgaussian* that you might have encountered before.

The result of the Lemma is characteristic of sums of i.i.d. subexponential random variables: they have "Gaussian tails" scaling as $\exp(-ct^2/d)$ up to a certain cutoff of $t \sim d$, beyond which they only have "exponential tails" scaling as $\exp(-ct)$. *Bernstein's inequality* is the general tool expressing this behavior; see Chapter 1 of [RH17] for more on that. A rough intuitive explanation for this is that, when x_i themselves have exponential tails, then large deviations of $\sum_{i=1}^{d} x_i$ of order $t \ll d$ are driven by the x_i each being slightly unusually large, while large deviations of order $t \gg d$ are driven by the largest of x_i being unusually large, whereby the tail behavior becomes the same as that of an individual x_i.

1.4 Consequences for High-Dimensional Geometry

Lemma 1.3.1 is one of the fundamental facts of high-dimensional geometry. Let us unpack its geometric meaning: informally, the result says that $\|\boldsymbol{x}\|^2 = d + O(\sqrt{d})$ with high probability. Taking square roots and a Taylor approximation, we see that

$$\|\boldsymbol{x}\| = \sqrt{d + O(\sqrt{d})} = \sqrt{d} \cdot \sqrt{1 + O\left(\frac{1}{\sqrt{d}}\right)} = \sqrt{d}\left(1 + O\left(\frac{1}{\sqrt{d}}\right)\right) = \sqrt{d} + O(1).$$

That is, a random standard Gaussian vector usually falls close to the spherical shell of width $O(1)$ around the sphere of radius $\sqrt{d}$.

This is quite counterintuitive if you have not seen it before: we think of a one- or two-dimensional Gaussian as having its "typical set" being a solid blob around the origin.[3] But, a high-dimensional Gaussian actually has a *non-convex* typical set, consisting of a hollow spherical shell![4]

Some other "paradoxes" of high-dimensional geometry that you might be aware of may be explained in the same way. For instance, consider $\boldsymbol{x} \sim \mathsf{Unif}([-1,+1])^{\otimes d}$, a uniformly random point in the solid box $[-1,+1]^d$. An integral calculation gives $\mathbb{E}\|\boldsymbol{x}\|^2 = d\mathbb{E}x_1^2 = d/3$. Similar concentration inequalities to the Gaussian case show that $\|\boldsymbol{x}\| = \sqrt{d/3} + O(1)$ with high probability. In particular, then, unlike in low dimension, almost all of the mass of this

[3] Let us not try to define a precise notion of typical set; think of it as meaning the smallest set in which a random variable falls with high probability.

[4] Sometimes such random vectors are called *thin-shell* in convex geometry and probability.

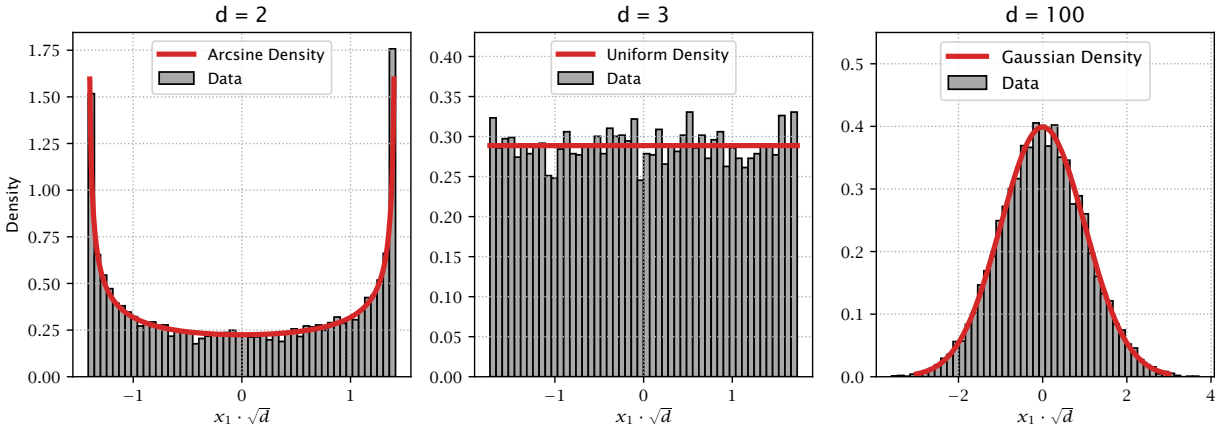

Figure 1.2: An illustration of Borel's theorem, Theorem 1.4.1, showing two low-dimensional examples $d = 2, 3$ where the distribution of $x_1 \cdot \sqrt{d}$ is known exactly (and is not Gaussian) as well as the Gaussian limit for large d.

box lies very close to its corners, far from the inscribed solid unit ball $\mathbb{B}^d := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq 1\}$. Using the analogous concentration inequality also implies that the ball $\mathbb{B}^d$ relative to its circumscribing box $[-1,+1]^d$ is bounded by

$$\begin{aligned} \frac{\mathsf{vol}(\mathbb{B}^d)}{\mathsf{vol}([-1,+1]^d)} &= \mathop{\mathbb{P}}_{\boldsymbol{x} \sim \mathsf{Unif}([-1,+1]^d)} [\|\boldsymbol{x}\| \leq 1] \\ &\leq \mathop{\mathbb{P}}_{\boldsymbol{x} \sim \mathsf{Unif}([-1,+1]^d)} [|\|\boldsymbol{x}\|^2 - d| \geq d - 1] \\ &= \exp(-\Omega(d)). \end{aligned} \tag{1.4.1}$$

Thus, these same probabilistic tools show that the volume of the unit sphere is exponentially smaller than the volume of its circumscribing box.[5]

Another way to look at our result is as implying the following approximation of laws, which we emphasize is an informal one:

$$\boxed{\text{“ } \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d) \approx \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d})) \text{ ”}} \tag{1.4.2}$$

We arrive at this heuristic by combining our observations that, when $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, then $\mathsf{Law}(\boldsymbol{x}/\|\boldsymbol{x}\|) = \mathsf{Unif}(\mathbb{S}^{d-1})$ (Corollary 1.2.5) and $\|\boldsymbol{x}\| \approx \sqrt{d}$ (Lemma 1.3.1). This idea can be made rigorous, for instance in the following result.

Theorem 1.4.1 (Borel's central limit theorem)**.** *Let $\boldsymbol{x} \sim \mathsf{Unif}(\mathbb{S}^{d-1})$. Then, $x_1 \cdot \sqrt{d}$ converges in distribution as $d \to \infty$ to $\mathcal{N}(0,1)$. In fact, for any fixed $k \geq 1$, $(x_1 \cdot \sqrt{d}, \ldots, x_k \cdot \sqrt{d})$ converges in distribution as $d \to \infty$ to $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$.*

[5] Actually, as you may look up, the true volume is even smaller than that, but to show that requires being more careful than simply mimicking the proof of the bound for the Gaussian case in Lemma 1.3.1.

2 | First Steps With Rectangular Matrices

We now take our first steps into random matrix theory, using the results from the previous chapter to study the behavior of asymmetric or rectangular matrices with i.i.d. Gaussian entries. In particular, we will focus on the random matrix model

$$\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d \times m}, \tag{2.0.1}$$

a $d \times m$ random matrix with i.i.d. Gaussian entries.

2.1 Gaussian Random Field Viewpoint

Let us start by trying to gain some intuition about what $\boldsymbol{G}$ does to a single $\boldsymbol{y} \in \mathbb{R}^m$ as well as several given $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \in \mathbb{R}^m$. Immediately our observations about Gaussian random vectors become very useful, since we can easily completely characterize the laws of these images.

Proposition 2.1.1. $\mathsf{Law}(\boldsymbol{G}\boldsymbol{y}) = \mathcal{N}(\boldsymbol{0}, \|\boldsymbol{y}\|^2 \boldsymbol{I}_d)$.

Proof. Since $\boldsymbol{G}\boldsymbol{y}$ is linear in $\boldsymbol{G}$, whose entries form a Gaussian random vector, $\boldsymbol{G}\boldsymbol{y}$ is itself a Gaussian random vector. So, it suffices to compute its mean and covariance:

$$\begin{aligned}
\mathbb{E}(\boldsymbol{G}\boldsymbol{y})_i &= \mathbb{E} \sum_{a=1}^{m} G_{ia} y_a \\
&= 0, \\
\mathsf{Cov}((\boldsymbol{G}\boldsymbol{y})_i, (\boldsymbol{G}\boldsymbol{y})_j) &= \mathbb{E}(\boldsymbol{G}\boldsymbol{y})_i (\boldsymbol{G}\boldsymbol{y})_j \\
&= \mathbb{E} \left(\sum_{a=1}^{m} G_{ia} y_a \right) \left(\sum_{b=1}^{m} G_{jb} y_b \right) \\
&= \sum_{a,b=1}^{m} y_a y_b \cdot \mathbb{E} G_{ia} G_{jb} \\
&= \left\{ \begin{array}{ll} 0 & \text{if } i \neq j, \\ \|\boldsymbol{y}\|^2 & \text{if } i = j \end{array} \right\},
\end{aligned}$$

where we have used that the G_{ia} are i.i.d. and have mean zero.

Another approach is to work with matrices; it is good to get used to using matrices inside of expectations in the way written below. For the mean, the following operation is valid in general when an expectation of a matrix product only has one random matrix involved:

$$\mathbb{E}[\boldsymbol{G}\boldsymbol{y}] = \mathbb{E}[\boldsymbol{G}]\boldsymbol{y} = \boldsymbol{0}. \tag{2.1.1}$$

You may check that this is just another form of the linearity of expectation. For the covariance, if we write $\boldsymbol{g}_1, \dots, \boldsymbol{g}_m \in \mathbb{R}^d$ for the columns of $\boldsymbol{G}$, so that $\boldsymbol{g}_a \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ are i.i.d. standard Gaussian random vectors, we may treat the entire covariance using similar manipulations:

$$\begin{aligned}
\mathsf{Cov}(\boldsymbol{G}\boldsymbol{y}) &= \mathbb{E}(\boldsymbol{G}\boldsymbol{y})(\boldsymbol{G}\boldsymbol{y})^\top \\
&= \mathbb{E}\left(\sum_{a=1}^m y_a \boldsymbol{g}_a\right)\left(\sum_{b=1}^m y_b \boldsymbol{g}_b^\top\right) \\
&= \sum_{a,b=1}^m y_a y_b \cdot \mathbb{E}\boldsymbol{g}_a \boldsymbol{g}_b^\top \\
&= \sum_{a=1}^m y_a^2 \cdot \boldsymbol{I}_d \\
&= \|\boldsymbol{y}\|^2 \boldsymbol{I}_d,
\end{aligned}$$

now using that the $\boldsymbol{g}_a$ are i.i.d. and have mean zero and covariance identity. □

By the same kinds of calculations, we may also deduce the joint law of any collection of finitely many $\boldsymbol{G}\boldsymbol{y}_1, \dots, \boldsymbol{G}\boldsymbol{y}_n$.

Proposition 2.1.2. *Let $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \in \mathbb{R}^m$. Then,*

$$\mathsf{Law}\left(\begin{bmatrix} \boldsymbol{G}\boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{G}\boldsymbol{y}_n \end{bmatrix}\right) = \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \|\boldsymbol{y}_1\|^2 \boldsymbol{I}_d & \langle \boldsymbol{y}_1, \boldsymbol{y}_2\rangle \boldsymbol{I}_d & \cdots & \langle \boldsymbol{y}_1, \boldsymbol{y}_n\rangle \boldsymbol{I}_d \\ \langle \boldsymbol{y}_2, \boldsymbol{y}_1\rangle \boldsymbol{I}_d & \|\boldsymbol{y}_2\|^2 \boldsymbol{I}_d & \cdots & \langle \boldsymbol{y}_2, \boldsymbol{y}_n\rangle \boldsymbol{I}_d \\ \vdots & \vdots & \ddots & \vdots \\ \langle \boldsymbol{y}_n, \boldsymbol{y}_1\rangle \boldsymbol{I}_d & \langle \boldsymbol{y}_n, \boldsymbol{y}_2\rangle \boldsymbol{I}_d & \cdots & \|\boldsymbol{y}_n\|^2 \boldsymbol{I}_d \end{bmatrix}\right). \tag{2.1.2}$$

Written more concisely, if $\boldsymbol{Y} \in \mathbb{R}^{m\times n}$ has the $\boldsymbol{y}_i$ as its columns, then the covariance matrix above is $(\boldsymbol{Y}^\top \boldsymbol{Y}) \otimes \boldsymbol{I}_d$.

Proof. The result follows from identical calculations to the previous ones in each block of the covariance matrix. □

The previous result fully characterizes the *Gaussian random field* (a term for a higher-dimensional Gaussian process) that attaches the random vector $\boldsymbol{G}\boldsymbol{y} \in \mathbb{R}^d$ to each point $\boldsymbol{y} \in \mathbb{R}^m$, since we have characterized the joint law of any finitely many points of this field.

From this characterization, we may start to probe more particular geometric properties of the action of $\boldsymbol{G}$. We may, for example, compute what multiplication by $\boldsymbol{G}$ does to expected lengths:

$$\mathbb{E}\|\boldsymbol{G}\boldsymbol{y}\|^2 = \mathsf{Tr}(\mathsf{Cov}(\boldsymbol{G}\boldsymbol{y})) = d\|\boldsymbol{y}\|^2. \tag{2.1.3}$$

What about angles? By a similar calculation you can find

$$\mathbb{E}\langle \boldsymbol{G}\boldsymbol{y}_1, \boldsymbol{G}\boldsymbol{y}_2\rangle = \mathsf{Tr}(\mathsf{Cov}(\boldsymbol{G}\boldsymbol{y}_1, \boldsymbol{G}\boldsymbol{y}_2)) = d\langle \boldsymbol{y}_1, \boldsymbol{y}_2\rangle, \tag{2.1.4}$$

(Actually, this also follows directly from the calculation for distances by the polarization identity $\langle \boldsymbol{y}_1, \boldsymbol{y}_2\rangle = \frac{1}{4}\|\boldsymbol{y}_1 + \boldsymbol{y}_2\|^2 - \frac{1}{4}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|^2$. Try working it out.) Indeed, $\boldsymbol{G}$ acts in the same way on the entire *Gram matrix* of any finite collection of vectors in expectation: for $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \in \mathbb{R}^m$ organized into the columns of $\boldsymbol{Y}$, we have

$$\mathbb{E}(\boldsymbol{G}\boldsymbol{Y})^\top(\boldsymbol{G}\boldsymbol{Y}) = d\boldsymbol{Y}^\top\boldsymbol{Y}. \tag{2.1.5}$$

Remark 2.1.3 (Gram matrices)**.** *The Gram matrix is a fundamental geometric object that does not always get its due in a linear algebra class. For a set of vectors $\boldsymbol{y}_i$ as above, $\boldsymbol{Y}^\top\boldsymbol{Y}$ describes the entire relative geometry of this set. Looking at the entries, we see that the Gram matrix contains the lengths of the $\boldsymbol{y}_i$ (on the diagonal) and the angles between each pair (on the off-diagonal). Actually, this information fully specifies the geometry of this set of vectors: if any other $\boldsymbol{y}_1', \dots, \boldsymbol{y}_n' \in \mathbb{R}^m$ have the same Gram matrix, then there must be an orthogonal $\boldsymbol{Q} \in \mathcal{O}(m)$ such that $\boldsymbol{Q}\boldsymbol{y}_i = \boldsymbol{y}_i'$; that is, it is possible to get from one point cloud to the other by rotations and reflections.*

The above calculations suggest that we should normalize away the factors of d throughout by considering instead

$$\widehat{\boldsymbol{G}} := \frac{1}{\sqrt{d}}\boldsymbol{G}, \quad \mathsf{Law}(\widehat{\boldsymbol{G}}) = \mathcal{N}\left(0, \frac{1}{d}\right)^{\otimes d\times m}. \tag{2.1.6}$$

This matrix will satisfy the appealing properties

$$\mathbb{E}\|\widehat{\boldsymbol{G}}\boldsymbol{y}\|^2 = \|\boldsymbol{y}\|^2, \tag{2.1.7}$$

$$\mathbb{E}\langle \widehat{\boldsymbol{G}}\boldsymbol{y}_1, \widehat{\boldsymbol{G}}\boldsymbol{y}_2\rangle = \langle \boldsymbol{y}_1, \boldsymbol{y}_2\rangle, \tag{2.1.8}$$

$$\mathbb{E}(\widehat{\boldsymbol{G}}\boldsymbol{Y})^\top(\widehat{\boldsymbol{G}}\boldsymbol{Y}) = \boldsymbol{Y}^\top\boldsymbol{Y}. \tag{2.1.9}$$

That is, in expectation, $\widehat{\boldsymbol{G}}$ preserves lengths, angles, and Gram matrices, and thus looks as if it preserves all of the geometry of $\mathbb{R}^m$! We will see in short order that this is true in a specific quantitative sense with high probability as well, making multiplication by $\widehat{\boldsymbol{G}}$ a very useful algorithmic and computational tool.

First, though, let us clarify an important caveat. Again, the above makes it look like $\widehat{\boldsymbol{G}}$ preserves the relative geometry of any given point cloud $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \in \mathbb{R}^m$ while mapping it to $\mathbb{R}^d$. But, this holds for any d, including $d \ll m$, and even $d = 1$! Thus, part of the above optimistic intuition is a kind of illusion created by taking expectations.

There are two important cautions about this intuition. First, once $d \ll m$, once we draw $\widehat{\boldsymbol{G}}$ at random, of course we can find $\boldsymbol{y} \neq \boldsymbol{0}$ such that $\widehat{\boldsymbol{G}}\boldsymbol{y} = \boldsymbol{0}$ (any $\boldsymbol{y}$ in the kernel of $\widehat{\boldsymbol{G}}$), and for such $\boldsymbol{y}$ we will have $0 = \|\widehat{\boldsymbol{G}}\boldsymbol{y}\| \ll \|\boldsymbol{y}\|$. Thus, $\widehat{\boldsymbol{G}}$ cannot and will not preserve the relative geometry of *any* vectors we multiply it by, and in particular we can easily find vectors *depending* on $\widehat{\boldsymbol{G}}$ whose relative geometry it badly distorts. On the other hand, if we *first* pick $\boldsymbol{y}$ and then draw a random $\widehat{\boldsymbol{G}}$, then we have reason to hope that $\|\widehat{\boldsymbol{G}}\boldsymbol{y}\|$ will concentrate

around the value $\|\boldsymbol{y}\|$ suggested by the previous calculations; indeed, our calculations show that $\mathrm{Law}(\widehat{\boldsymbol{G}}\boldsymbol{y}) = \mathcal{N}(\boldsymbol{0}, \frac{1}{d}\|\boldsymbol{y}\|^2\boldsymbol{I}_d)$, to which Lemma 1.3.1 should apply.

Second, even provided we make sure $\widehat{\boldsymbol{G}}$ is independent of the $\boldsymbol{y}$ that we multiply by it, the extent to which the above properties hold with high probability rather than merely in expectation will depend on how large d is: when $d = 1$, then $\widehat{\boldsymbol{G}} = \boldsymbol{g}^\top$ for $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, and we still have $\mathbb{E}\|\widehat{\boldsymbol{G}}\boldsymbol{y}_i\|^2 = \mathbb{E}\langle \boldsymbol{g}, \boldsymbol{y}_i\rangle^2 = \|\boldsymbol{y}_i\|^2$. But, it should be intuitively clear that we cannot actually have $\langle \boldsymbol{g}, \boldsymbol{y}_i\rangle^2 \approx \|\boldsymbol{y}_i\|^2$ simultaneously for very many $\boldsymbol{y}_i$, even if we choose them in advance of drawing $\widehat{\boldsymbol{G}}$. (Exercise: Try constructing concrete counterexamples.)

We will see next that what we have mentioned above are the only obstructions to $\widehat{\boldsymbol{G}}$ behaving as we wish: provided that (1) d is not too small and (2) that we fix a point cloud $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ in advance of drawing $\widehat{\boldsymbol{G}}$, we will have that multiplication by $\widehat{\boldsymbol{G}}$ approximately preserves the relative geometry of the $\boldsymbol{y}_i$ (that is, $\widehat{\boldsymbol{G}}$ acts as an *approximate isometry* on these points), even for $d \ll m$.

2.2 Application: Dimensionality Reduction and Johnson-Lindenstrauss Lemma

Validating our heuristic reasoning above, the Johnson-Lindenstrauss lemma states that, in a certain sense, a deterministic point cloud that is typically deformed substantially by $\widehat{\boldsymbol{G}}$ *must* be quite large relative to the dimension we are mapping it into, so the above obstruction to $\widehat{\boldsymbol{G}}$ preserving geometric structure is the only one. However, and this is an important condition that we will return to later, this result only concerns whether $\widehat{\boldsymbol{G}}$ deforms *pairwise* distances between the $\boldsymbol{y}_i$. This is not the same as preserving all aspects of the global geometry of the $\boldsymbol{y}_i$, but is an intuitive notion and is relevant to various applications.

Definition 2.2.1. *Given $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \in \mathbb{R}^m$, a function $f : \mathbb{R}^m \to \mathbb{R}^d$ is* pairwise ϵ-faithful[1] *on the $\boldsymbol{y}_i$ if, for all $i, j \in [n]$,*

$$(1-\epsilon)\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 \le \|f(\boldsymbol{y}_i) - f(\boldsymbol{y}_j)\|^2 \le (1+\epsilon)\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2. \tag{2.2.1}$$

Theorem 2.2.2 (Johnson-Lindenstrauss [JL82, IM98])**.** *Let $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \in \mathbb{R}^m$ be arbitrary and $\epsilon \in (0,1)$. Suppose that*

$$d \ge 24\frac{\log n}{\epsilon^2}. \tag{2.2.2}$$

Note that the lower bound on d does not depend on the dimension m at all! *Then, with probability at least $1 - \frac{1}{n}$, multiplication by $\widehat{\boldsymbol{G}} \sim \mathcal{N}(0, \frac{1}{d})^{\otimes d \times m}$ is pairwise ϵ-faithful on the $\boldsymbol{y}_i$.*

In fact, as the proof will make clear, by increasing the constant C in the assumption $d \ge C\frac{\log n}{\epsilon^2}$ to some $C = C(K)$, you can achieve a success probability $1 - \frac{1}{n^K}$ for any given $K > 0$.

Proof. The proof actually has little to do with the $\boldsymbol{y}_i$. Notice that being pairwise ϵ-faithful is a matter of preserving $\binom{n}{2}$ many vector norms to within a factor in $[1-\epsilon, 1+\epsilon]$. We will

[1] My own non-standard terminology.

show that this is true with high probability for *any* choice of $\binom{n}{2}$ vectors, and therefore also for the $\widehat{G}(\boldsymbol{y}_i - \boldsymbol{y}_j) = \widehat{G}\boldsymbol{y}_i - \widehat{G}\boldsymbol{y}_j$.

Consider an arbitrary $\boldsymbol{x} \in \mathbb{R}^m$. Recall that $\mathsf{Law}(\widehat{G}\boldsymbol{x}) = \mathcal{N}(0, \frac{1}{d}\|\boldsymbol{x}\|^2 \boldsymbol{I}_d)$. We then have

$$\begin{aligned}
\mathop{\mathbb{P}}_{\widehat{G}\sim\mathcal{N}(0,\frac{1}{d})^{\otimes d\times m}}\left[\,\big|\,\|\widehat{G}\boldsymbol{x}\|^2 - \|\boldsymbol{x}\|^2\,\big| > \epsilon\|\boldsymbol{x}\|^2\right] &= \mathop{\mathbb{P}}_{\boldsymbol{g}\sim\mathcal{N}(0,\frac{1}{d}\|\boldsymbol{x}\|^2)}\left[\,\big|\,\|\boldsymbol{g}\|^2 - \|\boldsymbol{x}\|^2\,\big| > \epsilon\|\boldsymbol{x}\|^2\right] \\
&= \mathop{\mathbb{P}}_{\boldsymbol{g}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\left|\,\frac{1}{d}\|\boldsymbol{x}\|^2\cdot\|\boldsymbol{g}\|^2 - \|\boldsymbol{x}\|^2\,\right| > \epsilon\|\boldsymbol{x}\|^2\right] \\
&= \mathop{\mathbb{P}}_{\boldsymbol{g}\sim\mathcal{N}(0,\boldsymbol{I}_d)}\left[\,\big|\,\|\boldsymbol{g}\|^2 - d\,\big| > \epsilon d\right]
\end{aligned}$$

and, using that $\epsilon < 1$ whereby $\epsilon d < d$, by Lemma 1.3.1 we have

$$\leq 2\exp\left(-\frac{1}{8}\epsilon^2 d\right). \tag{2.2.3}$$

This gives a bound on the probability of embedding a single vector with low distortion of the length:

$$\mathop{\mathbb{P}}_{\widehat{G}\sim\mathcal{N}(0,\frac{1}{d})^{\otimes d\times m}}\left[(1-\epsilon)\|\boldsymbol{x}\|^2 \leq \|\widehat{G}\boldsymbol{x}\|^2 \leq (1+\epsilon)\|\boldsymbol{x}\|^2\right] \geq 1 - 2\exp\left(-\frac{1}{8}\epsilon^2 d\right). \tag{2.2.4}$$

Finally, let E_{ij} be the event that $(1-\epsilon)\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 \leq \|\widehat{G}\boldsymbol{y}_i - \widehat{G}\boldsymbol{y}_j\|^2 \leq (1+\epsilon)\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2$. Then, by the union bound we have that

$$\begin{aligned}
\mathbb{P}[\text{some } E_{ij} \text{ does not occur}] &\leq \binom{n}{2}\cdot 2\exp\left(-\frac{1}{8}\epsilon^2 d\right) \\
&\leq n^2\exp\left(-\frac{1}{8}\epsilon^2 d\right) \\
&= \frac{1}{n}\exp\left(3\log n - \frac{1}{8}\epsilon^2 d\right),
\end{aligned} \tag{2.2.5}$$

and the result follows since, if $d \geq 24\frac{\log n}{\epsilon^2}$, then this is at most $1/n$. □

Multiplication by a random matrix for the purposes of dimensionality reduction is sometimes called the *Johnson-Lindenstrauss transform (JLT)* after this result. The original JLT and its algorithmic applications sparked many directions of further research. Let us briefly go over a few of these.

2.2.1 Simple Extensions

It is easy to extend the proof to show that we also have $(1-\epsilon)\|\boldsymbol{y}_i\| \leq \|f(\boldsymbol{y}_i)\| \leq (1+\epsilon)\|\boldsymbol{y}_i\|$ with the same high probability up to constants (say, by increasing n by one and adding $\boldsymbol{0}$ to the $\boldsymbol{y}_i$, or by including further appropriate "bad events" in the union bound argument above). As mentioned before, it is also possible to get an error bound of $1/n^K$ for any $K > 0$ by suitably increasing the constant 24 in the statement of the Theorem and adjusting the last step accordingly.

2.2.2 Lower Bounds

It is natural to ask if the result is optimal: could it be possible to reduce d even further? In fact it is not. The original paper of Johnson and Lindenstrauss [JL82] showed that dimension $\Omega(\log n)$ is needed for any given constant ϵ, but with no dependence on ϵ. Only relatively recently was it shown that the Johnson-Lindenstrauss result is actually tight up to constants. This was first shown for embeddings produced by applying a linear map to the $\boldsymbol{y}_i$ (i.e., multiplying them by a matrix) [LN14], and finally extended to arbitrary embeddings [LN17].

2.2.3 Sparse and Fast JLTs

Another line of work has explored how to speed up the matrix-vector multiplication required to form the embeddings $f(\boldsymbol{y}_i) = \widehat{\boldsymbol{G}}\boldsymbol{y}_i$. All such work is basically predicated on finding special structured matrices $\widehat{\boldsymbol{G}}$ that allow for faster matrix-vector multiplication. The simplest such structure is *sparsity* of the matrix involved; see [DKS10, KN14] for some results in this direction. Perhaps the most elegant idea studied in this direction is quite different: the algorithm of [AC09] uses that the fast Fourier transform algorithm may be used to multiply quickly by the discrete Fourier transform matrix. Their version of the JLT is based on performing this multiplication (which does not reduce the dimension of the vector, as the discrete Fourier transform matrix is square) and then subsampling a random subset of entries of the result. There is a further caveat that some vectors have sparse Fourier transforms, so subsampling their Fourier transform runs a risk of ending up with an unusually small vector. Thus, a further preprocessing step is required to avoid this particular set of "bad" vectors; see the paper for more details.

2.2.4 Proof Technique: First Moment Method

As we have seen, the proof technique is merely an application of the union bound. However, it is perhaps worth emphasizing some aspects of how specifically we are using the union bound that also hold in many modern applications. While the union bound might seem trivial, and indeed is when applied to just a few events, the union bound when applied to many carefully chosen events is actually a quite sophisticated tool.[2] If one wishes to bound the probability of a "bad event" E, one may decompose it into $E = E_1 \cup \cdots \cup E_N$ and bound

$$\mathbb{P}[E] \leq \sum_{i=1}^{N} \mathbb{P}[E_i] \leq N \cdot \max_{i=1}^{N} \mathbb{P}[E_i]. \tag{2.2.6}$$

Often, the $\mathbb{P}[E_i]$ are roughly the same. Then, the above becomes a clear "competition" between N, the number of ways that a bad event can happen, and $\mathbb{P}[E_i]$, the probability of a given way that it can happen.[3] In high-dimensional probability, often N is very large and

[2] See in particular the book [Tal05], whose results show that the values of a tremendous number of random optimization-like quantities including the norms of very general Gaussian random matrices can, in principle, be controlled accurately by nothing more than a sufficiently careful use of the union bound.

[3] Borrowing the language of statistical physics, these quantities are sometimes called "entropy" and "energy," respectively.

$\mathbb{P}[E_i]$ is very small, leading to the kind of competition on an exponential scale that we saw in the proof.

As an aside, it is also worth being aware that this approach is also often called the "first moment method," because one may phrase the bound as

$$\mathbb{P}[\text{some } E_i \text{ occurs}] = \mathbb{P}\left[\sum_{i=1}^{N} \mathbb{1}\{E_i \text{ occurs}\} \geq 1\right] \leq \mathbb{E}\left[\sum_{i=1}^{N} \mathbb{1}\{E_i \text{ occurs}\}\right] = \sum_{i=1}^{N} \mathbb{P}[E_i] \tag{2.2.7}$$

by Markov's inequality. Thus one is computing the first moment (or expectation) of the random variable $\#\{i : E_i \text{ occurs}\}$. There is also a widely used "second moment method" which involves computing the second moment and thus variance of this random variable, using which you can show that some E_i *does* occur with high probability. For instance, you could use this to show that, for a suitable point cloud, if d is too small then multiplication by $\widehat{\boldsymbol{G}}$ is *not* pairwise ϵ-faithful. (Note that this does not answer the question about lower bounds posed above, since it only applies to this particular method of embedding by multiplication by a random matrix.)

2.3 Rectangular Gaussian Matrices: Singular Values

Let us next see how to extend our approach to the JLT to understand more intrinsic properties of $\widehat{\boldsymbol{G}}$, namely its singular values and vectors. This will both give us a first treatment of our first fundamental random matrix model—that of rectangular matrices with i.i.d. Gaussian entries—and will hopefully further clarify why the JLT works well. We focus on "short fat" matrices, that is, the setting $d \leq m$ (often we will also mention how our results specialize to the asymptotic regime $d \ll m$). We also return to working with $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d \times m}$ again in order to simplify the notation.

As we have mentioned, the same argument as used in the proof of Theorem 2.2.2 actually implies the following more general result.

Theorem 2.3.1. *There is a constant $C > 0$ such that the following holds. For any $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^m$, if $d \geq C\frac{\log N}{\epsilon^2}$ and $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d \times m}$, then*

$$\mathbb{P}\left[\left|\|\boldsymbol{G}\boldsymbol{x}_i\|^2 - d\|\boldsymbol{x}_i\|^2\right| \leq \epsilon d\|\boldsymbol{x}_i\|^2 \text{ for all } i \in [N]\right] \geq 1 - \frac{1}{N}. \tag{2.3.1}$$

Before, we took $N \asymp n^2$ and the $\boldsymbol{x}_i$ to be all possible $\boldsymbol{y}_j - \boldsymbol{y}_k$, but we can apply the same tools to any $\boldsymbol{x}_i$.

We may reframe this by rewriting

$$\left|\|\boldsymbol{G}\boldsymbol{x}_i\|^2 - \|\boldsymbol{x}_i\|^2\right| = \left|\boldsymbol{x}_i^\top(\boldsymbol{G}^\top\boldsymbol{G})\boldsymbol{x}_i - \boldsymbol{x}_i^\top(d\boldsymbol{I}_m)\boldsymbol{x}_i\right| = \left|\boldsymbol{x}_i^\top(\boldsymbol{G}^\top\boldsymbol{G} - d\boldsymbol{I}_m)\boldsymbol{x}_i\right|. \tag{2.3.2}$$

Thus it seems that, from the "point of view" of a small number of deterministic quadratic form evaluations, the random matrix $\boldsymbol{G}^\top\boldsymbol{G}$ behaves like $d\boldsymbol{I}_m$, and indeed the JLT is based on this fact. Had we stuck with our previous normalization, we would see the different normalization that $\widehat{\boldsymbol{G}}^\top\widehat{\boldsymbol{G}}$ behaves like $\boldsymbol{I}_m$. In any case, this of course cannot be true without

restricting the number of quadratic form evaluations, since the matrices $\boldsymbol{G}^\top \boldsymbol{G}$ and $\widehat{\boldsymbol{G}}^\top \widehat{\boldsymbol{G}}$ have rank at most $d \ll m$.

This is another "paradox" (that we will resolve in short order) of high-dimensional probability and geometry: somehow, a very low-rank random matrix can actually act like the identity against a large number of test quadratic forms. This is basically just a variation on the same paradox we saw in $\widehat{\boldsymbol{G}}$ preserving lengths and angles in expectation, but note that the above statement does actually state that an event happens with high probability, so it cannot merely be accounted for by saying that the random variables involved are not close enough to their expectations.

We will now work towards a more sophisticated and "more spectral" (or more intrinsic and geometric) explanation of this phenomenon than the union bound based proof we saw, which is relatively simple but maybe not very enlightening. To do that, we will undertake an analysis of the eigendecomposition of the matrix $\boldsymbol{G}^\top \boldsymbol{G} \in \mathbb{R}^{m \times m}_{\mathsf{sym}}$, seeking to make claims about the distribution of its eigenvectors and eigenvalues (the same information as the right singular vectors and singular values of $\boldsymbol{G}$).

Note that $\boldsymbol{G}^\top \boldsymbol{G} \succeq \boldsymbol{0}$, so its eigenvalues are non-negative. Call these $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$. Since $\mathsf{rank}(\boldsymbol{G}^\top \boldsymbol{G}) \leq d$, we have $\lambda_{d+1} = \cdots = \lambda_m = 0$. The remaining eigenvalues are related to the singular values of $\boldsymbol{G}$ by $\lambda_i = \sigma_i(\boldsymbol{G})^2$. We begin by studying these eigenvalues, or equivalently the singular values of $\boldsymbol{G}$.

2.3.1 The **Geometric Method** of Random Matrix Theory

It will be more convenient to instead consider the eigenvalues of $\boldsymbol{G}\boldsymbol{G}^\top \in \mathbb{R}^{d \times d}_{\mathsf{sym}}$. Since the eigenvalues of $\boldsymbol{G}\boldsymbol{G}^\top$ and $\boldsymbol{G}^\top \boldsymbol{G}$ are equal except that the latter has $m - d$ extra eigenvalues equal to 0, $\boldsymbol{G}\boldsymbol{G}^\top$ will provide us with the same information; at the end, we will return to $\boldsymbol{G}^\top \boldsymbol{G}$ and deduce some geometric intuitions about it. The probabilistic reason for working with $\boldsymbol{G}\boldsymbol{G}^\top$ instead is that, if $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_m \in \mathbb{R}^d$ are the columns of $\boldsymbol{G}$ (distributed i.i.d. as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$), then we have

$$\boldsymbol{G}\boldsymbol{G}^\top = \sum_{i=1}^{m} \boldsymbol{g}_i \boldsymbol{g}_i^\top$$

which is a sum of $m \gg d$ i.i.d. random matrices of dimension d. As we will see, in such a situation it is justified to hope for a matrix-valued law of large numbers to hold, giving us

$$\begin{aligned} &\approx m \cdot \mathbb{E}\boldsymbol{g}_1 \boldsymbol{g}_1^\top \\ &= m\boldsymbol{I}_d \\ &= \mathbb{E}\boldsymbol{G}\boldsymbol{G}^\top. \end{aligned} \tag{2.3.3}$$

Roughly speaking, this heuristic is more accurate than the same applied to $\boldsymbol{G}^\top \boldsymbol{G}$ where the roles of d and m are reversed because we want in our appeal to the law of large numbers to work with an expression having "as much averaging as possible."

Remark 2.3.2. *To emphasize the difference in how we are discussing these two matrices: above we observed that* $\boldsymbol{G}^\top \boldsymbol{G}$ looks like $d\boldsymbol{I}_m$ *from the point of view of a small enough number*

of quadratic forms, but we also observed that it cannot actually *be close to this matrix (for instance by simple rank considerations). On the other hand, the above heuristic suggests that $\boldsymbol{G}\boldsymbol{G}^\top$ looks like $m\boldsymbol{I}_d$, and we will show below that this is actually true, and use that to understand the trickier behavior of $\boldsymbol{G}^\top\boldsymbol{G}$.*

Concretely, we expect the difference of operator norms $\|\boldsymbol{G}\boldsymbol{G}^\top - m\boldsymbol{I}_d\|$ to be small. This kind of statement that a sum of independent or weakly dependent random matrices is close in operator norm to its expectation is often called a *matrix concentration inequality*. Let us write $\boldsymbol{\Delta} := \boldsymbol{G}\boldsymbol{G}^\top - m\boldsymbol{I}_d$.

Remark 2.3.3 (Statistical interpretation)**.** *Expanding in a sum of rank one terms again, we may also view*

$$\frac{1}{m}\boldsymbol{\Delta} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{g}_i\boldsymbol{g}_i^\top - \mathbb{E}\boldsymbol{g}_1\boldsymbol{g}_1^\top \tag{2.3.4}$$

as the difference between a sample covariance *and the true covariance of several i.i.d. draws $\boldsymbol{g}_1, \dots, \boldsymbol{g}_m$ from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Thus, our investigation will also address the natural statistical question of how good of a statistical estimator the sample covariance is (in this particular case).*

How can we show that $\|\boldsymbol{\Delta}\|$ is relatively small? The chief trouble is that the operator norm is a quantity whose definition involves an uncountable quantification over test vectors. We adopt the following notation for closed balls in $\mathbb{R}^d$, which will be useful soon:

$$\mathbb{B}^d(\boldsymbol{x}, r) := \{\boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{y} - \boldsymbol{x}\| \le r\}, \tag{2.3.5}$$

$$\mathbb{B}^d := \mathbb{B}^d(\boldsymbol{0}, 1) \supset \mathbb{S}^{d-1}. \tag{2.3.6}$$

The operator norm can be formulated in either of the equivalent ways

$$\|\boldsymbol{\Delta}\| = \sup_{\boldsymbol{x}\in\mathbb{S}^{d-1}} |\boldsymbol{x}^\top\boldsymbol{\Delta}\boldsymbol{x}| = \sup_{\boldsymbol{x}\in\mathbb{B}^d} |\boldsymbol{x}^\top\boldsymbol{\Delta}\boldsymbol{x}|. \tag{2.3.7}$$

In either case, if $\mathbb{S}^{d-1}$ or $\mathbb{B}^d$ were replaced by a finite set, we could use the first moment method, bounding $\mathbb{P}[|\boldsymbol{x}^\top\boldsymbol{\Delta}\boldsymbol{x}| > t]$ (which as we will see boils down to yet another application of Lemma 1.3.1 on the concentration of norms of Gaussian random vectors) and using a union bound to control $\mathbb{P}[\|\boldsymbol{\Delta}\| > t]$, but this is not the case.

But, it will turn out to be useful to take this hypothetical situation more seriously, so let us define notation for these notions.

Definition 2.3.4. *For any bounded set $\mathcal{X} \subset \mathbb{R}^d$ and $\boldsymbol{M} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$, we define*

$$\|\boldsymbol{M}\|_{\mathcal{X}} := \sup_{\boldsymbol{x}\in\mathcal{X}} |\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x}|. \tag{2.3.8}$$

This is not always actually a norm; depending on $\mathcal{X}$, it may only be a *seminorm*, where $\|\boldsymbol{M}\|_{\mathcal{X}} = 0$ can hold for $\boldsymbol{M} \neq \boldsymbol{0}$. The definition will be useful in any case, however. Applying the first moment method to a general random matrix gives the following:

Proposition 2.3.5. *Let $\boldsymbol{M} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ be a random matrix and $\mathcal{X} \subset \mathbb{R}^d$ be a finite set. Then,*

$$\mathbb{P}[\|\boldsymbol{M}\|_{\mathcal{X}} > t] \le |\mathcal{X}| \cdot \max_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}[|\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x}| > t]. \tag{2.3.9}$$

We will apply this to $\boldsymbol{M} = \boldsymbol{\Delta}$. As mentioned above, controlling the second factor above involving scalar probabilities will be possible using Lemma 1.3.1. Thus, the key point will be the tradeoff between the factor of $|\mathcal{X}|$ in the bound above, which makes the bound deteriorate as $|\mathcal{X}|$ grows, versus the error in approximating $\|\boldsymbol{\Delta}\|$ by $\|\boldsymbol{\Delta}\|_{\mathcal{X}}$, which we will soon see decreases as $|\mathcal{X}|$ increases. We will therefore reduce the matter of bounding $\|\boldsymbol{\Delta}\|$ to merely that of finding a sufficiently small set $|\mathcal{X}|$ that behaves like a sufficiently dense "grid" of $\mathbb{S}^{d-1}$ or $\mathbb{B}^d$ to approximate the operator norm accurately (the sets we work with will not actually be grids but will behave similarly). We call this general approach the **Geometric Method** of random matrix theory, because it reduces the matter of understanding norms of random matrices to (possibly more complicated variants of) the above purely geometric question.

2.3.2 Nets, Coverings, and Packings

To do this, our first task will be to develop some tools for relating $\|\boldsymbol{M}\|_{\mathcal{X}}$ for finite $\mathcal{X}$ to $\|\boldsymbol{M}\|$, *discretizing* the continuous optimization appearing in the operator norm.

Definition 2.3.6 (ϵ-net)**.** *A set $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \mathcal{Y} \subset \mathbb{R}^d$ is an ϵ-*net *or ϵ-*covering *of $\mathcal{Y}$ if*

$$\mathcal{Y} \subseteq \bigcup_{i=1}^{N} \mathbb{B}^d(\boldsymbol{x}_i, \epsilon), \tag{2.3.10}$$

or equivalently if, for all $\boldsymbol{y} \in \mathcal{Y}$, there is $\boldsymbol{x}_i \in \mathcal{X}$ such that $\|\boldsymbol{y} - \boldsymbol{x}_i\| \leq \epsilon$.

The following key result shows that we may indeed discretize the operator norm and bound it by a maximum over a suitable net.

Lemma 2.3.7. *Suppose that $\boldsymbol{M} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ and $\mathcal{X}$ is an ϵ-net of $\mathbb{S}^{d-1}$ for $\epsilon < \frac{1}{2}$. Then,*

$$\|\boldsymbol{M}\|_{\mathcal{X}} \leq \|\boldsymbol{M}\| \leq \frac{1}{1-2\epsilon}\|\boldsymbol{M}\|_{\mathcal{X}}. \tag{2.3.11}$$

The same is true if $\mathcal{X}$ is instead an ϵ-net of $\mathbb{B}^d$.

Proof. The first inequality is immediate. For the second, let $\boldsymbol{x} \in \mathbb{S}^{d-1}$ be such that $|\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}| = \|\boldsymbol{M}\|$, and let $\boldsymbol{x}_i \in \mathcal{X}$ be such that $\|\boldsymbol{x} - \boldsymbol{x}_i\| \leq \epsilon$. We then have

$$\begin{aligned}
\|\boldsymbol{M}\|_{\mathcal{X}} &\geq \left|\boldsymbol{x}_i^\top \boldsymbol{M}\boldsymbol{x}_i\right| \\
&= \left|(\boldsymbol{x} + \boldsymbol{x}_i - \boldsymbol{x})^\top \boldsymbol{M}(\boldsymbol{x} + \boldsymbol{x}_i - \boldsymbol{x})\right| \\
&= \left|\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{x}) + (\boldsymbol{x}_i - \boldsymbol{x})^\top \boldsymbol{M}\boldsymbol{x}_i\right| \\
&\geq \left|\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}\right| - \left|\boldsymbol{x}^\top \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{x})\right| - \left|(\boldsymbol{x}_i - \boldsymbol{x})^\top \boldsymbol{M}\boldsymbol{x}_i\right| \\
&\geq (1 - 2\epsilon)\|\boldsymbol{M}\|,
\end{aligned}$$

and rearranging gives the result. The same argument applies verbatim in the case of a net of $\mathbb{B}^d$. □

Next, we need to develop some further tools for constructing ϵ-nets in a way such that we have some control over their size. As you can convince yourself, it is rather tricky to contrust explicit ϵ-nets by hand by completely describing the vectors of $\mathcal{X}$. Fortunately, there is a more abstract approach that lets us non-constructively show that small ϵ-nets exist. This proceeds by considering a complementary notion:

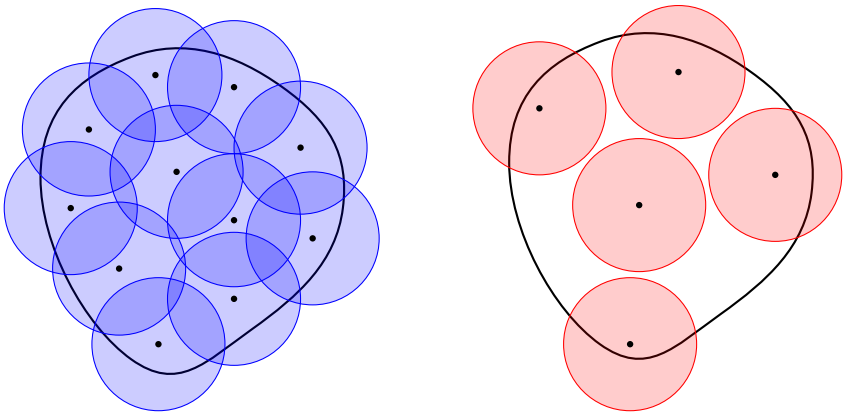

Figure 2.1: An illustration of a covering (left) and a packing (right) of a given set in $\mathbb{R}^2$ by balls of the same radius.

Definition 2.3.8 (ϵ-packing)**.** *A set $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \mathcal{Y} \subset \mathbb{R}^d$ is an ϵ-*packing of $\mathcal{Y}$ *if the balls $\mathbb{B}^d(\boldsymbol{x}_i, \epsilon)$ are pairwise disjoint, or equivalently if $\|\boldsymbol{x}_i - \boldsymbol{x}_j\| > 2\epsilon$ for all $i, j \in [N]$ distinct.*

Proposition 2.3.9. *Let $\mathcal{X}$ be a maximal ϵ-packing of $\mathcal{Y}$, one that cannot be extended to a larger packing by adding more points. Then, $\mathcal{X}$ is a 2ϵ-net of $\mathcal{Y}$.*

Proof. Let $\boldsymbol{y} \in \mathcal{Y} \setminus \mathcal{X}$. Since adding $\boldsymbol{y}$ to $\mathcal{X}$ must yield a set that is not an ϵ-packing, there must be some $\boldsymbol{x} \in \mathcal{X}$ such that $\mathbb{B}^d(\boldsymbol{x}, \epsilon) \cap \mathbb{B}^d(\boldsymbol{y}, \epsilon) \neq \emptyset$. Letting $\boldsymbol{z}$ be in the intersection, we have $\|\boldsymbol{x} - \boldsymbol{y}\| \leq \|\boldsymbol{x} - \boldsymbol{z}\| + \|\boldsymbol{y} - \boldsymbol{z}\| \leq 2\epsilon$. Thus, either $\boldsymbol{y} \in \mathcal{X}$ or $\boldsymbol{y} \in \mathbb{B}^d(\boldsymbol{x}, 2\epsilon)$ for some $\boldsymbol{x} \in \mathcal{X}$, giving the result. □

Thus, to show that small nets exist, it suffices to show that small maximal packings exist. But, there is a fundamental *volumetric* bound on the size of arbitrary packings, including maximal ones: roughly speaking, the total volume of the union of the balls in a packing cannot be much bigger than the volume of the set being packed (at least for a nice full-dimensional convex set).

Proposition 2.3.10. *Let $\mathcal{X}$ be any ϵ-packing of $\mathbb{B}^d$. Then, $|\mathcal{X}| \leq (1 + \frac{1}{\epsilon})^d$.*

Proof. We have

$$\bigcup_{\boldsymbol{x} \in \mathcal{X}} \mathbb{B}^d(\boldsymbol{x}, \epsilon) \subseteq \mathbb{B}^d(\boldsymbol{0}, 1 + \epsilon), \tag{2.3.12}$$

and the union on the left-hand side is disjoint. Thus, taking volumes,

$$|\mathcal{X}| \cdot \mathsf{vol}(\mathbb{B}^d(\boldsymbol{0}, \epsilon)) \leq \mathsf{vol}(\mathbb{B}^d(\boldsymbol{0}, 1 + \epsilon)). \tag{2.3.13}$$

Rearranging,

$$|\mathcal{X}| \leq \frac{\mathsf{vol}(\mathbb{B}^d(\boldsymbol{0}, 1 + \epsilon))}{\mathsf{vol}(\mathbb{B}^d(\boldsymbol{x}, \epsilon))} = \left(\frac{1 + \epsilon}{\epsilon}\right)^d = \left(1 + \frac{1}{\epsilon}\right)^d, \tag{2.3.14}$$

by scaling considerations of the volumes of d-dimensional balls. □

Corollary 2.3.11. *For any $\epsilon > 0$, there exists an ϵ-net $\mathcal{X}$ of $\mathbb{B}^d$ of size $|\mathcal{X}| \leq (1 + \frac{2}{\epsilon})^d$.*

Proof. Let $\mathcal{X}$ be a maximal $\frac{\epsilon}{2}$-packing of $\mathbb{B}^d$. By Proposition 2.3.10, $|\mathcal{X}| \leq (1 + \frac{2}{\epsilon})^d$, while by Proposition 2.3.9, $\mathcal{X}$ is an ϵ-net. □

2.3.3 Coarse Non-Asymptotic Bound on Singular Values

Finally, we may put together the pieces to bound $\|\boldsymbol{\Delta}\|$: we use the existence of a not-too-large ϵ-net by the above non-constructive method, approximate the operator norm by discretizing to test vectors in this net, and use the first moment method.

Theorem 2.3.12. *If $d \le m$, then*

$$\mathbb{P}[\|\boldsymbol{\Delta}\| > 64\sqrt{dm}] \le 2\exp\left(-\frac{d}{8}\right). \tag{2.3.15}$$

Proof. Let us fix $\epsilon := 1/4$. Let $\mathcal{X}$ be an ϵ-net of $\mathbb{B}^d$ as in Corollary 2.3.11, by which we can assume $|\mathcal{X}| \le (1+\frac{2}{\epsilon})^d = 9^d$. By Lemma 2.3.7, we have $\|\boldsymbol{\Delta}\| \le \frac{1}{1-2\epsilon}\|\boldsymbol{\Delta}\|_{\mathcal{X}} = 2\|\boldsymbol{\Delta}\|_{\mathcal{X}}$. Thus we may bound

$$\begin{aligned}
\mathbb{P}[\|\boldsymbol{\Delta}\| \ge t] &\le \mathbb{P}\left[\|\boldsymbol{\Delta}\|_{\mathcal{X}} \ge \frac{t}{2}\right] && \text{(Lemma 2.3.7)} \\
&\le 9^d \cdot \max_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}\left[|\boldsymbol{x}^\top \boldsymbol{\Delta}\boldsymbol{x}| \ge \frac{t}{2}\right] && \text{(Proposition 2.3.5)}
\end{aligned}$$

and writing $\hat{\boldsymbol{x}} := \boldsymbol{x}/\|\boldsymbol{x}\|$, since $\|\boldsymbol{x}\| \le 1$ for all $\boldsymbol{x} \in \mathcal{X}$, we may bound

$$\le 9^d \cdot \max_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}\left[|\hat{\boldsymbol{x}}^\top \boldsymbol{\Delta}\hat{\boldsymbol{x}}| \ge \frac{t}{2}\right]. \tag{2.3.16}$$

Let us work with this inner probability for any given $\boldsymbol{x}$ (and its corresponding normalization $\hat{\boldsymbol{x}}$) for a moment: expanding the definition of Δ,

$$\begin{aligned}
\mathbb{P}\left[|\hat{\boldsymbol{x}}^\top \boldsymbol{\Delta}\hat{\boldsymbol{x}}| \ge \frac{t}{2}\right] &= \mathbb{P}\left[\left|\|\boldsymbol{G}^\top\hat{\boldsymbol{x}}\|^2 - m\right| > \frac{t}{2}\right] \\
&= \mathbb{P}_{\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_m)}\left[\left|\|\boldsymbol{g}\|^2 - m\right| > \frac{t}{2}\right].
\end{aligned} \tag{2.3.17}$$

These probabilities are all equal, not depending on $\boldsymbol{x}$, so we have

$$\mathbb{P}[\|\boldsymbol{\Delta}\| \ge t] \le 9^d \cdot \mathbb{P}_{\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_m)}\left[\left|\|\boldsymbol{g}\|^2 - m\right| > \frac{t}{2}\right]$$

To the remaining probability we can apply Lemma 1.3.1, which it will be convenient to do specifically in the form (1.3.5). We can also bound $9 \le \exp(3)$. Together, these observations give

$$\le 2\exp\left(3d - \frac{1}{8}\min\left\{\frac{t^2}{4m}, \frac{t}{2}\right\}\right) \tag{2.3.18}$$

Now, let us take $t := C\sqrt{dm}$ for some C to be chosen momentarily. Noting that $\sqrt{dm} \ge d$ since $d \le m$, we may further bound

$$\begin{aligned}
&\le 2\exp\left(3d - \frac{1}{8}\min\left\{\frac{C^2}{4}d, \frac{C}{2}d\right\}\right) \\
&= 2\exp\left(d\cdot\left[3 - \min\left\{\frac{C^2}{32}, \frac{C}{16}\right\}\right]\right)
\end{aligned} \tag{2.3.19}$$

and finally we see that taking $C := 64$ gives the result. □

To give a corollary in a form that is maybe easier and more intuitive to understand concerning the singular values of $\boldsymbol{G}$, we will use the following basic eigenvalue perturbation inequalities.

Proposition 2.3.13. *Let $\boldsymbol{A}, \boldsymbol{\Delta} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$. Then,*

$$\lambda_d(\boldsymbol{A}) - \|\boldsymbol{\Delta}\| \le \lambda_d(\boldsymbol{A}+\boldsymbol{\Delta}) \le \lambda_1(\boldsymbol{A}+\boldsymbol{\Delta}) \le \lambda_1(\boldsymbol{A}) + \|\boldsymbol{\Delta}\|. \tag{2.3.20}$$

Proof. For both bounds, we use the variational description of the eigenvalues. For the upper bound, we have

$$\begin{aligned}
\lambda_1(\boldsymbol{A}+\boldsymbol{\Delta}) &= \sup_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \boldsymbol{x}^\top(\boldsymbol{A}+\boldsymbol{\Delta})\boldsymbol{x} \\
&\le \sup_{\boldsymbol{x}\in\mathbb{S}^{d-1}} (\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} + \|\boldsymbol{\Delta}\|) \\
&= \left(\sup_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}\right) + \|\boldsymbol{\Delta}\| \\
&= \lambda_1(\boldsymbol{A}) + \|\boldsymbol{\Delta}\|.
\end{aligned}$$

Similarly, for the lower bound,

$$\begin{aligned}
\lambda_d(\boldsymbol{A}+\boldsymbol{\Delta}) &= \inf_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \boldsymbol{x}^\top(\boldsymbol{A}+\boldsymbol{\Delta})\boldsymbol{x} \\
&\ge \inf_{\boldsymbol{x}\in\mathbb{S}^{d-1}} (\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} - \|\boldsymbol{\Delta}\|) \\
&= \left(\inf_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}\right) - \|\boldsymbol{\Delta}\| \\
&= \lambda_d(\boldsymbol{A}) - \|\boldsymbol{\Delta}\|.
\end{aligned}$$

□

These are special cases of *Weyl's inequalities*, which are a good more general tool to be aware of. The proof of the more general inequalities similarly uses a more general variational form of arbitrary eigenvalues given by the *Courant-Fischer min-max theorem*.

Corollary 2.3.14. *For any $d \le m$,*

$$\mathbb{P}\left[\sqrt{m} - 64\sqrt{d} \le \sigma_d(\boldsymbol{G}) \le \sigma_1(\boldsymbol{G}) \le \sqrt{m} + 32\sqrt{d}\right] \ge 1 - 2\exp(-d).$$

Proof. We will use the basic bounds

$$\sqrt{1+x} \le 1 + \frac{x}{2} \quad \text{for all } x \ge 0, \tag{2.3.21}$$

$$\sqrt{1-x} \ge 1 - x \quad \text{for all } 0 \le x \le 1. \tag{2.3.22}$$

Both may be viewed as consequences of the concavity of the square root function: the first says that one affine transformation of the square root lies below a tangent line to it, while the second says that another affine transformation lies above a secant segment connecting

two points on its graph. Then, on the event that $\|\boldsymbol{\Delta}\| > 64\sqrt{dm}$ from Theorem 2.3.12, using that $\boldsymbol{G}\boldsymbol{G}^\top = m\boldsymbol{I}_d + \boldsymbol{\Delta}$, we have

$$\begin{aligned}
\sigma_1(\boldsymbol{G}) &= \sqrt{\lambda_1(\boldsymbol{G}\boldsymbol{G}^\top)} \\
&\le \sqrt{m + \|\boldsymbol{\Delta}\|} && \text{(Proposition 2.3.13)} \\
&\le \sqrt{m + 64\sqrt{dm}} \\
&= \sqrt{m} \cdot \sqrt{1 + 64\sqrt{\frac{d}{m}}} \\
&\le \sqrt{m} \cdot \left(1 + 32\sqrt{\frac{d}{m}}\right) && \text{(2.3.21)} \\
&= \sqrt{m} + 32\sqrt{d}.
\end{aligned}$$

Likewise, for $\sigma_d(\boldsymbol{G})$ we have a similar argument. First, though, note that if $m - 64\sqrt{dm} < 0$, then $\sqrt{m} - 64\sqrt{d} < 0$ as well, so the result holds vacuously in this case. Otherwise,

$$\begin{aligned}
\sigma_d(\boldsymbol{G}) &= \sqrt{\lambda_d(\boldsymbol{G}\boldsymbol{G}^\top)} \\
&\ge \sqrt{m - \|\boldsymbol{\Delta}\|} && \text{(Proposition 2.3.13)} \\
&\ge \sqrt{m - 64\sqrt{dm}} \\
&= \sqrt{m} \cdot \sqrt{1 - 64\sqrt{\frac{d}{m}}} \\
&\ge \sqrt{m} \cdot \left(1 - 64\sqrt{\frac{d}{m}}\right) && \text{(2.3.22)} \\
&= \sqrt{m} - 64\sqrt{d},
\end{aligned}$$

completing the proof. □

Of course, what you should remember is just the softer statement that the singular values of $\boldsymbol{G}$ are, with high probability, all in the range

$$\boxed{\sqrt{m} - O(\sqrt{d}) \le \sigma_1(\boldsymbol{G}) \le \cdots \le \sigma_d(\boldsymbol{G}) \le \sqrt{m} + O(\sqrt{d}) \quad (\text{when } d \le m)} \tag{2.3.23}$$

We highlight this because it is a fundamental fact about the spectra of random matrices that you should try to commit to your intuition. We will see considerably more precise elaborations of this claim soon, too. Note that, indeed, once $m \gg d$, then all of these singular values are $\sqrt{m}$ to leading order and $\boldsymbol{G}\boldsymbol{G}^\top \approx m\boldsymbol{I}_d$, as we proposed earlier.

The top d eigenvalues of the larger matrix $\boldsymbol{G}^\top\boldsymbol{G} \in \mathbb{R}^{m\times m}_{\text{sym}}$ are also then approximately m; however, for this matrix, there are $m - d$ more eigenvalues that are identically zero. Thus, the eigenvalues of $\frac{1}{m}\boldsymbol{G}^\top\boldsymbol{G}$ are all either zero or close to 1; that is, it is close to a *projection matrix*. It is then natural to ask: a projection to what subspace? To answer this question we must understand the singular vectors of $\boldsymbol{G}$, to which we now turn our attention.

2.4 Rectangular Gaussian Matrices: Singular Vectors

While the ordered singular values $\sigma_1(\boldsymbol{G}) \geq \cdots \geq \sigma_d(\boldsymbol{G})$ are well-defined for any $\boldsymbol{G}$, the individual singular vectors are not always well-defined. Suppose we take the singular value decomposition (SVD),

$$\boldsymbol{G} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \sum_{i=1}^{d} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top,$$

for $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d\} \subset \mathbb{R}^d$ and $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d\} \subset \mathbb{R}^m$ orthonormal sets. When are we justified in viewing $\boldsymbol{u}_i(\boldsymbol{G})$ and $\boldsymbol{v}_i(\boldsymbol{G})$ as well-defined functions?

Actually, the answer is *never*, because the SVD does not disambiguate between the pair of singular vectors $(\boldsymbol{u}_i, \boldsymbol{v}_i)$ and the pair $(-\boldsymbol{u}_i, -\boldsymbol{v}_i)$; either pair would give rise to the same decomposition. To get around this minor issue, we may instead consider the rank-one matrices associated to the pairs of singular vectors, $\boldsymbol{Z}_i = \boldsymbol{u}_i \boldsymbol{v}_i^\top$. These matrices are unchanged by a simultaneous change of sign, and from them we may also recover the more geometrically meaningful projection matrices

$$\boldsymbol{L}_i := \boldsymbol{Z}_i \boldsymbol{Z}_i^\top = \boldsymbol{u}_i \boldsymbol{u}_i^\top, \tag{2.4.1}$$

$$\boldsymbol{R}_i := \boldsymbol{Z}_i^\top \boldsymbol{Z}_i = \boldsymbol{v}_i \boldsymbol{v}_i^\top. \tag{2.4.2}$$

It is sometimes useful for the sake of intuition to identify a projection matrix with the subspace it projects to; thus, $\boldsymbol{u}_i \boldsymbol{u}_i^\top$ and $\boldsymbol{v}_i \boldsymbol{v}_i^\top$ describe not the singular *vectors* but the entire *lines* in those directions. $\boldsymbol{Z}_i$ contains the information of this pair of lines, and its being one of the rank-one matrices in the SVD says that $\boldsymbol{G}$ maps the line spanned by $\boldsymbol{v}_i$ to that spanned by $\boldsymbol{u}_i$. $\boldsymbol{Z}_i$ also contains one more bit of information, saying whether that mapping maps $\boldsymbol{v}_i$ to a positive or negative multiple of $\boldsymbol{u}_i$.

However, there are more nuances. For instance, if $\sigma_i = 0$, then $\boldsymbol{Z}_i$ is equivalent to $-\boldsymbol{Z}_i$ in the SVD. And, more substantially, if $\sigma_i = \sigma_j$, even if neither are zero, then $\boldsymbol{Z}_i$ and $\boldsymbol{Z}_j$ are not uniquely determined. Indeed, writing

$$\widetilde{\boldsymbol{U}} := \begin{bmatrix} | & | \\ \boldsymbol{u}_i & \boldsymbol{u}_j \\ | & | \end{bmatrix} \in \mathbb{R}^{2\times d}, \quad \widetilde{\boldsymbol{V}} := \begin{bmatrix} | & | \\ \boldsymbol{v}_i & \boldsymbol{v}_j \\ | & | \end{bmatrix} \in \mathbb{R}^{2\times m},$$

we have that the sum of the two corresponding terms in the SVD is

$$\sigma_i(\boldsymbol{R}_i + \boldsymbol{R}_j) = \sigma_i(\boldsymbol{u}_i \boldsymbol{v}_i + \boldsymbol{u}_j \boldsymbol{v}_j^\top) = \sigma_i \widetilde{\boldsymbol{U}} \widetilde{\boldsymbol{V}}^\top = \sigma_i (\widetilde{\boldsymbol{U}}\boldsymbol{Q})(\widetilde{\boldsymbol{V}}\boldsymbol{Q})^\top$$

for any $\boldsymbol{Q} \in \mathcal{O}(2)$. This gives rise to equivalent decompositions of $\boldsymbol{R}_i + \boldsymbol{R}_j$ as a sum of rank-two matrices. Since $\sigma_i = \sigma_j$, the expression $\sigma_i(\boldsymbol{R}_i + \boldsymbol{R}_j)$ is what appears in the SVD for these two components, so these equivalent decompositions mean that $\boldsymbol{R}_i$ and $\boldsymbol{R}_j$ are not uniquely determined in this case.

As an extreme example, consider the case $d = m$ and $\boldsymbol{G} = \boldsymbol{I}_d$, the identity matrix. All d of its singular values are 1, and it admits many SVDs of the form $\boldsymbol{G} = \sum_{i=1}^{d} \boldsymbol{u}_i \boldsymbol{u}_i^\top$ for *any* orthonormal basis $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d\}$. Thus, the more repeated singular values there are, the more dire this issue becomes. However, as the following fact of linear algebra shows, these are the only issues.

Proposition 2.4.1. *Suppose that $\boldsymbol{G} \in \mathbb{R}^{d\times m}$ with $d \le m$, $\sigma_d(\boldsymbol{G}) > 0$, and the singular values $\sigma_1(\boldsymbol{G}), \dots, \sigma_d(\boldsymbol{G})$ are distinct. Then, there exist unique $\boldsymbol{Z}_i = \boldsymbol{Z}_i(\boldsymbol{G}) \in \mathbb{R}^{d\times m}$ such that each $\boldsymbol{R}_i$ has rank one, has $\sigma_1(\boldsymbol{Z}_i) = 1$, and*

$$\boldsymbol{G} = \sum_{i=1}^{d} \sigma_i(\boldsymbol{G})\boldsymbol{Z}_i(\boldsymbol{G}).$$

2.4.1 Distinctness of Singular Values

Remarkably, the conditions of the above Proposition 2.4.1 hold not merely with high probability but *almost surely* for Gaussian random matrices (and, though we will not discuss it here, for other matrices whose entries are independent random variables whose laws are absolutely continuous with respect to the Lebesgue measure).

Theorem 2.4.2. *Let $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d\times m}$ with $d \le m$. Then, almost surely, the following hold:*

1. *$\sigma_d(\boldsymbol{G}) > 0$.*
2. *The singular values $\sigma_1(\boldsymbol{G}), \dots, \sigma_d(\boldsymbol{G})$ are distinct, i.e., we have the strict inequalities $\sigma_1(\boldsymbol{G}) > \cdots > \sigma_d(\boldsymbol{G})$.*

We will use the following result to establish both parts, which should be intuitive but is a little bit clunky to prove. You can prove it as an exercise—one approach is to show the more geometric fact that any hypersurface has Lebesgue measure zero.

Lemma 2.4.3. *Suppose $p \in \mathbb{R}[x_1, \dots, x_N]$ is a non-constant polynomial, and $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_N)$. Then, for any $a \in \mathbb{R}$, $\mathbb{P}[p(\boldsymbol{g}) = a] = 0$.*

The first part of Theorem 2.4.2 can be proved quite directly using this result, as follows.

Proof of Theorem 2.4.2, Claim 1. Note that we have $\sigma_d(\boldsymbol{G}) = 0$ if and only if $\boldsymbol{G}$ is rank-deficient, which occurs if and only if $P(\boldsymbol{G}) := \det(\boldsymbol{G}\boldsymbol{G}^\top) = 0$. (Note that this is not the case if we take $\det(\boldsymbol{G}^\top\boldsymbol{G})$ instead, which is zero whenever $d < m$.) It is easy to show that p is a non-constant polynomial of the dm entries of the matrix argument: it is non-constant because, for example, for rank-deficient matrix inputs it is zero while for full-rank inputs it is non-zero. Thus $\mathbb{P}[\sigma_d(\boldsymbol{G}) = 0] = \mathbb{P}[P(\boldsymbol{G}) = 0] = 0$ by Lemma 2.4.3. □

For the second part we will need to establish some tools from the theory of symmetric polynomials as well as a relationship between polynomials of the *eigenvalues* of a matrix and polynomials of the *entries* of a matrix. First, the following are the basic notions of the general theory of symmetric polynomials.

Definition 2.4.4. *A polynomial $f \in \mathbb{R}[t_1, \dots, t_d]$ is* symmetric *if, for all permutations $\pi \in S_d$, we have $f(t_{\pi(1)}, \dots, t_{\pi(d)}) = f(t_1, \dots, t_d)$.*

Theorem 2.4.5 (Fundamental theorem of symmetric polynomials)**.** *Define the* power sum *polynomials*

$$s_k(t_1, \dots, t_d) := \sum_{i=1}^{d} t_i^k.$$

Then, if $f \in \mathbb{R}[t_1, \ldots, t_d]$ is a symmetric polynomial, then there exists some polynomial g of d variables such that

$$f(t_1, \ldots, t_d) = g(s_1(t_1, \ldots, t_d), \ldots, s_d(t_1, \ldots, t_d)).$$

In words, this result says that any symmetric polynomial is some polynomial of the power sum polynomials, or that the power sum polynomials *algebraically generate* the symmetric polynomials (which form a subring of the ring of all polynomials).

The power sum polynomials have an important role in random matrix theory, because they serve as a bridge between the properties of the entries of a matrix and the properties of the eigenvalues, a connection that we will see much more sophisticated uses of later. The basic underlying fact is quite simple, however. Suppose here and below that $\boldsymbol{X} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$. Let us define the *power trace* polynomials of $\boldsymbol{X}$,

$$S_k(\boldsymbol{X}) := \mathsf{Tr}(\boldsymbol{X}^k) = \sum_{a_1,\ldots,a_k=1}^{d} X_{a_1a_2}X_{a_2a_3}\cdots X_{a_{k-1}a_k}X_{a_ka_1}.$$

The latter expression shows that $S_k(\boldsymbol{X})$ is a polynomial of the entries of $\boldsymbol{X}$.

Proposition 2.4.6. $s_k(\lambda_1(\boldsymbol{X}), \ldots, \lambda_d(\boldsymbol{X})) = S_k(\boldsymbol{X})$.

This is just a simple consequence of the fact that the eigenvalues of $\boldsymbol{X}^k$ are the $\lambda_i(\boldsymbol{X})^k$. However, it leads to the following important corollary.

Corollary 2.4.7. *Suppose that $F(\boldsymbol{X}) = f(\lambda_1(\boldsymbol{X}), \ldots, \lambda_d(\boldsymbol{X}))$ for f a symmetric polynomial (i.e., $F(\boldsymbol{X})$ is a symmetric polynomial of the eigenvalues of $\boldsymbol{X}$). Then, $F(\boldsymbol{X})$ is also a (not necessarily symmetric) polynomial of the entries of $\boldsymbol{X}$.*

Remark 2.4.8. *The assumption that f is symmetric is crucial. For instance, you may check as an exercise that $F(\boldsymbol{X}) = \lambda_1(\boldsymbol{X})$ is not a polynomial of $\boldsymbol{X}$.*

With these tools, we are now ready for the proof of the second part of the main Theorem.

Proof of Theorem 2.4.2, Claim 2. Consider first a function of $\boldsymbol{X} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$, defined as:

$$\begin{aligned} f(t_1, \ldots, t_d) &:= \prod_{1\le i<j\le d} (t_i - t_j)^2, \\ F(\boldsymbol{X}) &:= f(\lambda_1(\boldsymbol{X}), \ldots, \lambda_d(\boldsymbol{X})) \\ &= \prod_{1\le i<j\le d} (\lambda_i(\boldsymbol{X}) - \lambda_j(\boldsymbol{X}))^2. \end{aligned}$$

Clearly we have that $\boldsymbol{X}$ has a repeated eigenvalue if and only if $F(\boldsymbol{X}) = 0$. Further, $F(\boldsymbol{X})$ satisfies the conditions of Corollary 2.4.7. Therefore, $F(\boldsymbol{X})$ is a polynomial of the entries of $\boldsymbol{X}$. (Note that, remarkably, we can draw this conclusion non-constructively from the tools above.)

Now, returning to the setting of the Theorem, we see that $\boldsymbol{G}$ has a repeated singular value if and only if $P(\boldsymbol{G}) = 0$, where $P(\boldsymbol{G}) := F(\boldsymbol{G}\boldsymbol{G}^\top)$. As before, P is non-constant merely

because some matrices do have repeated singular values and some do not. Thus, we may conclude as before by Lemma 2.4.3 that

$$\mathbb{P}[\boldsymbol{G} \text{ has a repeated singular value}] = \mathbb{P}[P(\boldsymbol{G}) = 0] = 0,$$

as claimed. □

Remark 2.4.9. *The proof above shows that the set of symmetric matrices with a repeated eigenvalue is a hypersurface in the space $\mathbb{R}^{d\times d}_{\mathsf{sym}}$, which may be identified with $\mathbb{R}^{d(d+1)/2}$.* A priori, *since we described this set as $\{\boldsymbol{X} : F(\boldsymbol{X}) = 0\}$ for a polynomial F, we only know that its codimension is at most 1 (i.e., the dimension of this hypersurface is at most $\frac{d(d+1)}{2} - 1$). Actually, its codimension is exactly 2. For example, the dimension of $\mathbb{R}^{2\times 2}_{\mathsf{sym}}$ is $\frac{2(2+1)}{2} = 3$, while the 2×2 matrices with a repeated eigenvalue are merely multiples of the identity, whose dimension is clearly $1 = 3 - 2$ (indeed, these matrices form a line in the space of all symmetric matrices). Note that one polynomial constraint can in fact encode several, for instance if $F(\boldsymbol{X}) = F_1(\boldsymbol{X})^2 + F_2(\boldsymbol{X})^2 = 0$, then $F_1(\boldsymbol{X}) = F_2(\boldsymbol{X}) = 0$. As discussed for instance by [BKL18], a more complicated version of this sum-of-squares structure is indeed what happens in the case of $F(\boldsymbol{X})$.*

We have established that, on an almost sure event, we are justified in speaking of *the* singular vectors of a Gaussian random matrix, at least via the rank-one matrices $\boldsymbol{Z}_i(\boldsymbol{G})$ and the projections to the left and right singular vectors $\boldsymbol{L}_i(\boldsymbol{G})$ and $\boldsymbol{R}_i(\boldsymbol{G})$. We are now ready to study the laws of these objects.

2.4.2 Orthogonal Invariance for Matrices

We will see that orthogonal invariance, as we discussed for vectors earlier, is crucial to understanding these distributions. The appropriate definitions for random matrices are as follows:

Definition 2.4.10 (Orthogonal invariance of matrices)**.** *We say that a random matrix $\boldsymbol{X} \in \mathbb{R}^{d\times m}$ (or its law $\mathsf{Law}(\boldsymbol{x})$) is:*

- Left-orthogonally invariant *if, for each (deterministic) orthogonal matrix $\boldsymbol{Q} \in \mathcal{O}(d)$, we have $\mathsf{Law}(\boldsymbol{QX}) = \mathsf{Law}(\boldsymbol{X})$.*
- Right-orthogonally invariant *if, for each (deterministic) orthogonal matrix $\boldsymbol{Q} \in \mathcal{O}(m)$, we have $\mathsf{Law}(\boldsymbol{XQ}) = \mathsf{Law}(\boldsymbol{X})$.*
- Bi-orthogonally invariant *if it is both left- and right-orthogonally invariant.*

Further, we say that a random symmetric matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ is symmetric-orthogonally invariant *if $\mathsf{Law}(\boldsymbol{QXQ}^\top) = \mathsf{Law}(\boldsymbol{X})$ for each $\boldsymbol{Q} \in \mathcal{O}(d)$.*

The following property is then simple to check using either explicit mean and covariance calculations or by using that the rows and columns of an i.i.d. Gaussian random matrix are i.i.d. standard Gaussian vectors, which are themselves orthogonally invariant.

Proposition 2.4.11. *$\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d\times m}$ is bi-orthogonally invariant.*

Corollary 2.4.12. *If $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d\times m}$, then $\boldsymbol{L}_i(\boldsymbol{G})$ and $\boldsymbol{R}_i(\boldsymbol{G})$ are both symmetric-orthogonally invariant, for each $i = 1,\dots,d$. Further, the stronger equalities of law hold that*

$$\mathsf{Law}\Big((\boldsymbol{Q}\boldsymbol{L}_1(\boldsymbol{G})\boldsymbol{Q}^\top,\dots,\boldsymbol{Q}\boldsymbol{L}_d(\boldsymbol{G})\boldsymbol{Q}^\top)\Big) = \mathsf{Law}\Big((\boldsymbol{L}_1(\boldsymbol{G}),\dots,\boldsymbol{L}_d(\boldsymbol{G}))\Big) \quad \text{for all } \boldsymbol{Q}\in\mathcal{O}(d),$$
$$\mathsf{Law}\Big((\boldsymbol{Q}\boldsymbol{R}_1(\boldsymbol{G})\boldsymbol{Q}^\top,\dots,\boldsymbol{Q}\boldsymbol{R}_d(\boldsymbol{G})\boldsymbol{Q}^\top)\Big) = \mathsf{Law}\Big((\boldsymbol{R}_1(\boldsymbol{G}),\dots,\boldsymbol{R}_d(\boldsymbol{G}))\Big) \quad \text{for all } \boldsymbol{Q}\in\mathcal{O}(m).$$

The following basic linear algebra observation is a useful preliminary to the proof.

Proposition 2.4.13 (Invariance and equivariance of SVD)**.** *Let $\boldsymbol{G}\in\mathbb{R}^{d\times m}$ be a (deterministic) matrix such that $\sigma_1(\boldsymbol{G}) > \cdots > \sigma_d(\boldsymbol{G}) > 0$. Let $\boldsymbol{Q}_L\in\mathcal{O}(d)$ and $\boldsymbol{Q}_R\in\mathcal{O}(m)$. Then,*

$$\begin{aligned}
\sigma_i(\boldsymbol{Q}_L\boldsymbol{G}\boldsymbol{Q}_R^\top) &= \sigma_i(\boldsymbol{G}),\\
\boldsymbol{Z}_i(\boldsymbol{Q}_L\boldsymbol{G}\boldsymbol{Q}_R^\top) &= \boldsymbol{Q}_L\boldsymbol{Z}_i(\boldsymbol{G})\boldsymbol{Q}_R^\top,\\
\boldsymbol{L}_i(\boldsymbol{Q}_L\boldsymbol{G}\boldsymbol{Q}_R^\top) &= \boldsymbol{Q}_L\boldsymbol{L}_i(\boldsymbol{G})\boldsymbol{Q}_L^\top,\\
\boldsymbol{R}_i(\boldsymbol{Q}_L\boldsymbol{G}\boldsymbol{Q}_R^\top) &= \boldsymbol{Q}_R\boldsymbol{R}_i(\boldsymbol{G})\boldsymbol{Q}_R^\top.
\end{aligned}$$

Proof of Corollary 2.4.12. Let us prove the simpler claim about the individual $\boldsymbol{L}_i$: we have, by combining the above results,

$$\begin{aligned}
\mathsf{Law}(\boldsymbol{Q}\boldsymbol{L}_i(\boldsymbol{G})\boldsymbol{Q}^\top) &= \mathsf{Law}(\boldsymbol{L}_i(\boldsymbol{Q}\boldsymbol{G})) && \text{(Proposition 2.4.13)}\\
&= \mathsf{Law}(\boldsymbol{L}_i(\boldsymbol{G})). && \text{(Proposition 2.4.11)}
\end{aligned}$$

The result for the individual $\boldsymbol{R}_i$ and the collections of all $\boldsymbol{L}_i$ and $\boldsymbol{R}_i$ follows similarly. □

Let us comment on the intuition behind what we have derived so far. Consider the $\boldsymbol{L}_i(\boldsymbol{G})$. Recall that, for $\boldsymbol{u}_i$ some choice of the left singular vectors in the SVD (which for $\boldsymbol{G}$ are almost surely well-defined up to a sign flip), $\boldsymbol{L}_i(\boldsymbol{G}) = \boldsymbol{u}_i\boldsymbol{u}_i^\top$. As we have discussed, we may identify this rank-one projection matrix with the line formed by $\mathsf{span}(\boldsymbol{u}_i)$. Thus the collection $(\boldsymbol{L}_1(\boldsymbol{G}),\dots,\boldsymbol{L}_d(\boldsymbol{G}))$ is a collection of d orthogonal lines, i.e., a coordinate system for $\mathbb{R}^d$ consisting of d orthogonal axes, although without a "positive direction" chosen for any of the axes. The above result says that this object for $\boldsymbol{G}\sim\mathcal{N}(0,1)^{\otimes d\times m}$ has an orthogonally invariant distribution. We will next try to formally explain in what sense this makes $(\boldsymbol{L}_1(\boldsymbol{G}),\dots,\boldsymbol{L}_d(\boldsymbol{G}))$ have the law of the unique "uniformly random" choice of coordinate system.

2.5 Application: Compressed Sensing

We will now see one last application of the ideas developed so far. This is for the problem of *compressed sensing*, recovering $\boldsymbol{x}\in\mathbb{R}^m$ from $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{x}\in\mathbb{R}^d$ with $\boldsymbol{G}\in\mathbb{R}^{d\times m}$ (the same dimensions as before).

For general $\boldsymbol{x}\in\mathbb{R}^m$, we need $\boldsymbol{G}$ to be injective, which requires $d\geq m$. Compressed sensing concerns making d much smaller, provided we are promised that $\boldsymbol{x}$ is *sparse*. In this setting, while we will discuss taking $\boldsymbol{G}$ random, we view ourselves as having control over $\boldsymbol{G}$ (the so-called *sensing matrix*) in general, so the sparsity assumption should be seen as in a basis of our choosing. Thus to take advantage of compressed sensing we should encode any prior knowledge of $\boldsymbol{x}$ in a basis that makes $\boldsymbol{x}$ sparse. (For example, if $\boldsymbol{x}$ encodes an image, we might choose a wavelet basis.)

2.5.1 Null Space and Restricted Isometry Properties

Let us denote sparsity by

$$\|\boldsymbol{x}\|_0 := \#\{i \in [m] : x_i \neq 0\}. \tag{2.5.1}$$

Note that this "ℓ^0-norm" is not actually a norm, because it is not homogeneous in $\boldsymbol{x}$.

Assuming that $\|\boldsymbol{x}\|_0 \leq k$ makes the task of recovering $\boldsymbol{x}$ easier. At the extreme, if $k = 1$, then $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{x}$ is just (up to scaling) one of the columns of $\boldsymbol{G}$. Thus provided the columns of $\boldsymbol{G}$ are well-separated, it will be easy to recover $\boldsymbol{x}$. You can show that it is possible to construct $\exp(\Omega(d))$ unit vectors in $\mathbb{R}^d$ that have, say, pairwise inner products of magnitude each at most $1/2$, which we may use as the columns of $\boldsymbol{G}$ (and such a choice will be robust to a small amount of noise). Thus when $k = 1$ then we may take d as small as $O(\log m)$ while still being able to recover any 1-sparse $\boldsymbol{x}$.

How does the situation change for larger k? We will show below that it actually does not change that much.

First, let us specify what we mean by $\boldsymbol{x}$ being "recoverable." The following definition is not standard but useful.

Definition 2.5.1. *We say that* $\boldsymbol{G}$ distinguishes k-sparse vectors *if* $\boldsymbol{G}\boldsymbol{x} \neq \boldsymbol{G}\boldsymbol{x}'$ *for any* $\boldsymbol{x} \neq \boldsymbol{x}'$ *with* $\|\boldsymbol{x}\|_0, \|\boldsymbol{x}'\|_0 \leq k$.

If $\boldsymbol{G}$ distinguishes k-sparse vectors, then compressed sensing is *information-theoretically* possible: by, say, brute force search over an ϵ-net of sparse vectors followed by a rounding procedure, we may exactly recover any k-sparse $\boldsymbol{x}$ from $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{x}$. We will not go into the *computational* feasibility of compressed sensing here, which is a deep area in its own right. You may look up the role of ℓ^1-norm minimization for such algorithms to get started with computational approaches.

The following more linear-algebraic definition is actually equivalent to distinguishing k-sparse vectors.

Definition 2.5.2 (Null space property)**.** *We say that* $\boldsymbol{G}$ *has the* k-null space property (k-NSP) *if, for all* $\boldsymbol{x} \neq \boldsymbol{0}$ *with* $\|\boldsymbol{x}\|_0 \leq k$, $\boldsymbol{G}\boldsymbol{x} \neq \boldsymbol{0}$.

Proposition 2.5.3. $\boldsymbol{G}$ *distinguishes* k*-sparse vectors if and only if* $\boldsymbol{G}$ *has the* $2k$*-NSP.*

Proof. If $\boldsymbol{G}$ does not distinguish k-sparse vectors, then there exist $\boldsymbol{x} \neq \boldsymbol{x}'$ with $\|\boldsymbol{x}\|_0, \|\boldsymbol{x}'\|_0 \leq k$ such that $\boldsymbol{G}\boldsymbol{x} = \boldsymbol{G}\boldsymbol{x}'$. In particular, $\boldsymbol{G}(\boldsymbol{x} - \boldsymbol{x}') = \boldsymbol{0}$, and $\|\boldsymbol{x} - \boldsymbol{x}'\|_0 \leq 2k$, so $\boldsymbol{G}$ does not have the $2k$-NSP. Conversely, if $\boldsymbol{G}$ does distinguish k-sparse vectors and $\boldsymbol{x}'' \neq \boldsymbol{0}$ has $\|\boldsymbol{x}''\| \leq 2k$, then we may write $\boldsymbol{x}'' = \boldsymbol{x} - \boldsymbol{x}'$ for $\boldsymbol{x} \neq \boldsymbol{x}'$ and $\|\boldsymbol{x}\|_0, \|\boldsymbol{x}'\|_0 \leq k$ (by partitioning the $2k$ indices on which $\boldsymbol{x}''$ is non-zero in some arbitrary way). Reversing the above argument then shows that $\boldsymbol{G}$ has the $2k$-NSP. □

We will in fact be able to show the following, more quantitative property.

Definition 2.5.4 (Restricted isometry property)**.** *We say that* $\boldsymbol{G}$ *has the* (k, δ)-restricted isometry property ((k, δ)-RIP) *if, for all* $\boldsymbol{x} \neq \boldsymbol{0}$ *with* $\|\boldsymbol{x}\|_0 \leq k$,

$$(1 - \delta)\|\boldsymbol{x}\|^2 \leq \|\boldsymbol{G}\boldsymbol{x}\|^2 \leq (1 + \delta)\|\boldsymbol{x}\|^2. \tag{2.5.2}$$

The following implication is immediate.

Proposition 2.5.5. *If $\boldsymbol{G}$ has the (k,δ)-RIP for any $\delta < 1$, then $\boldsymbol{G}$ has the k-NSP.*

The RIP is more useful than the NSP both for analyzing compressed sensing when some noise is further added to $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{x}$, and for analyzing algorithms for recovering $\boldsymbol{x}$, but here we will just view it as an analytic tool for establishing the NSP.

2.5.2 Random Sensing Matrices

The following is the main result that we will show.

Theorem 2.5.6. *For any $1 \le k \le m$ and $\delta \in (0,1)$, if*

$$d \ge \frac{160}{\delta^2} k \left(1 + \log\left(\frac{m}{k}\right)\right)$$

and $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d \times m}$, then

$$\mathbb{P}[\boldsymbol{G} \text{ has the } k\text{-NSP}] \ge \mathbb{P}[\boldsymbol{G} \text{ has the } (k,\delta)\text{-RIP}] \ge 1 - \frac{2}{\binom{m}{k}} \ge 1 - 2\left(\frac{k}{m}\right)^k. \tag{2.5.3}$$

Note that this scaling of d is consistent with our earlier observation that $d \gtrsim \log m$ was the right condition for $k = 1$. Remarkably, the result implies that, while

$$d \ge m$$

is needed for *arbitrary* $\boldsymbol{x}$ to be recoverable from $\boldsymbol{G}\boldsymbol{x}$, for the specific case of k-*sparse* vectors it suffices to have only the far smaller

$$\boxed{d \gtrsim k \log\left(\tfrac{m}{k}\right)} \tag{2.5.4}$$

Proof. For $S \subseteq [m]$ with $|S| = k$ and $\boldsymbol{x} \in \mathbb{R}^m$, let $\boldsymbol{x}^{(S)} \in \mathbb{R}^k$ be the restriction of $\boldsymbol{x}$ to the indices in S. Likewise, for $\boldsymbol{G} \in \mathbb{R}^{d\times m}$, let $\boldsymbol{G}^{(S)} \in \mathbb{R}^{d\times k}$ be the restriction of $\boldsymbol{G}$ to the columns whose indices are in S. If $\|\boldsymbol{x}\|_0 \le k$ and the non-zero indices of $\boldsymbol{x}$ are contained in such S, then

$$\|\boldsymbol{x}\|^2 = \|\boldsymbol{x}^{(S)}\|^2, \tag{2.5.5}$$

$$\boldsymbol{G}\boldsymbol{x} = \boldsymbol{G}^{(S)}\boldsymbol{x}^{(S)}. \tag{2.5.6}$$

We may then view the (k,δ)-RIP as requiring that, for all $S \subseteq [m]$ with $|S| = k$ and all $\boldsymbol{x}^{(S)} \in \mathbb{R}^k$, we have

$$(1-\delta)\|\boldsymbol{x}^{(S)}\|^2 \le \|\boldsymbol{G}^{(S)}\boldsymbol{x}^{(S)}\|^2 \le (1+\delta)\|\boldsymbol{x}^{(S)}\|^2. \tag{2.5.7}$$

But this just amounts to asking that

$$1 - \delta \le \lambda_k(\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)}) \le \lambda_1(\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)}) \le 1 + \delta, \tag{2.5.8}$$

or again equivalently that

$$\|\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)} - \boldsymbol{I}_k\| \le \delta. \tag{2.5.9}$$

We will then proceed by union bounding,

$$\mathbb{P}[\boldsymbol{G} \text{ does not have the } (k,\delta)\text{-RIP}] \le \sum_{S\in\binom{[m]}{k}} \mathbb{P}[\|\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)} - \boldsymbol{I}_k\| > \delta]. \tag{2.5.10}$$

This is almost exactly the kind of deviation that we bounded before in Theorem 2.3.12. First, let's decode our statement into that form. Introduce $\boldsymbol{H} \sim \mathcal{N}(0,1)^{\otimes k\times d}$, the scaling and aspect ratio that Theorem 2.3.12 addresses. We have $\mathsf{Law}(\boldsymbol{G}^{(S)}) = \mathsf{Law}(d^{-1/2}\boldsymbol{H}^\top)$. Thus,

$$\begin{aligned}\mathbb{P}[\|\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)} - \boldsymbol{I}_k\| > \delta] &= \mathbb{P}\left[\left\|\frac{1}{d}\boldsymbol{H}\boldsymbol{H}^\top - \boldsymbol{I}_k\right\| > \delta\right] \\ &= \mathbb{P}\left[\|\boldsymbol{H}\boldsymbol{H}^\top - d\boldsymbol{I}_k\| > \delta d\right].\end{aligned}$$

The only difference between this and what Theorem 2.3.12 covered is that there we were concerned with deviations of the order $O(\sqrt{kd})$, while here we are concerned with $O(d)$, which is much larger since $d \gg k$ under our assumption. But, we may still apply the intermediate result (2.3.18) from that proof, which gives

$$\begin{aligned}&\le 2\exp\left(3k - \frac{1}{8}\min\left\{\frac{\delta^2 d^2}{4d}, \frac{\delta d}{2}\right\}\right) \\ &= 2\exp\left(3k - \frac{\delta^2}{32}d\right),\end{aligned} \tag{2.5.11}$$

using that $\delta < 1$ so $\delta^2/4 = (\delta/2)^2 < \delta/2$. Returning to the union bound, we first recall the non-asymptotic bounds on binomial coefficients

$$\left(\frac{m}{k}\right)^k \le \binom{m}{k} \le \left(\frac{em}{k}\right)^k,$$

which we will use twice below. We may bound the terms in the union bound uniformly as

$$\begin{aligned}\mathbb{P}[\boldsymbol{G} \text{ does not have the } (k,\delta)\text{-RIP}] &\le \binom{m}{k}\cdot 2\exp\left(3k - \frac{\delta^2}{32}d\right) \\ &\le \left(\frac{em}{k}\right)^k \cdot 2\exp\left(3k - \frac{\delta^2}{32}d\right) \\ &\le 2\exp\left(k\left(1 + \log\left(\frac{m}{k}\right)\right) + 3k - \frac{\delta^2}{32}d\right) \\ &\le 2\exp\left(4k\left(1 + \log\left(\frac{m}{k}\right)\right) - \frac{\delta^2}{32}d\right) \\ &\le 2\exp\left(-k\left(1 + \log\left(\frac{m}{k}\right)\right)\right) \\ &\le 2\left(\frac{em}{k}\right)^{-k} \\ &\le \frac{2}{\binom{m}{k}}\end{aligned} \tag{2.5.12}$$

completing the proof. The alternate form of the bound follows from the lower bound on binomial coefficients above. □

Bibliography

[AC09] Nir Ailon and Bernard Chazelle. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

[BKL18] Paul Breiding, Khazhgali Kozhasov, and Antonio Lerario. On the geometry of the set of symmetric matrices with repeated eigenvalues. *Arnold Mathematical Journal*, 4(3):423–443, 2018.

[DKS10] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 341–350, 2010.

[IM98] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998.

[JL82] William Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Conference in Modern Analysis and Probability*, 26:189–206, 01 1982.

[KN14] Daniel M Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):1–23, 2014.

[LN14] Kasper Green Larsen and Jelani Nelson. The johnson-lindenstrauss lemma is optimal for linear dimensionality reduction. *arXiv preprint arXiv:1411.2404*, 2014.

[LN17] Kasper Green Larsen and Jelani Nelson. Optimality of the johnson-lindenstrauss lemma. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 633–638. IEEE, 2017.

[RH17] Phillippe Rigollet and Jan-Christian Hütter. Lecture notes on high dimensional statistics. 2017.

[Tal05] Michel Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer, 2005.