Lecture Notes on Random Matrix Theory in Data Science and Statistics

Dmitriy (Tim) Kunisky

Fall 2026 (Last Updated: September 22, 2026)

Contents

1 Random Vector Theory **2**
- 1.1 Independent Entries and Subgaussian Concentration . . . 2
- 1.2 Gaussian Vectors and Subexponential Concentration . . . 5
- 1.3 Orthogonal Invariance . . . 8
- 1.4 Exercises . . . 10
  - 1.4.1 High-Dimensional Geometry . . . 10
  - 1.4.2 Concentration Inequalities . . . 11
  - 1.4.3 Orthogonal Invariance for Vectors . . . 13

2 Geometry of Rectangular Random Matrices **14**
- 2.1 Gaussian Random Field Viewpoint . . . 14
- 2.2 The **Geometric Method** of Random Matrix Theory . . . 17
- 2.3 Singular Values . . . 18
  - 2.3.1 Applying the Geometric Method . . . 18
  - 2.3.2 Nets, Coverings, and Packings . . . 19
  - 2.3.3 Coarse Non-Asymptotic Bound on Singular Values . . . 21
- 2.4 Singular Vectors . . . 24
  - 2.4.1 Distinctness of Singular Values . . . 25
  - 2.4.2 Orthogonal Invariance for Matrices and Haar Measures . . . 27
  - 2.4.3 Connection to Random Projections . . . 30
- 2.5 Application: Compressed Sensing . . . 31
  - 2.5.1 Null Space and Restricted Isometry Properties . . . 31
  - 2.5.2 Random Sensing Matrices . . . 32
- 2.6 Application: Neural Network Initialization . . . 34
  - 2.6.1 General Setup . . . 34
  - 2.6.2 Initialization and Local Analysis . . . 35
  - 2.6.3 Jacobian Calculations and Kaiming Initialization . . . 36

Bibliography **38**

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.

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. These are random vectors with *independent and identically distributed (i.i.d.) entries* and random vectors *uniformly random on the sphere*, respectively. In this chapter, we will look at these models and their behaviors, and alongside them introduce some of the basic tools and intuitions for probability and geometry in high dimension.

1.1 Independent Entries and Subgaussian Concentration

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 $\mathbb{R}$. The symbol $\otimes$ here and throughout will denote forming the product measure, i.e., the above notation means $x_i \sim \mu$ are drawn i.i.d. The only exception is when it is applied to vectors or matrices, in which case it denotes the Kronecker or tensor product; which one we mean should always be clear from context.

The behavior of such vectors is closely linked to *high-dimensional geometry*. Consider, for example, the case $\mu = \mathsf{Unif}([-1,1])$. In this case, $\boldsymbol{x}$ is drawn uniformly at random from the solid box $[-1,1]^d \subset \mathbb{R}^d$, obeying the measure $\mathsf{Unif}([-1,1]^d)$, and so properties that $\boldsymbol{x}$ has with high probability can equivalently be viewed as properties that "most" vectors in this box have, with respect to Lebesgue measure of volume.

We will see that this measure behaves rather paradoxically in high dimension. If you imagine a picture of the measures $\mathsf{Unif}([-1,1]^d)$ for $d = 1,2,3$ as densities, they look like solid convex "blobs" of mass. But, this intuition breaks down in high dimension: for large d, the mass of the box is instead "fragmented" and "hollow", and in fact this follows from basic statements in scalar probability theory that you have likely encountered already.

[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.

Consider the radius at which the random point $\boldsymbol{x}$ lies,

$$R^2 := \|\boldsymbol{x}\|^2 = \sum_{i=1}^{d} x_i^2 =: \sum_{i=1}^{d} Y_i.$$

This is a sum of i.i.d. random variables $Y_i = x_i^2$, precisely the situation to which the law of large numbers (LLN), central limit theorem (CLT), large deviations principles (LDP), and other tools of scalar probability apply. For instance, the law of large numbers gives

$$\frac{1}{d}R^2 = \frac{1}{d}\sum_{i=1}^{d} Y_i \xrightarrow[d\to\infty]{(\mathbb{P})} \mathbb{E}Y_1 = \frac{1}{3},$$

the last equality following by a simple integral calculation. Thus already we see that, in high dimension, $\boldsymbol{x}$ usually lies close to a sphere of radius $\frac{1}{\sqrt{3}}\sqrt{d} \approx 0.577\sqrt{d}$.

The central limit theorem gives more detail, stating that

$$\frac{1}{\sqrt{d}}\left(R^2 - \frac{d}{3}\right) = \frac{1}{\sqrt{d}}\sum_{i=1}^{d}(Y_i - \mathbb{E}Y_i) \xrightarrow[d\to\infty]{(\text{law})} \mathcal{N}(0, \mathsf{Var}[Y_1]) = \mathcal{N}\left(0, \frac{4}{45}\right),$$

where $\frac{4}{45} = \frac{1}{5} - (\frac{1}{3})^2$ comes from similar integral calculations to before. We will not care that much about the detailed constants here; the important part is that the fluctuations of R^2 around $\frac{1}{3}d$ are of order $O(\sqrt{d})$. Thus heuristically we have $R^2 = \frac{1}{3}d + O(\sqrt{d})$, and so $R = \frac{1}{\sqrt{3}}\sqrt{d} + O(1)$.[2]

Most important for our purposes will be *non-asymptotic* concentration inequalities, which give concrete bounds on probabilities of deviation from the mean as a function of d. We develop some of the basic tools for working with such inequalities in this section, the next section, and the chapter exercises.

Our basic goal is to show bounds of the form

$$\mathbb{P}\left[\left|R^2 - \frac{d}{3}\right| \geq t\right] \leq \delta(t, d)$$

for some function $\delta(t, d)$ as small as possible. You could try using any inequalities you are already aware of for this purpose. For example, Markov's inequality gives

$$\mathbb{P}\left[R^2 \geq \frac{d}{3} + t\right] \leq \frac{\frac{d}{3}}{\frac{d}{3} + t} = \frac{1}{1 + \frac{3t}{d}}, \tag{1.1.2}$$

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

$$\mathbb{P}\left[R^2 \geq \frac{d}{3} + t\right] = \mathbb{P}[R^2 \geq \mathbb{E}R^2 + t] \leq \frac{\mathsf{Var}[R^2]}{t^2} = \frac{\frac{4}{45}d}{t^2}, \tag{1.1.3}$$

where we calculate $\mathsf{Var}[R^2] = d\,\mathsf{Var}[Y_1] = \frac{4}{45}d$. 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.

Sharper such inequalities require some more tools. We say a random scalar $Y \in \mathbb{R}$ is *centered* if $\mathbb{E}Y = 0$ and similarly for random vectors.

[2] To see this, write $R = \frac{1}{\sqrt{3}}\sqrt{d} \cdot \sqrt{1 + \frac{3}{d}(R^2 - \frac{1}{3}d)}$ and take a Taylor expansion of the square root.

Definition 1.1.1 (Subgaussianity)**.** A centered random scalar $Y \in \mathbb{R}$ is σ^2*-subgaussian* if $\mathbb{E}\exp(\lambda Y) \leq \exp(\frac{\sigma^2}{2}\lambda^2)$ for all $\lambda \in \mathbb{R}$.

The function

$$\phi_Y(\lambda) := \mathbb{E}\exp(\lambda Y)$$

is called the *moment generating function* of Y. The terminology "subgaussian" is justified by the following.

Proposition 1.1.2. If $Y \sim \mathcal{N}(0, \sigma^2)$, then for all $\lambda \in \mathbb{R}$ we have $\phi_Y(\lambda) = \exp(\frac{\sigma^2}{2}\lambda^2)$.

Exercise 1.4.5 will develop some further tools for working with subgaussianity.

The following is a form of *Hoeffding's inequality* giving concentration for sums of subgaussian random variables.

Lemma 1.1.3 (Subgaussian sum concentration)**.** Suppose that $Y_1, \ldots, Y_d$ are independent and each is centered and σ^2-subgaussian. Then, for all $t \geq 0$,

$$\mathbb{P}\left[\left|\sum_{i=1}^d Y_i\right| \geq t\right] \leq 2\exp\left(-\frac{1}{2\sigma^2} \cdot \frac{t^2}{d}\right).$$

Proof. We use a general proof technique that you can call the "Chernoff method" or the "nonlinear Markov inequality" that is behind the proofs of all Chebyshev-, Hoeffding-, and Chernoff-like inequalities. The idea is to "preprocess" a random variable with a nonlinearity before using Markov's inequality. In our case, for any $\lambda > 0$ we have

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^d Y_i \geq t\right] &= \mathbb{P}\left[\exp\left(\lambda \sum_{i=1}^d Y_i\right) \geq \exp(\lambda t)\right] \\
&\leq \frac{\mathbb{E}[\exp(\lambda \sum_{i=1}^d Y_i)]}{\exp(\lambda t)} && \text{(Markov's inequality)} \\
&= \frac{\prod_{i=1}^d \mathbb{E}[\exp(\lambda Y_i)]}{\exp(\lambda t)} && \text{(independence)} \\
&\leq \exp\left(\frac{\sigma^2 d}{2}\lambda^2 - t\lambda\right). && \text{(subgaussianity)}
\end{aligned}$$

Optimizing over λ, we choose $\lambda = \frac{t}{\sigma^2 d}$, and obtain

$$\mathbb{P}\left[\sum_{i=1}^d Y_i \geq t\right] \leq \exp\left(-\frac{1}{2\sigma^2} \cdot \frac{t^2}{d}\right),$$

and applying the same argument to the lower tail and using the union bound gives the result. □

In particular, if σ^2 is a constant as d grows, then this probability indeed becomes small once $t \gg \sqrt{d}$, as our softer argument with the central limit theorem above predicted. As a corollary, we find the following concrete concentration inequality about the uniform measure over the box.

Corollary 1.1.4 (Typical set of high-dimensional cube)**.** Let $\boldsymbol{x} \sim \mathsf{Unif}([-1,1]^d)$ and $R^2 = \|\boldsymbol{x}\|^2$. Then, for all $t \geq 0$,

$$\mathbb{P}\left[\left|R^2 - \frac{d}{3}\right| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{d}\right).$$

This follows from Lemma 1.1.3 together with the following, which follows from a calculus argument that you can find in [RH17, Lemma 1.8].

Lemma 1.1.5 (Hoeffding's lemma)**.** Suppose that $Y \in \mathbb{R}$ is centered with $Y \in [A,B]$ almost surely. Then, Y is σ^2-subgaussian for $\sigma^2 = (\frac{B-A}{2})^2$.

Above we use this on $Y_i = x_i^2 - \frac{1}{3}$, so we can take $[A,B] = [-\frac{1}{3}, \frac{2}{3}]$, giving the specific constants above.

Thus, we see that most of the volume of the box $[-1,1]^d$ lies around a thin spherical *shell* of thickness $O(1)$ around the sphere of radius $\frac{1}{\sqrt{3}}\sqrt{d}$. In particular, you should not think of this typical set[3] as being convex anymore; further, you should think of it as being rather close to the corners of the cube, which lie at distance $\sqrt{d}$ from the origin, very far from the inscribed solid unit ball $\mathbb{B}^d := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq 1\}$. We give a schematic picture of this in two dimension in Figure 1.1.

Another way to phrase this phenomenon is that, for $d > 3$, the volume of $\mathbb{B}^d$ relative to its circumscribing box $[-1,+1]^d$ is bounded by

$$\begin{aligned}\frac{\mathsf{vol}(\mathbb{B}^d)}{\mathsf{vol}([-1,+1]^d)} &= \mathop{\mathbb{P}}_{\boldsymbol{x}\sim\mathsf{Unif}([-1,+1]^d)}[\|\boldsymbol{x}\| \leq 1] \\ &\leq \mathop{\mathbb{P}}_{\boldsymbol{x}\sim\mathsf{Unif}([-1,+1]^d)}[|\|\boldsymbol{x}\|^2 - d/3| \geq d/3 - 1] \\ &= \exp(-\Omega(d)). \end{aligned} \tag{1.1.4}$$

Thus, these same probabilistic tools show that the volume of the unit ball is exponentially smaller than the volume of its circumscribing box.[4]

1.2 Gaussian Vectors and Subexponential Concentration

The above model is useful for getting a sense of high-dimensional geometry, but the most important kind of random vector is the following, that you have likely seen before.

Definition 1.2.1 (Gaussian random vector)**.** $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for parameters $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}_{\succeq \mathbf{0}}$ is the probability measure on $\mathbb{R}^d$ 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 $\boldsymbol{\Sigma}$ translated by $\boldsymbol{\mu}$. Here, $\boldsymbol{\Sigma}^+$ is the

[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] Actually, as you may look up, the true volume is even smaller than that, but to show that requires a more careful calculation.

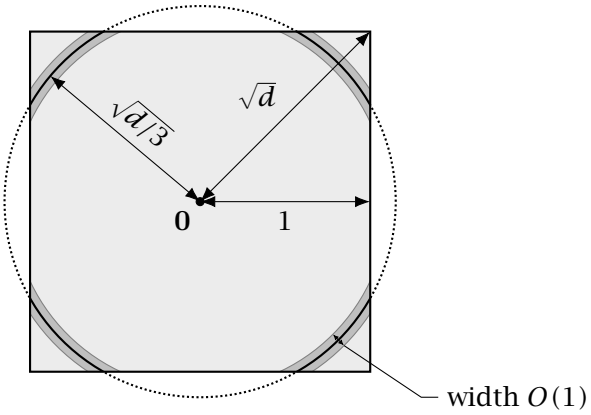


Figure 1.1: Schematic picture of the cube $[-1,1]^d$ and the thin spherical shell giving the typical set of its uniform measure, per Corollary 1.1.4. Distances are labeled for large dimension d and are not drawn to scale.

Moore-Penrose pseudoinverse, and $\det^*$ is the product of all non-zero eigenvalues. When Σ is invertible (the most typical situation for us), then $\det^*$ above is the ordinary determinant, $\Sigma^+ = \Sigma^{-1}$, 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 $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ for some $\boldsymbol{\mu}$ and Σ, and a *standard Gaussian random vector* if $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = \boldsymbol{I}_d$.

In particular, we will spend lots of time working with the standard Gaussian random vector, which is equivalently $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d) = \mathcal{N}(0,1)^{\otimes d}$.

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

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

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

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

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 (Affine closure)**.** 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 $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, then the corresponding parameters after this linear transformation are

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

These two facts taken together make it easy to do certain computations with Gaussian random vectors: provided we only make affine transformations, we both remain within the class of Gaussian

random vectors and can keep track of all distributions involved merely by linear algebra.[5]

We will see in the next section the key symmetry that makes the standard Gaussian random vector so fundamental. For now, let us take it for granted that this is an important distribution to understand and try to replay the concentration argument from before. The following calculation shows that something will go wrong:

Proposition 1.2.4. Suppose $X \sim \mathcal{N}(0,1)$ and $Y = X^2$. Then,

$$\phi_Y(\lambda) = \mathbb{E}\exp(\lambda Y) = \left\{ \begin{array}{ll} (1-2\lambda)^{-1/2} & \text{if } \lambda < 1/2, \\ +\infty & \end{array} \right\}.$$

In particular, Y is not σ^2-subgaussian for any choice of $\sigma^2 > 0$.

We can still repeat the argument of Lemma 1.1.3, but we will have to be more careful to choose a value of λ where we can control the above expression. This situation, which happens often with sums of heavier-tailed random variables, motivates the following definition and argument. We use a two-parameter version of [RH17, Definition 1.11].

Definition 1.2.5 (Subexponentiality)**.** Let $\sigma^2 > 0$ and $K > 0$. A centered random scalar $Y \in \mathbb{R}$ is (σ^2, K)*-subexponential* if

$$\mathbb{E}[\exp(\lambda Y)] \le \exp\left(\frac{\sigma^2}{2}\lambda^2\right) \text{ for all } |\lambda| \le K.$$

One way to think of this is as a form of *local subgaussianity*: σ^2 is the variance proxy, and K is the radius of the interval $[-K, K]$ on which the subgaussian bound holds. The corresponding bound for sums is a form of *Bernstein's inequality* (compare with [RH17, Theorem 1.13]).

Lemma 1.2.6 (Subexponential sum concentration)**.** Suppose that $Y_1, \dots, Y_d$ are independent and each is (σ^2, K)-subexponential. Then, for all $t \ge 0$,

$$\mathbb{P}\left[\left|\sum_{i=1}^d Y_i\right| \ge t\right] \le 2\exp\left(-\min\left\{\frac{1}{2\sigma^2}\cdot\frac{t^2}{d}, \frac{K}{2}\cdot t\right\}\right).$$

Proof. The same argument as in Lemma 1.1.3 gives, for $0 < \lambda \le K$,

$$\mathbb{P}\left[\sum_{i=1}^d Y_i \ge t\right] \le \exp\left(\frac{d\sigma^2}{2}\lambda^2 - t\lambda\right).$$

Now, however, we cannot choose $\lambda > K$, so for $t > 0$ we instead optimize by taking

$$\lambda := \min\left\{\frac{t}{d\sigma^2}, K\right\}.$$

For this choice, $d\sigma^2\lambda \le t$, so the exponent is at most $-t\lambda/2$. Applying the same argument to the lower tail and using the union bound gives the result. The case $t = 0$ is immediate. □

[5] 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.

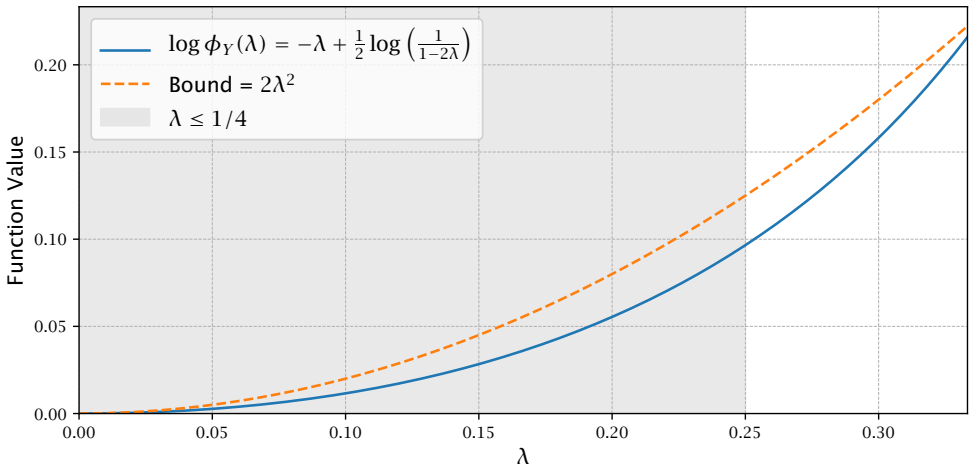


Figure 1.2: The cumulant generating function of the random variable Y appearing in Proposition 1.2.8, our bound on it, and the region where we apply this bound.

Thus the bound is Gaussian when $t \le d\sigma^2 K$ and exponential when $t \ge d\sigma^2 K$.

The following is the corollary of this general result for specifically standard Gaussian vectors, which we will use several times.

Corollary 1.2.7. For $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ and any $t \ge 0$,

$$\begin{aligned} \mathbb{P}\left[\left|\|\boldsymbol{x}\|^2 - d\right| \ge t\right] &= \mathbb{P}\left[\left|\sum_{i=1}^{d} x_i^2 - d\right| \ge t\right] \\ &\le 2\left\{\begin{array}{ll} \exp\left(-\frac{t^2}{8d}\right) & \text{if } t \le d, \\ \exp\left(-\frac{t}{8}\right) & \text{if } t \ge d \end{array}\right\} \qquad (1.2.5) \\ &= 2\exp\left(-\frac{1}{8}\min\left\{\frac{t^2}{d}, t\right\}\right). \qquad (1.2.6) \end{aligned}$$

This follows from Lemma 1.2.6 together with the following bound on the result of Proposition 1.2.4.

Proposition 1.2.8. For all $\lambda \le \frac{1}{4}$, $\exp(-\lambda)(1-2\lambda)^{-1/2} \le \exp(2\lambda^2)$. Thus, if $X \sim \mathcal{N}(0,1)$, then $Y = X^2 - 1$ is $(4, \frac{1}{4})$-subexponential.

We illustrate this estimate in Figure 1.2.

1.3 Orthogonal Invariance

Aside from independence of entries, the following is the other simple symmetry property a random vector can have.

Definition 1.3.1 (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 $\mathsf{Law}(\boldsymbol{Q}\boldsymbol{x}) = \mathsf{Law}(\boldsymbol{x})$.

Without going into more technical details of actually defining this probability measure explicitly, the following result states that there is a well-defined uniform 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).

Proposition 1.3.2. 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 1.3.3. Suppose that $\boldsymbol{x}$ is orthogonally invariant and $\boldsymbol{x} \neq \boldsymbol{0}$ almost surely. Then, $\|\boldsymbol{x}\|$ is independent of $\boldsymbol{x}/\|\boldsymbol{x}\|$, and $\mathsf{Law}(\boldsymbol{x}/\|\boldsymbol{x}\|) = \mathsf{Unif}(\mathbb{S}^{d-1})$.

The standard Gaussian random vector turns out to be a useful bridge between independent entries and concentration inequalities on the one hand and these measures on the other.

Proposition 1.3.4. 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 Proposition 1.2.3, for $\boldsymbol{Q} \in \mathcal{O}(d)$,

$$\mathsf{Law}(\boldsymbol{Q}\boldsymbol{x}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^\top \boldsymbol{I}_d \boldsymbol{Q}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d) = \mathsf{Law}(\boldsymbol{x}), \tag{1.3.1}$$

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

So, a standard Gaussian random vector is *both* an i.i.d. random vector *and* an orthogonally invariant one. Explicitly, we have the following, which also has the useful practical interpretation of giving an explicit way to sample uniformly at random from spheres of any dimension.

Corollary 1.3.5. 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 1.3.3. □

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.

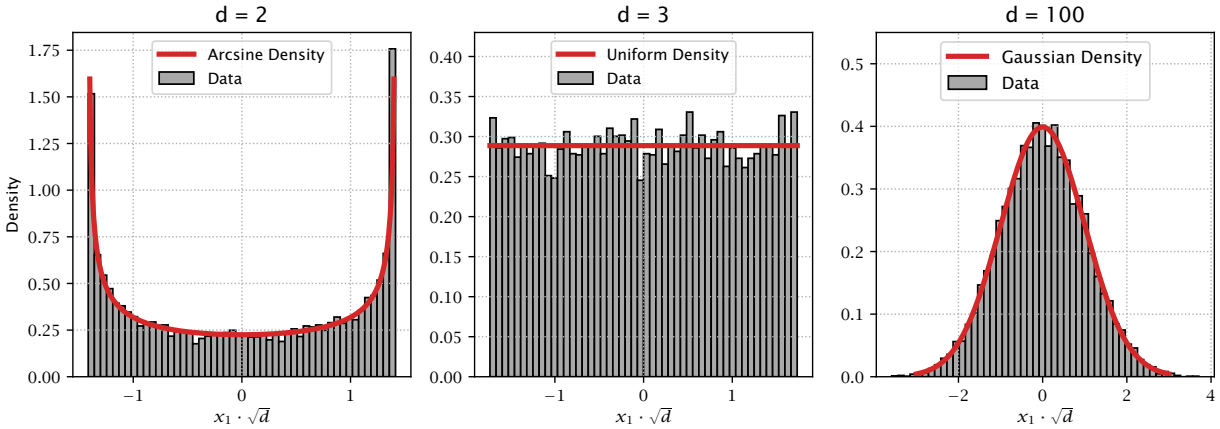


Figure 1.3: An illustration of Borel's theorem, Theorem 1.3.7, showing two low-dimensional examples $d = 2, 3$ where the distribution of $x_1 \cdot \sqrt{d}$ is known exactly (and is not Gaussian) as well as the Gaussian limit for large d.

Theorem 1.3.6 (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 $\sigma^2 \geq 0$.

These observations together give some of the fundamental intuitions of high-dimensional geometry. Analogously to random vectors in the box, Corollary 1.2.7 says that $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ has $\|\boldsymbol{x}\|^2 = d + O(\sqrt{d})$ with high probability, or $\|\boldsymbol{x}\| = \sqrt{d} + O(1)$. That is, a random standard Gaussian vector usually falls close to the spherical shell of width $O(1)$ around the sphere of radius $\sqrt{d}$. As in the case of the solid box, this typical set is non-convex and instead "hollow" and concentrated far from the origin.

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

$$\text{“ } \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d) \approx \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d})) \text{ ”} \tag{1.3.2}$$

We arrive at this 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 1.3.5) and $\|\boldsymbol{x}\| \approx \sqrt{d}$ (Corollary 1.2.7). This idea can be made rigorous, for instance in the following result.

Theorem 1.3.7 (Borel)**.** Let $\boldsymbol{x} \sim \mathsf{Unif}(\mathbb{S}^{d-1})$. Then, 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 $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$.

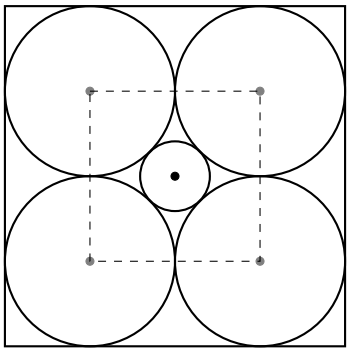
1.4 Exercises

1.4.1 High-Dimensional Geometry

Exercise 1.4.1. Consider the box B in $\mathbb{R}^d$ of side length 2, centered at the origin, with vertices at the points $(\pm 1, \ldots, \pm 1)$. For each $\boldsymbol{s} \in \{\pm 1\}^d$, let $S_{\boldsymbol{s}}$ be the sphere of radius $\frac{1}{2}$ centered at the point $\frac{1}{2}\boldsymbol{s}$.

These are 2^d spheres packed on a cubic lattice into the box B. Consider the sphere S' centered at the origin that is tangent to every S_s. Find d_0 such that, if $d < d_0$, then S' is contained in B, but if $d \geq d_0$, then S' is *not* contained in B.

The case $d = 2$ looks as follows. The innermost circle is S', and the four around it are $S_{(\pm 1, \pm 1)}$. The larger outermost square is B; the smaller square in a dotted line is just for reference to show how the centers of the latter circles are arranged. Your task is to show that, in high dimension, the innermost circle is not contained in the outermost square (!).



Exercise 1.4.2. Show that there is a constant $c > 0$ such that, for all $\epsilon \in (0, 1)$, for d sufficiently large, there are at least $N = \exp(c\epsilon^2 d)$ unit vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ in $\mathbb{R}^d$ such that $|\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle| \leq \epsilon$ for all $1 \leq i < j \leq N$ (i.e., such that the $\boldsymbol{v}_i$ are almost orthogonal).

(**Hint:** Consider random vectors. Choose a convenient distribution to work with. Note, though, that the question is not making a probabilistic statement.)

Exercise 1.4.3. Consider the shape $\Delta^d \subset \mathbb{R}^d$ that is the convex hull of the points $\mathbf{0}, \boldsymbol{e}_1, \boldsymbol{e}_1 + \boldsymbol{e}_2, \ldots, \boldsymbol{e}_1 + \boldsymbol{e}_2 + \cdots + \boldsymbol{e}_d$ (the origin plus the "partial sums" of the standard basis vectors). This is a *simplex* or high-dimensional tetrahedron, though not an equilateral one: edges have lengths varying among $1 = \sqrt{1}, \sqrt{2}, \ldots, \sqrt{d}$. Compute the volume of Δ^d. What is the side length of a cube in $\mathbb{R}^d$ with the same volume? (Give an asymptotic approximation as $d \to \infty$.)

(**Hint:** For the volume computation, consider gluing several copies of Δ^d together to tile a more familiar object.)

Exercise 1.4.4. Suppose that $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. Show the non-asymptotic inequalities

$$\left(1 - \frac{2}{d}\right)\sqrt{d} \leq \mathbb{E}\|\boldsymbol{g}\| \leq \sqrt{d}.$$

(**Hint:** For the lower bound, that $\sqrt{x} \geq \frac{1}{2}(1 + x - (x-1)^2)$ for all $x \geq 0$. Apply this with $x := \|\boldsymbol{g}\|^2/d$.)

1.4.2 Concentration Inequalities

Exercise 1.4.5 (Subgaussianity)**.** Recall that a centered random scalar $X \in \mathbb{R}$ is called σ^2-*subgaussian* if $\mathbb{E}\exp(\lambda X) \leq \exp(\sigma^2\lambda^2/2)$ for all $\lambda \in \mathbb{R}$. A centered random vector $\boldsymbol{x} \in \mathbb{R}^d$ is called the same if, for all $\boldsymbol{u} \in \mathbb{R}^d$ with $\|\boldsymbol{u}\| = 1$, $\langle \boldsymbol{u}, \boldsymbol{x} \rangle$ is a σ^2-subgaussian random scalar.

1. Show that if X is a centered random scalar that is σ^2-subgaussian, then aX is $a^2\sigma^2$-subgaussian, and that if $X_1, \ldots, X_n$ are independent and each X_i is σ_i^2-subgaussian, then $\sum_{i=1}^n X_i$ is $(\sum_{i=1}^n \sigma_i^2)$-subgaussian. Conclude that a random vector with independent entries that are each centered and σ^2-subgaussian is itself σ^2-subgaussian.

2. Show that the following three conditions on a centered random scalar X are equivalent:

   (a) There exists $\sigma^2 > 0$ such that X is σ^2-subgaussian.

   (b) There exists $c > 0$ such that, for all $t > 0$, $\mathbb{P}[|X| \geq t] \leq 2\exp(-ct^2)$.

   (c) There exists $\lambda > 0$ such that $\mathbb{E}\exp(\lambda X^2) < \infty$.

3. For X a centered subgaussian scalar, define

$$\|X\|_{\mathsf{sg}} := \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\}.$$

   Note that this is well-defined by Part 2. Show that $\|\cdot\|_{\mathsf{sg}}$ is a norm on the space of centered subgaussian random variables (over a given probability space).

   (**Hint:** Prove and use that $(X+Y)^2 \leq (1+a)X^2 + (1+\frac{1}{a})Y^2$ for any $a > 0$ (you might be more familiar with the case $a = 1$), together with Hölder's inequality.)

Exercise 1.4.6 (Subexponentiality)**.** Recall that, for $\sigma^2, K > 0$, a centered random scalar $X \in \mathbb{R}$ is called (σ^2, K)-*subexponential* if

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\sigma^2\lambda^2/2) \text{ for all } |\lambda| \leq K.$$

A centered random vector $\boldsymbol{x} \in \mathbb{R}^d$ is called the same if, for every $\boldsymbol{u} \in \mathbb{R}^d$ with $\|\boldsymbol{u}\| = 1$, $\langle \boldsymbol{u}, \boldsymbol{x}\rangle$ is a (σ^2, K)-subexponential random scalar.

1. Show that multiplying a (σ^2, K)-subexponential scalar X by $a \neq 0$ gives a subexponential scalar with parameters $(a^2\sigma^2, K/|a|)$. If $X_1, \ldots, X_n$ are independent and each X_i is (σ_i^2, K_i)-subexponential, show that $\sum_{i=1}^n X_i$ is subexponential with parameters

$$\left(\sum_{i=1}^n \sigma_i^2, \min_{1\leq i\leq n} K_i\right).$$

   Conclude that a random vector with independent (σ^2, K)-subexponential entries is (σ^2, K)-subexponential.

2. Show that the following three conditions on a centered random scalar X are equivalent:

   (a) There exist $\sigma^2, K > 0$ such that X is (σ^2, K)-subexponential.

   (b) There exists $c > 0$ such that, for all $t > 0$, $\mathbb{P}[|X| \geq t] \leq 2\exp(-ct)$.

   (c) There exists $\lambda > 0$ such that $\mathbb{E}[\exp(\lambda|X|)] < \infty$.

3. For X a centered subexponential scalar, define

$$\|X\|_{\mathsf{se}} := \inf\{t > 0 : \mathbb{E}[\exp(|X|/t)] \leq 2\}.$$

   Note that this is well-defined by Part 2. Show that $\|\cdot\|_{\mathsf{se}}$ is a norm on the space of centered subexponential random variables (over a given probability space). Show also that, for every centered subgaussian scalar X,

$$\|X^2 - \mathbb{E}[X^2]\|_{\mathsf{se}} \leq 2\|X\|_{\mathsf{sg}}^2,$$

   where $\|\cdot\|_{\mathsf{sg}}$ is defined in Exercise 1.4.5.

   (**Hint:** Use Hölder's inequality for the triangle inequality and Jensen's inequality for the bound on centered squares.)

1.4.3 Orthogonal Invariance for Vectors

Exercise 1.4.7 (Factorization of distribution)**.** Suppose $\boldsymbol{x} \in \mathbb{R}^d$ is an orthogonally invariant random vector, i.e., one having $\mathsf{Law}(\boldsymbol{Q}\boldsymbol{x}) = \mathsf{Law}(\boldsymbol{x})$ for all $\boldsymbol{Q} \in \mathcal{O}(d)$. Suppose also that $\boldsymbol{x} \neq \boldsymbol{0}$ almost surely. Show that the scalar $\|\boldsymbol{x}\|$ and the vector $\widehat{\boldsymbol{x}} := \boldsymbol{x}/\|\boldsymbol{x}\|$ are independent random variables, i.e., that $\mathbb{P}[\|\boldsymbol{x}\| \in A, \widehat{\boldsymbol{x}} \in B] = \mathbb{P}[\|\boldsymbol{x}\| \in A] \cdot \mathbb{P}[\widehat{\boldsymbol{x}} \in B]$ for all pairs of Borel sets $A \subseteq \mathbb{R}_{\geq 0}$ and $B \subseteq \mathbb{S}^{d-1}$.

Exercise 1.4.8 (Moments)**.** Suppose $\boldsymbol{x} \in \mathbb{R}^d$ is an orthogonally invariant random vector whose entries have finite first and second moments. Show that $\mathbb{E}\boldsymbol{x} = \boldsymbol{0}$ and $\mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top = \sigma^2 \boldsymbol{I}_d$ for some $\sigma^2 \geq 0$.

2 | Geometry of Rectangular Random 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

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

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 $\mathsf{Law}(\widehat{\boldsymbol{G}}\boldsymbol{y}) = \mathcal{N}(\boldsymbol{0}, \frac{1}{d}\|\boldsymbol{y}\|^2\boldsymbol{I}_d)$, to which Corollary 1.2.7 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 $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$, 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.)

The next step is to understand the singular values of $\boldsymbol{G}$. This will require controlling a quadratic form uniformly over all unit vectors, using a finite approximation of the sphere or ball called an ϵ-net.

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

We will now move from the action of $\boldsymbol{G}$ on fixed vectors to its singular values. This will give us a first spectral analysis of rectangular matrices with i.i.d. Gaussian entries.

We will see below that this amounts to controlling the operator norm of a certain symmetric "error" matrix $\boldsymbol{M}$. The general approach we take to this task is one of the main methods of random matrix theory that we will study, so let us outline it separately here.

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:

$$\mathbb{B}^d(\boldsymbol{x}, r) := \{\boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{y} - \boldsymbol{x}\| \le r\}, \tag{2.2.1}$$

$$\mathbb{B}^d := \mathbb{B}^d(\boldsymbol{0}, 1) \supset \mathbb{S}^{d-1}. \tag{2.2.2}$$

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

$$\|\boldsymbol{M}\| = \sup_{\boldsymbol{x} \in \mathbb{S}^{d-1}} |\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}| = \sup_{\boldsymbol{x} \in \mathbb{B}^d} |\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}|. \tag{2.2.3}$$

In either case, if $\mathbb{S}^{d-1}$ or $\mathbb{B}^d$ were replaced by a finite set, we could use the union bound (also called the first moment method), bounding $\mathbb{P}[|\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}| > t]$ (which as we will see boils down to yet another application of Corollary 1.2.7 on the concentration of norms of Gaussian random vectors) and using a union bound to control $\mathbb{P}[\|\boldsymbol{M}\| > t]$, 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 2.2.1. For any nonempty bounded set $\mathcal{X} \subset \mathbb{R}^d$ and $\boldsymbol{M} \in \mathbb{R}^{d \times d}_{\mathsf{sym}}$, we define

$$\|\boldsymbol{M}\|_{\mathcal{X}} := \sup_{\boldsymbol{x} \in \mathcal{X}} |\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}|. \tag{2.2.4}$$

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

$$\mathbb{P}[\|\boldsymbol{M}\|_{\mathcal{X}} \ge t] \le |\mathcal{X}| \cdot \max_{\boldsymbol{x} \in \mathcal{X}} \mathbb{P}[|\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}| \ge t]. \tag{2.2.5}$$

As mentioned above, controlling the second factor above involving scalar probabilities will be possible in our case using Corollary 1.2.7; in general, it reduces to simpler questions involving scalar probability theory. 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 $\|\boldsymbol{M}\|$ by $\|\boldsymbol{M}\|_{\mathcal{X}}$, which we will soon see decreases as $|\mathcal{X}|$ increases. We will therefore reduce the matter of bounding $\|\boldsymbol{M}\|$ 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 Singular Values

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.

For any fixed unit vector $\boldsymbol{x} \in \mathbb{R}^m$, the vector $\boldsymbol{G}\boldsymbol{x}$ has law $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Thus Corollary 1.2.7 gives concentration of the quadratic form

$$\boldsymbol{x}^\top (\boldsymbol{G}^\top \boldsymbol{G} - d\boldsymbol{I}_m)\boldsymbol{x} = \|\boldsymbol{G}\boldsymbol{x}\|^2 - d. \tag{2.3.1}$$

When d is large, this is small compared with d with high probability. A union bound gives simultaneous control for a sufficiently small finite set of such test vectors. But, as we observed above, $\boldsymbol{G}^\top \boldsymbol{G}$ has rank at most d, so when $d < m$ it cannot be close to $d\boldsymbol{I}_m$ in operator norm relative to d. To understand the matrix itself, we will study its eigenvalues, or equivalently the 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 Applying the Geometric Method

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

$$\begin{aligned} &\approx m \cdot \mathbb{E}[\boldsymbol{g}_1 \boldsymbol{g}_1^\top] \\ &= m\boldsymbol{I}_d \\ &= \mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]. \end{aligned} \tag{2.3.2}$$

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 2.3.1. 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, when $m \gg d$, we expect the operator norm of the difference $\|\boldsymbol{G}\boldsymbol{G}^\top - m\boldsymbol{I}_d\|$ to be small compared with m. 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 2.3.2 (Statistical interpretation)**.** Expanding in a sum of rank one terms again, we may also view

$$\frac{1}{m}\boldsymbol{\Delta} = \left(\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{g}_i\boldsymbol{g}_i^\top\right) - \mathbb{E}[\boldsymbol{g}_1\boldsymbol{g}_1^\top] \tag{2.3.3}$$

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).

To show that $\|\boldsymbol{\Delta}\|$ is relatively small, we will apply the Geometric Method outlined in Section 2.2 with $\boldsymbol{M} = \boldsymbol{\Delta}$.

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 defining the operator norm.

Definition 2.3.3 (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

$$\mathcal{Y} \subseteq \bigcup_{i=1}^{N} \mathbb{B}^d(\boldsymbol{x}_i, \epsilon), \tag{2.3.4}$$

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

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

Lemma 2.3.4. 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 $0 < \epsilon < \frac{1}{2}$. Then,

$$\|\boldsymbol{M}\|_{\mathcal{X}} \le \|\boldsymbol{M}\| \le \frac{1}{1-2\epsilon}\|\boldsymbol{M}\|_{\mathcal{X}}. \tag{2.3.5}$$

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 $\|\boldsymbol{x} - \boldsymbol{x}_i\| \le \epsilon$. We then have

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

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

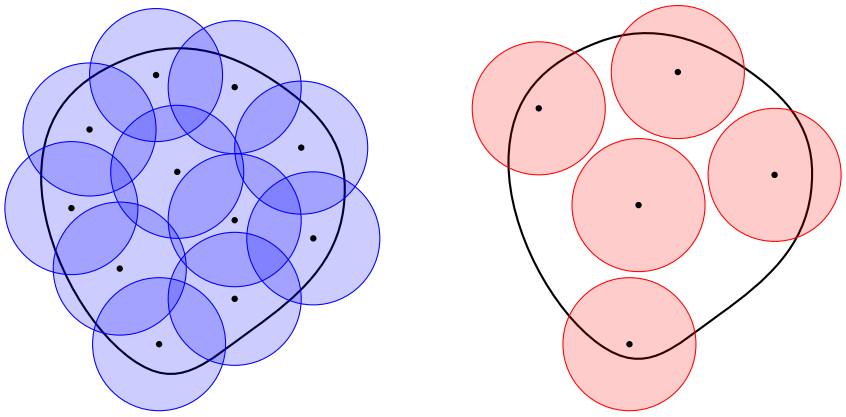


Figure 2.1: An illustration of a covering (left) and a packing (right) of a given set in $\mathbb{R}^2$ by balls of the same radius.

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 construct 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:

Definition 2.3.5 (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 2.3.6. 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 2.3.7. Let $\mathcal{X}$ be any ϵ-packing of $\mathbb{B}^d$. Then, $|\mathcal{X}| \leq (1 + \frac{1}{\epsilon})^d$.

Proof. We have

$$\bigcup_{\boldsymbol{x} \in \mathcal{X}} \mathbb{B}^d(\boldsymbol{x}, \epsilon) \subseteq \mathbb{B}^d(\boldsymbol{0}, 1 + \epsilon), \tag{2.3.6}$$

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

$$|\mathcal{X}| \cdot \mathsf{vol}(\mathbb{B}^d(\boldsymbol{0}, \epsilon)) \leq \mathsf{vol}(\mathbb{B}^d(\boldsymbol{0}, 1 + \epsilon)). \tag{2.3.7}$$

Rearranging,

$$|\mathcal{X}| \le \frac{\mathrm{vol}(\mathbb{B}^d(\mathbf{0}, 1+\epsilon))}{\mathrm{vol}(\mathbb{B}^d(\mathbf{0}, \epsilon))} = \left(\frac{1+\epsilon}{\epsilon}\right)^d = \left(1 + \frac{1}{\epsilon}\right)^d, \tag{2.3.8}$$

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

Corollary 2.3.8. For any $\epsilon > 0$, there exists an ϵ-net $\mathcal{X}$ of $\mathbb{B}^d$ of size $|\mathcal{X}| \le (1 + \frac{2}{\epsilon})^d$.

Proof. Let $\mathcal{X}$ be a maximal $\frac{\epsilon}{2}$-packing of $\mathbb{B}^d$. Such a packing exists: greedily add points while possible, and Proposition 2.3.7 guarantees that this process stops after finitely many steps. By Proposition 2.3.7, $|\mathcal{X}| \le (1 + \frac{2}{\epsilon})^d$, while by Proposition 2.3.6, $\mathcal{X}$ 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 2.3.9. If $d \le m$, then

$$\mathbb{P}[\|\boldsymbol{\Delta}\| > 64\sqrt{dm}] \le 2\exp(-d). \tag{2.3.9}$$

Proof. Let us fix $\epsilon := 1/4$. Let $\mathcal{X}$ be an ϵ-net of $\mathbb{B}^d$ as in Corollary 2.3.8, by which we can assume $|\mathcal{X}| \le (1 + \frac{2}{\epsilon})^d = 9^d$. By Lemma 2.3.4, we have $\|\boldsymbol{\Delta}\| \le \frac{1}{1-2\epsilon}\|\boldsymbol{\Delta}\|_{\mathcal{X}} = 2\|\boldsymbol{\Delta}\|_{\mathcal{X}}$. Thus, for $t > 0$, we may bound

$$\begin{aligned}
\mathbb{P}[\|\boldsymbol{\Delta}\| \ge t] &\le \mathbb{P}\left[\|\boldsymbol{\Delta}\|_{\mathcal{X}} \ge \frac{t}{2}\right] && \text{(Lemma 2.3.4)} \\
&\le 9^d \cdot \max_{\boldsymbol{x} \in \mathcal{X}} \mathbb{P}\left[|\boldsymbol{x}^\top \boldsymbol{\Delta} \boldsymbol{x}| \ge \frac{t}{2}\right]. && \text{(Proposition 2.2.2)}
\end{aligned}$$

For any fixed $\boldsymbol{x} \in \mathcal{X}$, the vector $\boldsymbol{G}^\top \boldsymbol{x}$ has law $\mathcal{N}(\mathbf{0}, \|\boldsymbol{x}\|^2 \boldsymbol{I}_m)$. Expanding the definition of $\boldsymbol{\Delta}$ and using $\|\boldsymbol{x}\| \le 1$, we therefore have

$$\begin{aligned}
\mathbb{P}\left[|\boldsymbol{x}^\top \boldsymbol{\Delta} \boldsymbol{x}| \ge \frac{t}{2}\right] &= \mathbb{P}\left[\left|\, \|\boldsymbol{G}^\top \boldsymbol{x}\|^2 - m\|\boldsymbol{x}\|^2 \,\right| \ge \frac{t}{2}\right] \\
&= \mathop{\mathbb{P}}_{\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)}\left[\|\boldsymbol{x}\|^2 \left|\, \|\boldsymbol{g}\|^2 - m \,\right| \ge \frac{t}{2}\right] \\
&\le \mathop{\mathbb{P}}_{\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)}\left[\left|\, \|\boldsymbol{g}\|^2 - m \,\right| \ge \frac{t}{2}\right].
\end{aligned} \tag{2.3.10}$$

The last bound is independent of $\boldsymbol{x}$, so we have

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

To the remaining probability we can apply Corollary 1.2.7, which it will be convenient to do specifically in the form (1.2.6). We can also bound $9 \le \exp(3)$. Together, these observations give

$$\le 2\exp\left(3d - \frac{1}{8}\min\left\{\frac{t^2}{4m}, \frac{t}{2}\right\}\right) \tag{2.3.11}$$

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

$$\begin{aligned} &\leq 2\exp\left(3d - \frac{1}{8}\min\left\{\frac{C^2}{4}d, \frac{C}{2}d\right\}\right) \\ &= 2\exp\left(d \cdot \left[3 - \min\left\{\frac{C^2}{32}, \frac{C}{16}\right\}\right]\right) \end{aligned} \tag{2.3.12}$$

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

$$\lambda_d(\boldsymbol{A}) - \|\boldsymbol{\Delta}\| \leq \lambda_d(\boldsymbol{A}+\boldsymbol{\Delta}) \leq \lambda_1(\boldsymbol{A}+\boldsymbol{\Delta}) \leq \lambda_1(\boldsymbol{A}) + \|\boldsymbol{\Delta}\|. \tag{2.3.13}$$

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

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

Similarly, for the lower bound,

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

□

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 2.3.11. 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

$$\sqrt{1+x} \le 1 + \frac{x}{2} \text{ for all } x \ge 0, \tag{2.3.14}$$

$$\sqrt{1-x} \ge 1 - x \text{ for all } 0 \le x \le 1. \tag{2.3.15}$$

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 $\|\boldsymbol{\Delta}\| \le 64\sqrt{dm}$ from Theorem 2.3.9, using that $\boldsymbol{G}\boldsymbol{G}^\top = m\boldsymbol{I}_d + \boldsymbol{\Delta}$, we have

$$\begin{aligned}
\sigma_1(\boldsymbol{G}) &= \sqrt{\lambda_1(\boldsymbol{G}\boldsymbol{G}^\top)} \\
&\le \sqrt{m + \|\boldsymbol{\Delta}\|} && \text{(Proposition 2.3.10)} \\
&\le \sqrt{m + 64\sqrt{dm}} \\
&= \sqrt{m} \cdot \sqrt{1 + 64\sqrt{\frac{d}{m}}} \\
&\le \sqrt{m} \cdot \left(1 + 32\sqrt{\frac{d}{m}}\right) && \text{(2.3.14)} \\
&= \sqrt{m} + 32\sqrt{d}.
\end{aligned}$$

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,

$$\begin{aligned}
\sigma_d(\boldsymbol{G}) &= \sqrt{\lambda_d(\boldsymbol{G}\boldsymbol{G}^\top)} \\
&\ge \sqrt{m - \|\boldsymbol{\Delta}\|} && \text{(Proposition 2.3.10)} \\
&\ge \sqrt{m - 64\sqrt{dm}} \\
&= \sqrt{m} \cdot \sqrt{1 - 64\sqrt{\frac{d}{m}}} \\
&\ge \sqrt{m} \cdot \left(1 - 64\sqrt{\frac{d}{m}}\right) && \text{(2.3.15)} \\
&= \sqrt{m} - 64\sqrt{d},
\end{aligned}$$

completing the proof. □

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

$$\boxed{\sqrt{m} - O(\sqrt{d}) \le \sigma_d(\boldsymbol{G}) \le \cdots \le \sigma_1(\boldsymbol{G}) \le \sqrt{m} + O(\sqrt{d})} \tag{2.3.16}$$

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 $\boldsymbol{G}^\top\boldsymbol{G} \in \mathbb{R}^{m\times m}_{\mathsf{sym}}$ 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 $\boldsymbol{G}$, which we will study next.

2.4 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 $\{\boldsymbol{u}_1,\dots,\boldsymbol{u}_d\} \subset \mathbb{R}^d$ and $\{\boldsymbol{v}_1,\dots,\boldsymbol{v}_d\} \subset \mathbb{R}^m$ 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

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

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

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

for any $\boldsymbol{Q} \in \mathcal{O}(2)$. This gives rise to equivalent decompositions of $\boldsymbol{Z}_i + \boldsymbol{Z}_j$ as a sum of rank-one matrices. Since $\sigma_i = \sigma_j$, the expression $\sigma_i(\boldsymbol{Z}_i + \boldsymbol{Z}_j)$ is what appears in the SVD for these two

components, so these equivalent decompositions mean that $\boldsymbol{Z}_i$ and $\boldsymbol{Z}_j$ 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 $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_d\}$. 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 $\boldsymbol{Z}_i$ 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 $f \in \mathbb{R}[t_1, \ldots, t_d]$ is *symmetric* if, for all permutations $\pi \in S_d$, we have $f(t_{\pi(1)}, \ldots, t_{\pi(d)}) = f(t_1, \ldots, t_d)$.

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

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

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 $\boldsymbol{X}$,

$$S_k(\boldsymbol{X}) := \mathsf{Tr}(\boldsymbol{X}^k) = \sum_{a_1, \ldots, a_k = 1}^{d} X_{a_1 a_2} X_{a_2 a_3} \cdots X_{a_{k-1} a_k} X_{a_k a_1}.$$

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 an algebraic set 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 non-zero polynomial F, we only know that its codimension is at least 1 (i.e., the dimension of this set 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 and Haar Measures

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 $\mathsf{Law}(\boldsymbol{Q}\boldsymbol{X}) = \mathsf{Law}(\boldsymbol{X})$.
- *Right-orthogonally invariant* if, for each (deterministic) orthogonal matrix $\boldsymbol{Q} \in \mathcal{O}(m)$, we have $\mathsf{Law}(\boldsymbol{X}\boldsymbol{Q}) = \mathsf{Law}(\boldsymbol{X})$.
- *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 $\mathsf{Law}(\boldsymbol{Q}\boldsymbol{X}\boldsymbol{Q}^\top) = \mathsf{Law}(\boldsymbol{X})$ 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.

To do this, we will show that after all it is possible to give a meaning to the left singular vectors "$\boldsymbol{u}_i(\boldsymbol{G})$" (and similarly for the right singular vectors), provided we allow for some further randomization.

Definition 2.4.14. Suppose $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d \times m}$ with $d \le m$. $\boldsymbol{Z}_i = \boldsymbol{Z}_i(\boldsymbol{G}) \in \mathbb{R}^{d \times m}$ is (almost surely) a well-defined rank-one matrix with $\sigma_1(\boldsymbol{Z}_i) = 1$. There then exist exactly two pairs of unit vectors $(\boldsymbol{u}_i^{(1)}, \boldsymbol{v}_i^{(1)}), (\boldsymbol{u}_i^{(2)}, \boldsymbol{v}_i^{(2)})$ such that $\boldsymbol{u}_i^{(a)}\boldsymbol{v}_i^{(a)\top} = \boldsymbol{Z}_i$, which are related by $\boldsymbol{u}_i^{(1)} = -\boldsymbol{u}_i^{(2)}$ and $\boldsymbol{v}_i^{(1)} = -\boldsymbol{v}_i^{(2)}$. Let us define further random variables $(\tilde{\boldsymbol{u}}_i, \tilde{\boldsymbol{v}}_i) \sim \mathsf{Unif}(\{(\boldsymbol{u}_i^{(1)}, \boldsymbol{v}_i^{(1)}), (\boldsymbol{u}_i^{(2)}, \boldsymbol{v}_i^{(2)})\})$.

These random variables have a well-defined law, and may be viewed as generated randomly from $\boldsymbol{G}$ along with one more bit of randomness to make the above choice. Finally, we define associated random matrices

$$\widetilde{\boldsymbol{U}} := \begin{bmatrix} | & & | \\ \tilde{\boldsymbol{u}}_1 & \cdots & \tilde{\boldsymbol{u}}_d \\ | & & | \end{bmatrix} \in \mathbb{R}^{d\times d}, \quad \widetilde{\boldsymbol{V}} := \begin{bmatrix} | & & | \\ \tilde{\boldsymbol{v}}_1 & \cdots & \tilde{\boldsymbol{v}}_d \\ | & & | \end{bmatrix} \in \mathbb{R}^{m\times d}.$$

The above is just a slightly elaborate way of formalizing that we break the sign ambiguity of $\boldsymbol{u}_i$ by choosing a random sign, and making that choice compatible between $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$. It is easy to check that

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

and we may also view the above as choosing a random such SVD of $\boldsymbol{G}$ among the 2^d equivalent ones. At a heuristic level, you may just think of

$$\widetilde{\boldsymbol{U}} := \begin{bmatrix} | & & | \\ \pm\boldsymbol{u}_1(\boldsymbol{G}) & \cdots & \pm\boldsymbol{u}_d(\boldsymbol{G}) \\ | & & | \end{bmatrix}, \quad \widetilde{\boldsymbol{V}} := \begin{bmatrix} | & & | \\ \pm\boldsymbol{v}_1(\boldsymbol{G}) & \cdots & \pm\boldsymbol{v}_d(\boldsymbol{G}) \\ | & & | \end{bmatrix},$$

with the signs chosen to make the above SVD hold.

We will now see how the following fact determines the laws of these matrices.

Proposition 2.4.15. The random matrices $\widetilde{\boldsymbol{U}}$ and $\widetilde{\boldsymbol{V}}$ are left-orthogonally invariant.

For $\widetilde{\boldsymbol{U}}$, first note that $\widetilde{\boldsymbol{U}} \in \mathcal{O}(d)$, i.e., this is an orthogonal matrix. The characterization of the law then follows from the next result.

Theorem 2.4.16 (Haar)**.** There exists a unique probability measure supported on $\mathcal{O}(d)$ that is left- or right-orthogonally invariant, called the *Haar measure* and which we denote $\mathsf{Haar}(\mathcal{O}(d))$. It is the same as the law of the matrix whose columns are the vectors formed by applying the Gram-Schmidt orthonormalization procedure to d i.i.d. random vectors sampled from either $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ or $\mathsf{Unif}(\mathbb{S}^{d-1})$.

Corollary 2.4.17. $\mathsf{Law}(\widetilde{\boldsymbol{U}}) = \mathsf{Haar}(\mathcal{O}(d))$.

For $\widetilde{\boldsymbol{V}}$ consisting of the right singular vectors, the above cannot quite work verbatim since this matrix is not square. Indeed, it belongs to a different special set of matrices, which is worth knowing about in its own right.

Definition 2.4.18 (Stiefel manifold)**.** The *Stiefel manifold* of dimensions $m \geq d$, denoted $\mathsf{Stief}(m, d)$, is the set

$$\mathsf{Stief}(m, d) := \{\boldsymbol{V} \in \mathbb{R}^{m\times d} : \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_d\},$$

or equivalently the set of $m \times d$ matrices with orthonormal columns.

This is a very natural geometric structure: it is the set to which the matrix of right singular vectors in the SVD belongs, but also it has an important relationship to parametrizing the set of d-dimensional subspaces of $\mathbb{R}^m$ (an object called the *Grassmannian manifold*), since it may be viewed as the set of all possible orthonormal bases of such subspaces.

One practical way to think of the Stiefel manifold is as the set of matrices that can be extended to orthogonal matrices in $\mathcal{O}(m)$ by adding $m - d$ more columns. Therefore, the following extension of Theorem 2.4.16 should not be surprising.

Corollary 2.4.19. For any $1 \le d \le m$, there exists a unique probability measure supported on $\mathsf{Stief}(m, d)$ that is left-orthogonally invariant, again called the Haar measure and which we denote $\mathsf{Haar}(\mathsf{Stief}(m, d))$. It is the same as the law of a matrix formed by sampling from $\mathsf{Haar}(\mathcal{O}(m))$ and keeping only the submatrix formed by the first d columns.

Remark 2.4.20. Note that $\mathsf{Stief}(m, 1) = \mathbb{S}^{m-1}$, and $\mathsf{Haar}(\mathsf{Stief}(m, 1)) = \mathsf{Unif}(\mathbb{S}^{m-1})$, the uniform measure we discussed in Chapter 1.

Corollary 2.4.21. $\mathsf{Law}(\widetilde{\boldsymbol{V}}) = \mathsf{Haar}(\mathsf{Stief}(m, d))$.

We have gotten as close as we might hope to defining "the laws of the singular vectors" of $\boldsymbol{G}$: though those singular vectors are not well-defined, we have described well-defined randomized versions $\widetilde{\boldsymbol{U}}$ and $\widetilde{\boldsymbol{V}}$ of the matrices formed by those vectors and identified their laws.

2.4.3 Connection to Random Projections

Previously, we raised the issue of the behavior of $\boldsymbol{G}^\top \boldsymbol{G} \in \mathbb{R}^{m \times m}_{\mathsf{sym}}$: this is a matrix of rank d, yet which "looks like" $d\boldsymbol{I}_m$ from the point of view of a small number of quadratic form evaluations. Using our findings about the singular values and vectors of $\boldsymbol{G}$, we may clarify this situation. Let $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{G}) \in \mathbb{R}^{d \times d}_{\mathsf{sym}}$ be the diagonal matrix of singular values. If $m \gg d$, then $\boldsymbol{\Sigma} \approx \sqrt{m}\boldsymbol{I}_d$.

We have $\boldsymbol{G} = \widetilde{\boldsymbol{U}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{V}}^\top$, so

$$\boldsymbol{G}^\top \boldsymbol{G} = \widetilde{\boldsymbol{V}}\boldsymbol{\Sigma}^2\widetilde{\boldsymbol{V}}^\top \approx m\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^\top.$$

Since $\widetilde{\boldsymbol{V}} \in \mathsf{Stief}(m, d)$, $\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^\top =: \boldsymbol{P}_V$ is a projection matrix to the d-dimensional column space $V := \mathsf{col}(\widetilde{\boldsymbol{V}})$. By the left-orthogonal invariance of $\widetilde{\boldsymbol{V}}$, this random d-dimensional subspace has law invariant under orthogonal transformations. You should think of this law as, in a similar spirit to the Haar measures above, the unique law of a "uniformly random" d-dimensional subspace of $\mathbb{R}^m$.

Now, consider a quadratic form: for $\boldsymbol{x} \in \mathbb{R}^m$,

$$\|\boldsymbol{G}\boldsymbol{x}\|^2 = \boldsymbol{x}^\top \boldsymbol{G}^\top \boldsymbol{G}\boldsymbol{x} \approx m\|\boldsymbol{P}_V \boldsymbol{x}\|^2.$$

We may view $\boldsymbol{P}_V \boldsymbol{x}$ as first expressing $\boldsymbol{x}$ in a random coordinate system, and then zeroing out all but d coordinates of $\boldsymbol{x}$ in this coordinate system. Since in a random coordinate system $\boldsymbol{x}$ should have entries of roughly the same size, we should expect

$$\|\boldsymbol{P}_V \boldsymbol{x}\|^2 \approx \frac{d}{m}\|\boldsymbol{x}\|^2,$$

and thus

$$\|\boldsymbol{G}\boldsymbol{x}\|^2 \approx d\|\boldsymbol{x}\|^2,$$

compatible with Corollary 1.2.7 applied to $\boldsymbol{G}\boldsymbol{x}$.

Thus the geometric mechanism behind the above observation about quadratic forms and the Johnson-Lindenstrauss transform can be seen as the principle that *suitably rescaled random projections preserve the norms of deterministic vectors*. As we have discussed, while for any given projection there will exist some vectors that "escape" this behavior and are either very close to or nearly orthogonal to the range of a random projection, one must fix a very large number of deterministic vectors in order to expect to find such a vector with respect to a random projection. (Indeed, you may reflect on how this is basically the same phenomenon as needing exponentially many vectors to produce an ϵ-net of $\mathbb{S}^{d-1}$.)

2.5 Application: Compressed Sensing

We will now see a few applications of the ideas developed so far. The first is to 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 $\|\boldsymbol{x}''\|_0 \leq 2k$, 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 its at most $2k$ non-zero indices into two sets of size at most k). 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 $\boldsymbol{G} \sim \mathcal{N}(0, 1/d)^{\otimes d \times m}$, 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

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

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 2.3.9. 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 2.3.9 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 2.3.9 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 (2.3.11) from that proof, which gives

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

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

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

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

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

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

2.6 Application: Neural Network Initialization

2.6.1 General Setup

We next give a quite different application to the training of *neural network* machine learning models. While this is a toy example, as we will mention below the result of our derivation is actually exactly what is used in some real model-training software.

A neural network forms a sequence of representations of its input by composing linear maps with nonlinear functions. In the simplest form without any constraints, this is a function $f : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ of the form

$$f(\boldsymbol{x}) = \rho(\boldsymbol{W}_L(\rho(\boldsymbol{W}_{L-1} \cdots \rho(\boldsymbol{W}_1 \boldsymbol{x}))))$$

for a sequence of rectangular matrices $\boldsymbol{W}_1 \in \mathbb{R}^{m_1 \times d_{\text{in}}}, \boldsymbol{W}_2 \in \mathbb{R}^{m_2 \times m_1}, \dots, \boldsymbol{W}_L \in \mathbb{R}^{d_{\text{out}} \times m_{L-1}}$ for some intermediate dimensions $m_1, \dots, m_{L-1}$. Here $\rho : \mathbb{R} \to \mathbb{R}$ is some nonlinear function, which is applied coordinatewise to vectors. A common choice that we focus on here is the *rectified linear unit (ReLU)*,

$$\rho(x) = \max\{0, x\}.$$

This kind of structure, with no constraints on the $\boldsymbol{W}_i$, is called a *multilayer perceptron (MLP)*. The function f is *parametrized* by the $\boldsymbol{W}_i$, which we may emphasize by writing $\boldsymbol{\theta} = (\boldsymbol{W}_1, \dots, \boldsymbol{W}_L)$ and writing $f(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta})$.

In *training* of such a network, we are given a dataset $(\boldsymbol{x}_1, \boldsymbol{y}_1), \dots, (\boldsymbol{x}_M, \boldsymbol{y}_M)$ of pairs of inputs and outputs, and wish to find $\boldsymbol{\theta}$ such that $f(\boldsymbol{x}_i; \boldsymbol{\theta}) \approx \boldsymbol{y}_i$. To do this, we define a *loss function*, a common choice being the ℓ^2 loss,

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{M} \|f(\boldsymbol{x}_i; \boldsymbol{\theta}) - \boldsymbol{y}_i\|^2.$$

We then attempt to minimize this over $\boldsymbol{\theta}$ by some form of local optimization such as gradient descent, stochastic gradient descent, and so forth. This yields a sequence of iterates $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \dots$, starting from some *initialization* $\boldsymbol{\theta}^{(0)}$ that we choose at the beginning.

There are many interesting questions about what happens during neural network training, why and when it works well, and how best to configure it. We will return to some others later, but for now we focus on the basic matter of how to choose the initialization $\boldsymbol{\theta}^{(0)}$.

2.6.2 Initialization and Local Analysis

Often the initial weight matrices are chosen to have i.i.d. Gaussian (or some other entry distribution) entries of various scales, $(\boldsymbol{W}_i)_{a,b} \sim \mathcal{N}(0, \sigma_i^2)$. This reduces our question to that of how to choose the σ_i^2. Let us consider a single nonlinear layer $f(\boldsymbol{x}) = \rho(\boldsymbol{W}\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{W} \in \mathbb{R}^{m \times d}$, and therefore $f(\boldsymbol{x}) \in \mathbb{R}^m$, but bearing in mind that we plan to compose several such layers. (Here the rows of $\boldsymbol{W}$ correspond to outputs, so the matrix dimensions are reversed from our earlier convention for rectangular matrices $\boldsymbol{G}$.) We then have just one parameter σ^2 to choose to set $\boldsymbol{W} \sim \mathcal{N}(0, \sigma^2)^{\otimes m \times d}$.

What is it that we want this choice of σ^2 to achieve? The following two criteria are reasonable:

1. First, we want the layer to roughly preserve norms, $\|f(\boldsymbol{x})\|^2 \approx \|\boldsymbol{x}\|^2$. If this fails, then composing many layers will either shrink of blow up the scale of the data, which may lead to issues with numerical stability at the beginning of gradient descent (among other issues).

2. Second, we want the layer to distinguish different points, mapping far apart $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ to far apart $f(\boldsymbol{x}_1)$ and $f(\boldsymbol{x}_2)$. If this fails, then we will "forget" some differences in the input data, and this will be impossible to undo by composing with further layers. If, say, we eventually want to solve a classification problem, then we risk "collapsing" two or more classes together in a way that they cannot be distinguished.

Let us focus on the *first* layer, where the second concern is most significant. In light of it, we should certainly choose $m \geq d$, or else we risk losing lots of information present in the data. (Choosing a very large m is also an instance of *overparametrization* of our model, which has many other effects and is an important aspect of most effective neural network architectures.)

Because $f(\boldsymbol{x})$ is nonlinear, these questions are tricky (though sometimes still possible) to address directly and globally. Instead, we consider a local version of the same question: we fix some $\boldsymbol{x} \in \mathbb{R}^d$ and consider small perturbations. Since f is differentiable almost everywhere, we have

$$f(\boldsymbol{x} + \boldsymbol{h}) \approx f(\boldsymbol{x}) + (\boldsymbol{J}f)(\boldsymbol{x})\boldsymbol{h},$$

where $(\boldsymbol{J}f)(\boldsymbol{x}) \in \mathbb{R}^{m\times d}$ is the *Jacobian matrix*,

$$(\boldsymbol{J}f)(\boldsymbol{x})_{ab} = \frac{\partial f_a}{\partial x_b}(\boldsymbol{x}).$$

A natural condition on the Jacobian that will achieve the above criteria is if it acts as an approximately isometric embedding, i.e., if its d singular values are all close to 1 with high probability for a given fixed $\boldsymbol{x}$.

2.6.3 Jacobian Calculations and Kaiming Initialization

Fix a deterministic input $\boldsymbol{x} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$. Write $\boldsymbol{w}_a^\top$ for row a of $\boldsymbol{W}$, and let

$$b_a := \mathbb{1}\{\langle \boldsymbol{w}_a, \boldsymbol{x}\rangle > 0\}.$$

Each $\langle \boldsymbol{w}_a, \boldsymbol{x}\rangle$ is a non-degenerate Gaussian random variable, so almost surely all of them are non-zero. On this event, $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}$, with Jacobian

$$\begin{aligned}(\boldsymbol{J}f)(\boldsymbol{x}) =: \boldsymbol{J}(\boldsymbol{x}) &= \boldsymbol{D}_{\boldsymbol{x}}\boldsymbol{W},\\ \boldsymbol{D}_{\boldsymbol{x}} &:= \mathsf{Diag}(b_1,\dots,b_m).\end{aligned}$$

Thus, the derivative keeps the rows corresponding to positive outputs and replaces the other rows by zero. We call those positive outputs *active*.

For all sufficiently small $\boldsymbol{h}$, none of these signs changes, and we actually have exactly

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{J}(\boldsymbol{x})\boldsymbol{h}.$$

The singular values therefore bound the change in the output by

$$\sigma_d(\boldsymbol{J}(\boldsymbol{x}))\|\boldsymbol{h}\| \le \|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x})\| \le \sigma_1(\boldsymbol{J}(\boldsymbol{x}))\|\boldsymbol{h}\|, \tag{2.6.1}$$

justifying the previous claims in case $\sigma_1(\boldsymbol{J}(\boldsymbol{x})), \sigma_d(\boldsymbol{J}(\boldsymbol{x})) \approx 1$.

To move towards analyzing this matrix, let $B := \sum_{a=1}^m b_a$ be the number of active outputs. The variables b_a are independent with law $\mathsf{Ber}(1/2)$, so $B \sim \mathsf{Bin}(m, 1/2)$. There is a dependence to handle carefully: a row of $\boldsymbol{W}$ is kept precisely when its inner product with $\boldsymbol{x}$ is positive. Conditional on being kept, that row is constrained to a half-space. However, the singular values depend on the row only through its outer product, which has a useful symmetry.

Proposition 2.6.1. Conditional on $B = k$, the matrix $\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x})$ has the same law as $\sigma^2 \boldsymbol{H}^\top \boldsymbol{H}$, where $\boldsymbol{H} \sim \mathcal{N}(0,1)^{\otimes k\times d}$. For $k = 0$, the latter matrix is understood to be zero.

Proof. Write $\boldsymbol{w}_a = \sigma \boldsymbol{g}_a$, so the $\boldsymbol{g}_a$ are independent standard Gaussian vectors in $\mathbb{R}^d$. For one such vector $\boldsymbol{g}$, the transformation $\boldsymbol{g} \mapsto -\boldsymbol{g}$ preserves its law and the outer product $\boldsymbol{g}\boldsymbol{g}^\top$, while reversing the sign of $\langle \boldsymbol{g}, \boldsymbol{x}\rangle$. Thus, for any measurable set $\mathcal{A}$ of symmetric matrices,

$$\begin{aligned}\mathbb{P}[\boldsymbol{g}\boldsymbol{g}^\top \in \mathcal{A},\, \langle \boldsymbol{g}, \boldsymbol{x}\rangle > 0] &= \mathbb{P}[\boldsymbol{g}\boldsymbol{g}^\top \in \mathcal{A},\, \langle \boldsymbol{g}, \boldsymbol{x}\rangle < 0]\\ &= \frac{1}{2}\mathbb{P}[\boldsymbol{g}\boldsymbol{g}^\top \in \mathcal{A}]\\ &= \mathbb{P}[\langle \boldsymbol{g}, \boldsymbol{x}\rangle > 0] \cdot \mathbb{P}[\boldsymbol{g}\boldsymbol{g}^\top \in \mathcal{A}].\end{aligned}$$

This proves that $\mathrm{sgn}(\langle \boldsymbol{g}, \boldsymbol{x}\rangle)$ is independent of $\boldsymbol{g}\boldsymbol{g}^\top$. Independence of the rows then shows that the collection $(\boldsymbol{g}_a\boldsymbol{g}_a^\top)_{a=1}^m$ is independent of the entire vector $(b_a)_{a=1}^m$. Since

$$\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x}) = \sigma^2 \sum_{a=1}^{m} b_a \boldsymbol{g}_a \boldsymbol{g}_a^\top,$$

conditioning on any set of k active rows gives a sum of k independent outer products of standard Gaussian vectors. This law depends only on k, which proves the claim. □

The Jacobian nonlinear layer has therefore led us back to a rectangular Gaussian matrix $\boldsymbol{H}$, but now with a random number of rows B. Since B is usually close to $m/2$, the bounds from earlier in this chapter suggest that all singular values of $\boldsymbol{J}(\boldsymbol{x})$ should be close to $\sigma\sqrt{m/2}$ in relative terms when m is much larger than d. We also obtain the exact identity

$$\mathbb{E}[\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x})] = \sigma^2 \mathbb{E}[B]\boldsymbol{I}_d = \frac{m\sigma^2}{2}\boldsymbol{I}_d. \tag{2.6.2}$$

Thus, the correct choice, most immediately the choice that preserves the squared length of every fixed perturbation in expectation, is

$$\boxed{\sigma^2 = \frac{2}{m}}$$

We will now show that sufficiently large m indeed makes this apply.

Theorem 2.6.2. There is a constant $C > 0$ such that the following holds. Let $\sigma^2 = 2/m$, $\epsilon, \delta \in (0,1)$, fix $\boldsymbol{x} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$, and suppose

$$m \geq \frac{C}{\epsilon^2}\left(d + \log\frac{1}{\delta}\right).$$

Then, with probability at least $1-\delta$,

$$\|\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x}) - \boldsymbol{I}_d\| \leq \epsilon. \tag{2.6.3}$$

Proof. By Hoeffding's lemma (Lemma 1.1.5), each $b_a - 1/2$ is $1/4$-subgaussian. Lemma 1.1.3 therefore gives

$$\mathbb{P}\left[\left|B - \frac{m}{2}\right| > \frac{\epsilon m}{4}\right] \leq 2\exp\left(-\frac{\epsilon^2 m}{8}\right).$$

On the complementary event, $m/4 \leq B \leq m$ and $|2B/m - 1| \leq \epsilon/2$. Taking C large enough ensures $m \geq 4d$, so $B \geq d$ on this event.

Since $\sigma^2 = 2/m$, Proposition 2.6.1 says that, conditional on $B = k$, the matrix $\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x})$ has the same law as $(2/m)\boldsymbol{H}^\top \boldsymbol{H}$, where $\boldsymbol{H}$ is a $k \times d$ standard Gaussian matrix. For every integer $k \in [m/4, m]$, applying (2.3.11) to $\boldsymbol{H}^\top$ with $t = \epsilon m/4$ gives

$$\mathbb{P}\left[\left\|\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x}) - \frac{2k}{m}\boldsymbol{I}_d\right\| > \frac{\epsilon}{2} \;\middle|\; B = k\right] \leq 2\exp\left(3d - \frac{\epsilon^2 m}{512}\right).$$

Here we used $t^2/(4k) \geq \epsilon^2 m/64$ and $t/2 \geq \epsilon^2 m/64$, since $k \leq m$ and $\epsilon < 1$. When both bounds hold, the triangle inequality gives

$$\begin{aligned}\|\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x}) - \boldsymbol{I}_d\| &\leq \left\|\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x}) - \frac{2B}{m}\boldsymbol{I}_d\right\| + \left|\frac{2B}{m} - 1\right| \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.\end{aligned}$$

The probability of failure is at most $2\exp(-\epsilon^2 m/8) + 2\exp(3d - \epsilon^2 m/512)$, which is at most δ under the assumed lower bound on m, for a sufficiently large constant C. □

The calculation gives two distinct roles to width m and variance σ^2. Large width makes the singular values close together, while choosing the "right" variance $\sigma^2 = 2/m$ places the singular values near 1. We note that the condition that the width be large is important: if, say, $m = d$, then the typical rank of $\boldsymbol{J}(\boldsymbol{x})$ is about $d/2$, and thus $\sigma_d(\boldsymbol{J}(\boldsymbol{x})) = 0$ usually. Working in an asymptotic regime of large width is a common assumption in the neural network literature and leads to many such limiting simplifications.

Theorem 2.6.2 concerns one fixed input at initialization, with simultaneous control over all perturbation directions at that input. A random input independent of the weights is also allowed, by conditioning on it, provided it is non-zero almost surely. Uniform control over inputs, or control after training changes the weights, requires further, considerably more complicated arguments. The broader question of making Jacobian singular values close to 1 is studied under the name *dynamical isometry*. See, for example, [PSG17].

The choice $\sigma^2 = 2/m$ is called the *fan-out* version of the *He* or *Kaiming initialization*,[1] as derived in [HZRS15, Section 2.2]. Precisely this normalization is used in actual implementations in various utilities in `PyTorch` and other neural network libraries.[2]

Remark 2.6.3 (Another criterion for the variance)**.** Symmetry of the Gaussian distribution also gives $\mathbb{E}[\|\boldsymbol{f}(\boldsymbol{x})\|^2] = (m\sigma^2/2)\|\boldsymbol{x}\|^2$. Thus our choice $\sigma^2 = 2/m$ preserves the total squared length of the input in expectation. One could instead ask to preserve the average squared coordinate, meaning $\mathbb{E}[\|\boldsymbol{f}(\boldsymbol{x})\|^2]/m = \|\boldsymbol{x}\|^2/d$. This criterion gives $\sigma^2 = 2/d$, the *fan-in* version of the Kaiming initialization. For a wide layer, the theorem then places the singular values near $\sqrt{m/d}$, and in particular the norm itself can be dramatically inflated over several such layers. The choice of variance therefore depends on which quantity we want to preserve.

[1] The last and first names of the first author of the reference.

[2] See, e.g., here and here.

Bibliography

[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.

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034. IEEE, 2015.

[PSG17] Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *arXiv preprint arXiv:1711.04735*, Nov 2017. Appeared at NeurIPS 2017.

[RH17] Philippe Rigollet and Jan-Christian Hütter. High dimensional statistics, 2017. Lecture notes.