Lecture Notes on Random Matrix Theory in Data Science and Statistics

Dmitriy (Tim) Kunisky

Fall 2026 (Last Updated: November 5, 2026)

Contents

1 Random Vector Theory **3**
- 1.1 Independent Entries and Subgaussian Concentration . . . 3
- 1.2 Gaussian Vectors and Subexponential Concentration . . . 6
- 1.3 Orthogonal Invariance . . . 9
- 1.4 Exercises . . . 11
  - 1.4.1 High-Dimensional Geometry . . . 11
  - 1.4.2 Concentration Inequalities . . . 13
  - 1.4.3 Orthogonal Invariance for Vectors . . . 14

2 Geometry of Rectangular Random Matrices **15**
- 2.1 Gaussian Random Field Viewpoint . . . 15
- 2.2 Application: Dimensionality Reduction . . . 18
  - 2.2.1 Johnson-Lindenstrauss Lemma . . . 18
  - 2.2.2 Simple Extensions . . . 19
  - 2.2.3 Lower Bounds . . . 19
  - 2.2.4 Sparse and Fast JLTs . . . 19
  - 2.2.5 Proof Technique: First Moment Method . . . 20
- 2.3 The **Geometric Method** of Random Matrix Theory . . . 20
- 2.4 Singular Values . . . 21
  - 2.4.1 Applying the Geometric Method . . . 22
  - 2.4.2 Nets, Coverings, and Packings . . . 23
  - 2.4.3 Coarse Non-Asymptotic Bound on Singular Values . . . 25
- 2.5 Singular Vectors . . . 28
  - 2.5.1 Distinctness of Singular Values . . . 29
  - 2.5.2 Orthogonal Invariance for Matrices and Haar Measures . . . 31
  - 2.5.3 Connection to Random Projections . . . 34
- 2.6 Application: Compressed Sensing . . . 35
  - 2.6.1 Null Space and Restricted Isometry Properties . . . 35
  - 2.6.2 Random Sensing Matrices . . . 36
- 2.7 Application: Neural Network Initialization . . . 38
  - 2.7.1 General Setup . . . 38
  - 2.7.2 Initialization and Local Analysis . . . 39
  - 2.7.3 Jacobian Calculations and Kaiming Initialization . . . 40
- 2.8 Exercises . . . 42

3 Classical Limit Theorems **45**
- 3.1 Main Phenomena and Motivation . . . 45
- 3.2 Convergence of Random Measures . . . 47
  - 3.2.1 Deterministic Weak Convergence . . . 48

3.2.2 Equivalent Conditions . . . 49
3.2.3 Random Weak Convergence . . . 50
3.3 The **Combinatorial Method** of Random Matrix Theory . . . 51
3.4 Warmup: Central Limit Theorem by Moments . . . 52
3.5 Wigner Semicircle Limit Theorem . . . 55
3.5.1 Convergence of Expected Moments . . . 56
3.5.2 Upgrading to Weak Convergence in Probability . . . 60
3.5.3 Universality . . . 63
3.6 Combinatorial Method for Extreme Eigenvalues . . . 65
3.7 Marchenko-Pastur Limit Theorem . . . 66
3.8 Exercises . . . 68

4 Free Probability **72**
4.1 Computing Joint Tracial Moments . . . 73
4.1.1 First Steps with Word Expansion . . . 74
4.1.2 Gaussian Orthogonal Ensemble Calculations . . . 77
4.1.3 Asymptotic Freeness . . . 82

Bibliography **83**

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 called σ^2*-subgaussian* if $\mathbb{E}\exp(\lambda Y) \le \exp(\frac{\sigma^2}{2}\lambda^2)$ 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 $Y_1, \dots, Y_d$ are independent and each is centered and σ^2-subgaussian. Then, for all $t \ge 0$,

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

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

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

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

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

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.

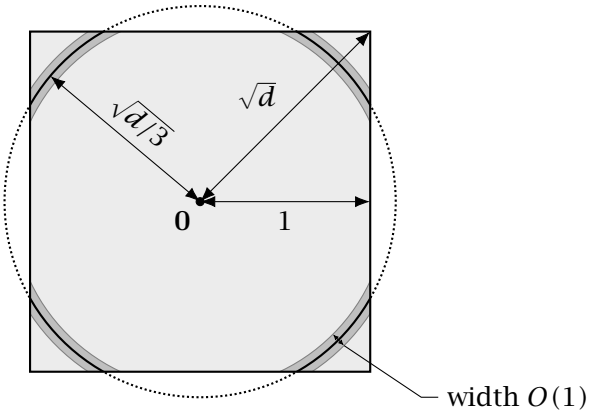


Figure 1.1: Schematic picture of the cube $[-1,1]^d$ and the thin spherical shell giving the typical set of its uniform measure, per Corollary 1.1.4. Distances are labeled for large dimension d and are not drawn to scale.

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.

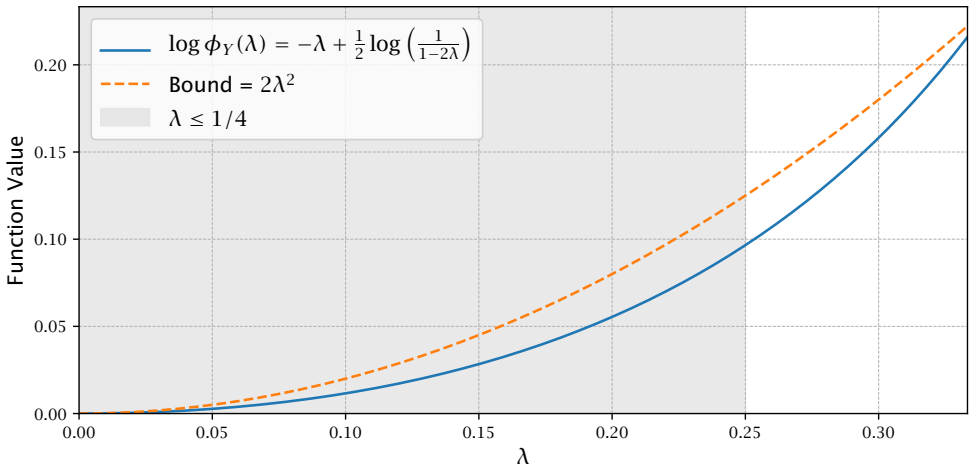


Figure 1.2: The cumulant generating function of the random variable Y appearing in Proposition 1.2.8, our bound on it, and the region where we apply this bound.

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.

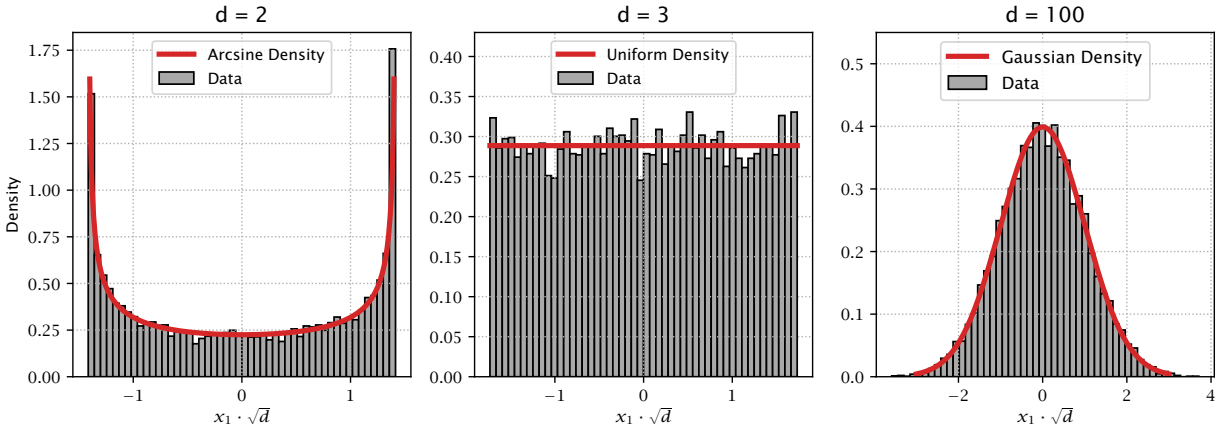


Figure 1.3: An illustration of Borel's theorem, Theorem 1.3.7, showing two low-dimensional examples $d = 2, 3$ where the distribution of $x_1 \cdot \sqrt{d}$ is known exactly (and is not Gaussian) as well as the Gaussian limit for large d.

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 $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$.

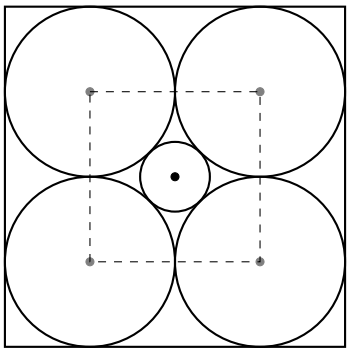
1.4 Exercises

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 (Almost orthogonal vectors and Gram matrices)**.**

1. Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ be independent uniform samples from $\{\pm 1/\sqrt{d}\}^d \subset \mathbb{S}^{d-1}$, where $d \geq 1$ and $N \geq 2$. For $\epsilon \in (0,1)$, show that

$$\mathbb{P}\left[\max_{i \neq j} |\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle| > \epsilon\right] \leq N^2 \exp(-d\epsilon^2/2).$$

   Deduce that there is a universal constant $c > 0$ such that, for every $\epsilon \in (0,1)$ and all sufficiently large d depending on ϵ, there exist $N \geq \exp(c\epsilon^2 d)$ unit vectors in $\mathbb{R}^d$ whose pairwise inner products have magnitude at most ϵ.

   (**HINT:** Apply subgaussian concentration to each inner product, then use the union bound.)

2. For any unit vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N \in \mathbb{R}^d$, define their *Gram matrix* $\boldsymbol{M} \in \mathbb{R}^{N \times N}_{\mathsf{sym}}$ by $M_{ij} := \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$. Show that

$$\lambda_1(\boldsymbol{M}) \geq \frac{N}{d}.$$

   Thus, when $N \gg d$, $\boldsymbol{M}$ is far from the identity matrix in operator norm, even though the construction in Part 1 makes every diagonal entry equal to 1 and every off-diagonal entry arbitrarily small in magnitude.

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 $g \sim \mathcal{N}(0, I_d)$. Show the non-asymptotic inequalities

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

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

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 $\mathbb{E}\exp(\lambda X) \le \exp(\sigma^2\lambda^2/2)$ for all $\lambda \in \mathbb{R}$. A centered random vector $x \in \mathbb{R}^d$ is called the same if, for all $u \in \mathbb{R}^d$ with $\|u\| = 1$, $\langle u, x\rangle$ 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 $X_1, \dots, X_n$ 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 $t > 0$, $\mathbb{P}[|X| \ge t] \le 2\exp(-ct^2)$.

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

3. For X a centered subgaussian scalar, define

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

   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 $(X+Y)^2 \le (1+a)X^2 + (1+\frac{1}{a})Y^2$ for any $a > 0$ (you might be more familiar with the case $a = 1$), together with Hölder's inequality.)

4. Show that there is a universal constant $C > 0$ such that, if X is a centered σ^2-subgaussian scalar, then

   $$\mathbb{E}[|X|^k] \le (C\sigma^2 k)^{k/2}$$

   for every integer $k \ge 1$.

   (**HINT:** Use the identity that, if $Y \ge 0$, then $\mathbb{E}[Y] = \int_0^\infty \mathbb{P}[Y > t]\,dt$. If you have not seen this before, sketch a proof of it as well.)

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

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

A centered random vector $x \in \mathbb{R}^d$ is called the same if, for every $u \in \mathbb{R}^d$ with $\|u\| = 1$, $\langle u, x\rangle$ 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 $X_1, \dots, X_n$ 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

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

   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 $t > 0$, $\mathbb{P}[|X| \ge t] \le 2\exp(-ct)$.

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

3. For X a centered subexponential scalar, define

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

   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,

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

   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 $\hat{\boldsymbol{x}} := \boldsymbol{x}/\|\boldsymbol{x}\|$ are independent random variables, i.e., that $\mathbb{P}[\|\boldsymbol{x}\| \in A, \hat{\boldsymbol{x}} \in B] = \mathbb{P}[\|\boldsymbol{x}\| \in A] \cdot \mathbb{P}[\hat{\boldsymbol{x}} \in B]$ for all pairs of Borel sets $A \subseteq \mathbb{R}_{\ge 0}$ 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 $\sigma^2 \ge 0$.

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.)

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

2.2 Application: Dimensionality Reduction

2.2.1 Johnson-Lindenstrauss Lemma

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

Definition 2.2.1. Given $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \in \mathbb{R}^m$, a function $f : \mathbb{R}^m \to \mathbb{R}^d$ is *pairwise ϵ-faithful* (this is not standard terminology) on the $\boldsymbol{y}_i$ if, for all $i, j \in [n]$,

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

Theorem 2.2.2 (Johnson-Lindenstrauss [JL82, IM98])**.** Let $n \ge 2$, $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \in \mathbb{R}^m$ be arbitrary, and $\epsilon \in (0,1)$. Suppose that

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

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

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

Proof. The proof actually has little to do with the $\boldsymbol{y}_i$. Notice that being pairwise ϵ-faithful is a matter of preserving $\binom{n}{2}$ many squared vector norms to within a factor in $[1-\epsilon, 1+\epsilon]$. We will show that this is true with high probability for *any* choice of $\binom{n}{2}$ vectors, and therefore also for the differences $\boldsymbol{y}_i - \boldsymbol{y}_j$, whose images satisfy $\widehat{G}(\boldsymbol{y}_i - \boldsymbol{y}_j) = \widehat{G}\boldsymbol{y}_i - \widehat{G}\boldsymbol{y}_j$.

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

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

and, using that $\epsilon < 1$ whereby $\epsilon d < d$, by Corollary 1.2.7 we have

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

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

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

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

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

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

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

2.2.2 Simple Extensions

It is easy to extend the proof to show that we also have $(1-\epsilon)\|\boldsymbol{y}_i\|^2 \le \|f(\boldsymbol{y}_i)\|^2 \le (1+\epsilon)\|\boldsymbol{y}_i\|^2$ with the same high probability up to constants (say, by increasing n by one and adding $\mathbf{0}$ to the $\boldsymbol{y}_i$, or by including further appropriate "bad events" in the union bound argument above). As mentioned before, it is also possible to get an error bound of $1/n^K$ for any $K > 0$ by suitably increasing the constant 24 in the statement of the Theorem and adjusting the last step accordingly.

2.2.3 Lower Bounds

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

2.2.4 Sparse and Fast JLTs

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

sparse Fourier transforms, so subsampling their Fourier transform runs a risk of ending up with an unusually small vector. Thus, a further preprocessing step is required to avoid this particular set of "bad" vectors; see the paper for more details.

2.2.5 Proof Technique: First Moment Method

As we have seen, the proof technique is merely an application of the union bound. However, it is perhaps worth emphasizing some aspects of how specifically we are using the union bound that also hold in many modern applications. While the union bound might seem trivial, and indeed is when applied to just a few events, the union bound when applied to many carefully chosen events is actually a quite sophisticated tool.[1] If one wishes to bound the probability of a "bad event" E, one may decompose it into $E = E_1 \cup \cdots \cup E_N$ and bound

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

Often, the $\mathbb{P}[E_i]$ are roughly the same. Then, the above becomes a clear "competition" between N, the number of ways that a bad event can happen, and $\mathbb{P}[E_i]$, the probability of a given way that it can happen.[2] In high-dimensional probability, often N is very large and $\mathbb{P}[E_i]$ is very small, leading to the kind of competition on an exponential scale that we saw in the proof.

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

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

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

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

We will now extend our approach to the JLT to understand the singular values of $\boldsymbol{G}$. 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

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

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

will be useful soon:

$$\mathbb{B}^d(\boldsymbol{x}, r) := \{\boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{y} - \boldsymbol{x}\| \le r\}, \tag{2.3.1}$$

$$\mathbb{B}^d := \mathbb{B}^d(\boldsymbol{0}, 1) \supset \mathbb{S}^{d-1}. \tag{2.3.2}$$

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

$$\|\boldsymbol{M}\| = \sup_{\boldsymbol{x} \in \mathbb{S}^{d-1}} |\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}| = \sup_{\boldsymbol{x} \in \mathbb{B}^d} |\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}|. \tag{2.3.3}$$

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

$$\|\boldsymbol{M}\|_{\mathcal{X}} := \sup_{\boldsymbol{x} \in \mathcal{X}} |\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}|. \tag{2.3.4}$$

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 2.3.2. 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,

$$\mathbb{P}[\|\boldsymbol{M}\|_{\mathcal{X}} \ge t] \le |\mathcal{X}| \cdot \max_{\boldsymbol{x} \in \mathcal{X}} \mathbb{P}[|\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}| \ge t]. \tag{2.3.5}$$

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.

2.4 Singular Values

We focus on "short fat" matrices, that is, the setting $d \le m$ (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

$$\boldsymbol{x}^\top(\boldsymbol{G}^\top\boldsymbol{G} - d\boldsymbol{I}_m)\boldsymbol{x} = \|\boldsymbol{G}\boldsymbol{x}\|^2 - d. \tag{2.4.1}$$

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 $\boldsymbol{G}$.

2.4.1 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

$$\begin{aligned} &\approx m \cdot \mathbb{E}[\boldsymbol{g}_1\boldsymbol{g}_1^\top] \\ &= m\boldsymbol{I}_d \\ &= \mathbb{E}[\boldsymbol{G}\boldsymbol{G}^\top]. \end{aligned} \tag{2.4.2}$$

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 2.4.1. 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 2.4.2 (Statistical interpretation)**.** Expanding in a sum of rank one terms again, we may also view

$$\frac{1}{m}\boldsymbol{\Delta} = \left(\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{g}_i \boldsymbol{g}_i^\top\right) - \mathbb{E}[\boldsymbol{g}_1 \boldsymbol{g}_1^\top] \tag{2.4.3}$$

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 $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. 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 2.3 with $\boldsymbol{M} = \boldsymbol{\Delta}$.

2.4.2 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 2.4.3 (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

$$\mathcal{Y} \subseteq \bigcup_{i=1}^{N} \mathbb{B}^d(\boldsymbol{x}_i, \epsilon), \tag{2.4.4}$$

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 2.4.4. 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,

$$\|\boldsymbol{M}\|_{\mathcal{X}} \le \|\boldsymbol{M}\| \le \frac{1}{1-2\epsilon}\|\boldsymbol{M}\|_{\mathcal{X}}. \tag{2.4.5}$$

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 $\mathbb{B}^d$. □

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:

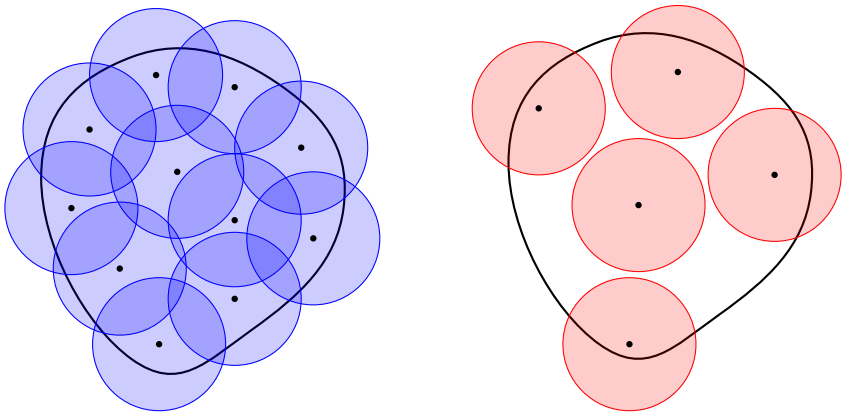


Figure 2.1: An illustration of a covering (left) and a packing (right) of a given set in $\mathbb{R}^2$ by balls of the same radius.

Definition 2.4.5 (ϵ-packing)**.** A set $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \mathcal{Y} \subset \mathbb{R}^d$ 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 2.4.6. 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 2.4.7. Let $\mathcal{X}$ be any ϵ-packing of $\mathbb{B}^d$. Then, $|\mathcal{X}| \leq (1 + \frac{1}{\epsilon})^d$.

Proof. We have

$$\bigcup_{\boldsymbol{x} \in \mathcal{X}} \mathbb{B}^d(\boldsymbol{x}, \epsilon) \subseteq \mathbb{B}^d(\boldsymbol{0}, 1 + \epsilon), \tag{2.4.6}$$

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

$$|\mathcal{X}| \cdot \mathsf{vol}(\mathbb{B}^d(\boldsymbol{0}, \epsilon)) \leq \mathsf{vol}(\mathbb{B}^d(\boldsymbol{0}, 1 + \epsilon)). \tag{2.4.7}$$

Rearranging,

$$|\mathcal{X}| \leq \frac{\mathsf{vol}(\mathbb{B}^d(\boldsymbol{0}, 1 + \epsilon))}{\mathsf{vol}(\mathbb{B}^d(\boldsymbol{0}, \epsilon))} = \left(\frac{1 + \epsilon}{\epsilon}\right)^d = \left(1 + \frac{1}{\epsilon}\right)^d, \tag{2.4.8}$$

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

Corollary 2.4.8. 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 2.4.7 guarantees that this process stops after finitely many steps. By Proposition 2.4.7, $|\mathcal{X}| \le (1 + \frac{2}{\epsilon})^d$, while by Proposition 2.4.6, $\mathcal{X}$ is an ϵ-net. □

2.4.3 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 2.4.9. If $d \le m$, then

$$\mathbb{P}[\|\boldsymbol{\Delta}\| > 64\sqrt{dm}] \le 2\exp(-d). \tag{2.4.9}$$

Proof. Let us fix $\epsilon := 1/4$. Let $\mathcal{X}$ be an ϵ-net of $\mathbb{B}^d$ as in Corollary 2.4.8, by which we can assume $|\mathcal{X}| \le (1 + \frac{2}{\epsilon})^d = 9^d$. By Lemma 2.4.4, 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

$$\begin{aligned}
\mathbb{P}[\|\boldsymbol{\Delta}\| \ge t] &\le \mathbb{P}\left[\|\boldsymbol{\Delta}\|_{\mathcal{X}} \ge \frac{t}{2}\right] && \text{(Lemma 2.4.4)} \\
&\le 9^d \cdot \max_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}\left[|\boldsymbol{x}^\top \boldsymbol{\Delta} \boldsymbol{x}| \ge \frac{t}{2}\right]. && \text{(Proposition 2.3.2)}
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}\left[|\boldsymbol{x}^\top \boldsymbol{\Delta} \boldsymbol{x}| \ge \frac{t}{2}\right] &= \mathbb{P}\left[\,\big|\,\|\boldsymbol{G}^\top \boldsymbol{x}\|^2 - m\|\boldsymbol{x}\|^2\,\big| \ge \frac{t}{2}\right] \\
&= \mathop{\mathbb{P}}_{\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_m)}\left[\|\boldsymbol{x}\|^2\,\big|\,\|\boldsymbol{g}\|^2 - m\,\big| \ge \frac{t}{2}\right] \\
&\le \mathop{\mathbb{P}}_{\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_m)}\left[\,\big|\,\|\boldsymbol{g}\|^2 - m\,\big| \ge \frac{t}{2}\right]. && (2.4.10)
\end{aligned}$$

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

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

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

$$\le 2\exp\left(3d - \frac{1}{8}\min\left\{\frac{t^2}{4m}, \frac{t}{2}\right\}\right) \tag{2.4.11}$$

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

$$\begin{aligned}
&\le 2\exp\left(3d - \frac{1}{8}\min\left\{\frac{C^2}{4}d, \frac{C}{2}d\right\}\right) \\
&= 2\exp\left(d \cdot \left[3 - \min\left\{\frac{C^2}{32}, \frac{C}{16}\right\}\right]\right) && (2.4.12)
\end{aligned}$$

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

$$\lambda_d(\boldsymbol{A}) - \|\boldsymbol{\Delta}\| \le \lambda_d(\boldsymbol{A}+\boldsymbol{\Delta}) \le \lambda_1(\boldsymbol{A}+\boldsymbol{\Delta}) \le \lambda_1(\boldsymbol{A}) + \|\boldsymbol{\Delta}\|. \tag{2.4.13}$$

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

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

Similarly, for the lower bound,

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

□

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

Corollary 2.4.11. 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

$$\sqrt{1+x} \le 1 + \frac{x}{2} \text{ for all } x \ge 0, \tag{2.4.14}$$

$$\sqrt{1-x} \ge 1 - x \text{ for all } 0 \le x \le 1. \tag{2.4.15}$$

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 2.4.9, using that $\boldsymbol{G}\boldsymbol{G}^\top = m\boldsymbol{I}_d + \boldsymbol{\Delta}$, we have

$$\begin{aligned}
\sigma_1(\boldsymbol{G}) &= \sqrt{\lambda_1(\boldsymbol{G}\boldsymbol{G}^\top)} \\
&\le \sqrt{m + \|\boldsymbol{\Delta}\|} && \text{(Proposition 2.4.10)} \\
&\le \sqrt{m + 64\sqrt{dm}} \\
&= \sqrt{m} \cdot \sqrt{1 + 64\sqrt{\frac{d}{m}}} \\
&\le \sqrt{m} \cdot \left(1 + 32\sqrt{\frac{d}{m}}\right) && (2.4.14) \\
&= \sqrt{m} + 32\sqrt{d}.
\end{aligned}$$

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,

$$\begin{aligned}
\sigma_d(\boldsymbol{G}) &= \sqrt{\lambda_d(\boldsymbol{G}\boldsymbol{G}^\top)} \\
&\ge \sqrt{m - \|\boldsymbol{\Delta}\|} && \text{(Proposition 2.4.10)} \\
&\ge \sqrt{m - 64\sqrt{dm}} \\
&= \sqrt{m} \cdot \sqrt{1 - 64\sqrt{\frac{d}{m}}} \\
&\ge \sqrt{m} \cdot \left(1 - 64\sqrt{\frac{d}{m}}\right) && (2.4.15) \\
&= \sqrt{m} - 64\sqrt{d},
\end{aligned}$$

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

$$\boxed{\sqrt{m} - O(\sqrt{d}) \le \sigma_d(\boldsymbol{G}) \le \cdots \le \sigma_1(\boldsymbol{G}) \le \sqrt{m} + O(\sqrt{d})} \tag{2.4.16}$$

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

2.5 Singular Vectors

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

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

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

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

$$\boldsymbol{L}_i := \boldsymbol{Z}_i \boldsymbol{Z}_i^\top = \boldsymbol{u}_i \boldsymbol{u}_i^\top, \tag{2.5.1}$$

$$\boldsymbol{R}_i := \boldsymbol{Z}_i^\top \boldsymbol{Z}_i = \boldsymbol{v}_i \boldsymbol{v}_i^\top. \tag{2.5.2}$$

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

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

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 $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d\}$. Thus, the more repeated singular values there are, the more dire this issue becomes. However, as the following fact of linear algebra shows, these are the only issues.

Proposition 2.5.1. Suppose that $\boldsymbol{G} \in \mathbb{R}^{d \times m}$ with $d \leq m$, $\sigma_d(\boldsymbol{G}) > 0$, and the singular values $\sigma_1(\boldsymbol{G}), \ldots, \sigma_d(\boldsymbol{G})$ 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

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

2.5.1 Distinctness of Singular Values

Remarkably, the conditions of the above Proposition 2.5.1 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 2.5.2. Let $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d \times m}$ with $d \leq m$. Then, almost surely, the following hold:

1. $\sigma_d(\boldsymbol{G}) > 0$.
2. The singular values $\sigma_1(\boldsymbol{G}), \ldots, \sigma_d(\boldsymbol{G})$ 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 2.5.3. Suppose $p \in \mathbb{R}[x_1, \ldots, x_N]$ 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 2.5.2 can be proved quite directly using this result, as follows.

Proof of Theorem 2.5.2, 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 2.5.3. □

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 2.5.4. 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 2.5.5 (Fundamental theorem of symmetric polynomials)**.** Define the *power sum* polynomials

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

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

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

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 $\boldsymbol{X}$,

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

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

Proposition 2.5.6. $s_k(\lambda_1(\boldsymbol{X}), \dots, \lambda_d(\boldsymbol{X})) = S_k(\boldsymbol{X})$.

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 2.5.7. Suppose that $F(\boldsymbol{X}) = f(\lambda_1(\boldsymbol{X}), \dots, \lambda_d(\boldsymbol{X}))$ 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 2.5.8. 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 2.5.2, Claim 2. Consider first a function of $\boldsymbol{X} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$, defined as:

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

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 2.5.7. 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 2.5.3 that

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

as claimed. □

Remark 2.5.9. 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.

2.5.2 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 2.5.10 (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 2.5.11. $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d \times m}$ is bi-orthogonally invariant.

Corollary 2.5.12. 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 2.5.13 (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 2.5.12. Let us prove the simpler claim about the individual $\boldsymbol{L}_i$: we have, by combining the above results,

$$\begin{aligned}
\mathsf{Law}(\boldsymbol{Q}\boldsymbol{L}_i(\boldsymbol{G})\boldsymbol{Q}^\top) &= \mathsf{Law}(\boldsymbol{L}_i(\boldsymbol{Q}\boldsymbol{G})) && \text{(Proposition 2.5.13)}\\
&= \mathsf{Law}(\boldsymbol{L}_i(\boldsymbol{G})). && \text{(Proposition 2.5.11)}
\end{aligned}$$

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 2.5.14. 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 2.5.15. 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 2.5.16 (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 2.5.17. $\mathsf{Law}(\widetilde{\boldsymbol{U}}) = \mathsf{Haar}(\mathcal{O}(d))$.

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 2.5.18 (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 2.5.16 should not be surprising.

Corollary 2.5.19. 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 2.5.20. 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 2.5.21. $\mathsf{Law}(\widetilde{\boldsymbol{V}}) = \mathsf{Haar}(\mathsf{Stief}(m, d))$.

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.

2.5.3 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 $\mathbb{S}^{d-1}$.)

2.6 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.)

2.6.1 Null Space and Restricted Isometry Properties

Let us denote sparsity by

$$\|\boldsymbol{x}\|_0 := \#\{i \in [m] : x_i \neq 0\}. \tag{2.6.1}$$

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 2.6.1. 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 2.6.2 (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 2.6.3. $\boldsymbol{G}$ 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 2.6.4 (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 $\|\boldsymbol{x}\|_0 \leq k$,

$$(1 - \delta)\|\boldsymbol{x}\|^2 \leq \|\boldsymbol{G}\boldsymbol{x}\|^2 \leq (1 + \delta)\|\boldsymbol{x}\|^2. \tag{2.6.2}$$

The following implication is immediate.

Proposition 2.6.5. 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.

2.6.2 Random Sensing Matrices

The following is the main result that we will show.

Theorem 2.6.6. 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

$$\mathbb{P}[\boldsymbol{G} \text{ has the } k\text{-NSP}] \ge \mathbb{P}[\boldsymbol{G} \text{ has the } (k,\delta)\text{-RIP}] \ge 1 - \frac{2}{\binom{m}{k}} \ge 1 - 2\left(\frac{k}{m}\right)^k. \tag{2.6.3}$$

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

$$\boxed{d \gtrsim k \log\left(\frac{m}{k}\right)} \tag{2.6.4}$$

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

$$\|\boldsymbol{x}\|^2 = \|\boldsymbol{x}^{(S)}\|^2, \tag{2.6.5}$$

$$\boldsymbol{G}\boldsymbol{x} = \boldsymbol{G}^{(S)}\boldsymbol{x}^{(S)}. \tag{2.6.6}$$

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

$$(1-\delta)\|\boldsymbol{x}^{(S)}\|^2 \le \|\boldsymbol{G}^{(S)}\boldsymbol{x}^{(S)}\|^2 \le (1+\delta)\|\boldsymbol{x}^{(S)}\|^2. \tag{2.6.7}$$

But this just amounts to asking that

$$1 - \delta \le \lambda_k(\boldsymbol{G}^{(S)\top}\boldsymbol{G}^{(S)}) \le \lambda_1(\boldsymbol{G}^{(S)\top}\boldsymbol{G}^{(S)}) \le 1 + \delta, \tag{2.6.8}$$

or again equivalently that

$$\|\boldsymbol{G}^{(S)\top}\boldsymbol{G}^{(S)} - \boldsymbol{I}_k\| \le \delta. \tag{2.6.9}$$

We will then proceed by union bounding,

$$\mathbb{P}[\boldsymbol{G} \text{ does not have the } (k,\delta)\text{-RIP}] \le \sum_{S \in \binom{[m]}{k}} \mathbb{P}[\|\boldsymbol{G}^{(S)\top}\boldsymbol{G}^{(S)} - \boldsymbol{I}_k\| > \delta]. \tag{2.6.10}$$

This is almost exactly the kind of deviation that we bounded before in Theorem 2.4.9. 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 2.4.9 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 2.4.9 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 (2.4.11) from that proof, which gives

$$\begin{aligned} &\leq 2\exp\left(3k - \frac{1}{8}\min\left\{\frac{\delta^2 d^2}{4d}, \frac{\delta d}{2}\right\}\right) \\ &= 2\exp\left(3k - \frac{\delta^2}{32}d\right), \end{aligned} \tag{2.6.11}$$

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

$$\begin{aligned} \mathbb{P}[\boldsymbol{G} \text{ does not have the } (k,\delta)\text{-RIP}] &\leq \binom{m}{k} \cdot 2\exp\left(3k - \frac{\delta^2}{32}d\right) \\ &\leq \left(\frac{em}{k}\right)^k \cdot 2\exp\left(3k - \frac{\delta^2}{32}d\right) \\ &\leq 2\exp\left(k\left(1 + \log\left(\frac{m}{k}\right)\right) + 3k - \frac{\delta^2}{32}d\right) \\ &\leq 2\exp\left(4k\left(1 + \log\left(\frac{m}{k}\right)\right) - \frac{\delta^2}{32}d\right) \\ &\leq 2\exp\left(-k\left(1 + \log\left(\frac{m}{k}\right)\right)\right) \\ &\leq 2\left(\frac{em}{k}\right)^{-k} \\ &\leq \frac{2}{\binom{m}{k}} \end{aligned} \tag{2.6.12}$$

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

2.7 Application: Neural Network Initialization

2.7.1 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 $\boldsymbol{\theta}^{(0)}$.

2.7.2 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 $\boldsymbol{x}$.

2.7.3 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

$$\sigma_d(\boldsymbol{J}(\boldsymbol{x}))\|\boldsymbol{h}\| \le \|\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x})\| \le \sigma_1(\boldsymbol{J}(\boldsymbol{x}))\|\boldsymbol{h}\|, \tag{2.7.1}$$

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 2.7.1. 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

$$\mathbb{E}[\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x})] = \sigma^2 \mathbb{E}[B] \boldsymbol{I}_d = \frac{m\sigma^2}{2} \boldsymbol{I}_d. \tag{2.7.2}$$

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 2.7.2. 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 $1-\delta$,

$$\|\boldsymbol{J}(\boldsymbol{x})^\top \boldsymbol{J}(\boldsymbol{x}) - \boldsymbol{I}_d\| \leq \epsilon. \tag{2.7.3}$$

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 2.7.1 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 (2.4.11) 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 2.7.2 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*,[3] 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.[4]

Remark 2.7.3 (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.

2.8 Exercises

Exercise 2.8.1 (Packing almost orthogonal vectors)**.** Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_N \in \mathbb{S}^{d-1}$ satisfy $|\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle| \le \epsilon$ for all $i \ne j$, where $\epsilon \in (0,1)$. Show that there is a constant $C = C(\epsilon) > 0$ such that $N \le \exp(Cd)$. Together with Exercise 1.4.2, this shows that, for fixed ϵ, the largest possible number of almost orthogonal vectors grows exponentially in the dimension.

(**Hint:** First bound the distance between each pair of vectors from below. Then compare the volumes of disjoint balls centered at the vectors with that of a ball containing their union.)

Exercise 2.8.2 (Subgaussian vector and matrix norms)**.** Recall the definition of a subgaussian random vector from Exercise 1.4.5.

1. Suppose that $\boldsymbol{x} \in \mathbb{R}^d$ is a centered σ^2-subgaussian random vector. Show that there are universal constants $c, C > 0$ such that
$$\mathbb{P}[\|\boldsymbol{x}\| > C\sigma\sqrt{d}] \le \exp(-cd).$$
Thus, a high-dimensional $O(1)$-subgaussian vector has norm $O(\sqrt{d})$ with high probability.

(**Hint:** Use ϵ-nets and adapt Lemma 2.4.4 to the Euclidean norm of a vector.)

[3] The last and first names of the first author of the reference.
[4] See, e.g., here and here.

2. Let $\boldsymbol{X} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ be a centered random symmetric matrix. Write $\mathsf{vec}(\boldsymbol{X}) \in \mathbb{R}^{d(d+1)/2}$ for the vector formed by concatenating the entries on and above the diagonal of $\boldsymbol{X}$, in any fixed order. Show that there are universal constants $c, C > 0$ such that, if $\mathsf{vec}(\boldsymbol{X})$ is a σ^2-subgaussian random vector, then
$$\mathbb{P}[\|\boldsymbol{X}\| > C\sigma\sqrt{d}] \le \exp(-cd).$$
In particular, a random symmetric matrix with independent centered $O(1)$-subgaussian entries on and above the diagonal has operator norm $O(\sqrt{d})$ with high probability.

(**Hint:** Use ϵ-nets again and express $\boldsymbol{v}^\top \boldsymbol{X} \boldsymbol{v}$ as $\langle \mathsf{vec}(\boldsymbol{X}), \boldsymbol{u}\rangle$ for a suitable $\boldsymbol{u} \in \mathbb{R}^{d(d+1)/2}$. Bound $\|\boldsymbol{u}\|$ uniformly over unit vectors $\boldsymbol{v}$.)

Exercise 2.8.3 (Johnson-Lindenstrauss lower bound)**.** For $\epsilon \in (0,1)$, Theorem 2.2.2 gives embeddings of any $n \ge 2$ points in Euclidean space into $O(\epsilon^{-2}\log n)$ dimensions that preserve all squared distances up to factors of $1 \pm \epsilon$. This exercise proves a lower bound on the required dimension that nearly matches this guarantee.

1. Let $\lambda_1, \dots, \lambda_n \ge 0$, not all zero. Show that
$$\#\{i : \lambda_i \ne 0\} \ge \frac{(\sum_{i=1}^n \lambda_i)^2}{\sum_{i=1}^n \lambda_i^2}.$$
Reinterpret this as a relationship between the rank, trace, and Frobenius norm of a non-zero positive semidefinite matrix.

2. Suppose $\boldsymbol{M} \in \mathbb{R}^{n\times n}_{\mathsf{sym}}$ has $\boldsymbol{M} \succeq \boldsymbol{0}$, $M_{ii} = 1$ for all $i \in [n]$, and $|M_{ij}| \le 1/\sqrt{n}$ for all $i \ne j$. Show that $\mathsf{rank}(\boldsymbol{M}) \ge n/2$.

3. For integers $r, k \ge 1$, show that the number of monomials $v_1^{a_1} \cdots v_r^{a_r}$ with non-negative integer exponents and total degree $a_1 + \cdots + a_r = k$ is $\binom{r+k-1}{k}$. Show also that
$$\binom{r+k-1}{k} \le \left(e \cdot \frac{r+k-1}{k}\right)^k.$$

(**Hint:** For the bound, use the series expansion of $\exp(k)$ to show that $k! \ge (k/e)^k$.)

4. For an integer $k \ge 1$ and a non-zero positive semidefinite matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}_{\mathsf{sym}}$, write $\boldsymbol{X}^{\odot k}$ for its entrywise kth power, with $(\boldsymbol{X}^{\odot k})_{ij} = X_{ij}^k$. Show that
$$\boldsymbol{X}^{\odot k} \succeq \boldsymbol{0},$$
$$\mathsf{rank}(\boldsymbol{X}^{\odot k}) \le \binom{\mathsf{rank}(\boldsymbol{X}) + k - 1}{k}.$$

(**Hint:** Write $X_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j\rangle$ for vectors $\boldsymbol{v}_i \in \mathbb{R}^{\mathsf{rank}(\boldsymbol{X})}$. Express X_{ij}^k as another Gram matrix using monomials of degree k.)

5. Let $\boldsymbol{y}_0 := \boldsymbol{0}$ and $\boldsymbol{y}_i := \boldsymbol{e}_i$ for $i = 1, \dots, n$, a point cloud of $n+1$ vectors in $\mathbb{R}^n$. Show that there are constants $c, \epsilon_0 > 0$ such that, for every $\epsilon \in (0, \epsilon_0)$ and all sufficiently large n depending on ϵ, if
$$d \le \frac{c \log n}{\epsilon^2 \log(1/\epsilon)},$$

then there do not exist $z_0, z_1, \ldots, z_n \in \mathbb{R}^d$ satisfying

$$(1-\epsilon)\|y_i - y_j\|^2 \leq \|z_i - z_j\|^2 \leq (1+\epsilon)\|y_i - y_j\|^2 \text{ for all } i, j \in \{0, \ldots, n\}.$$

Thus, the dimension in the Johnson-Lindenstrauss lemma is necessary up to a factor of $O(\log(1/\epsilon))$ for these point clouds.

(**HINT:** Form the Gram matrix X with entries

$$X_{ij} := \frac{\langle z_i - z_0, z_j - z_0 \rangle}{\|z_i - z_0\| \cdot \|z_j - z_0\|}$$

for $i, j \in \{1, \ldots, n\}$. Bound its off-diagonal entries, then raise it to a large enough entrywise power to obtain a contradiction between Parts 2 and 4, using the bound from Part 3.)

3 | Classical Limit Theorems

We now proceed, chronologically backwards, to the older asymptotic results that initiated random matrix theory, motivated then sometimes by physics and sometimes by statistics. On the one hand, this will involve *asymptotic* statements about convergences as dimensions tend to infinity, without concrete bounds on probabilities of fluctuations like in, e.g., our previous centerpiece Corollary 2.4.11 above. On the other hand, the statements to come will concern finer-grained information than our previous inequalities, much in the same way that the central limit theorem (CLT) is more precise or refined than the law of large numbers. We first explain what new information these statements will give us and why and when it might be useful.

3.1 Main Phenomena and Motivation

Let us switch to a scaling that is more conventional to classical random matrix theory and that will make some of the phenomena we are interested in clearer. Consider $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d \times m}$, with columns $\boldsymbol{g}_1, \dots, \boldsymbol{g}_m \in \mathbb{R}^d$, and let

$$\boldsymbol{M} := \frac{1}{m}\boldsymbol{G}\boldsymbol{G}^\top = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{g}_i \boldsymbol{g}_i^\top. \tag{3.1.1}$$

Let $\lambda_1 \geq \cdots \geq \lambda_d$ be the eigenvalues of $\boldsymbol{M}$. As we have seen, we expect $\boldsymbol{M} \approx \boldsymbol{I}_d$ by an informal law of large numbers argument, and Corollary 2.4.11 makes this precise, showing that if $m \gg d$, then with high probability we have $\lambda_1, \dots, \lambda_d \in [1 - O(\sqrt{d/m}), 1 + O(\sqrt{d/m})]$.

It is instructive to consider the case $d = 1$, in which case $\boldsymbol{M} = M$ is a scalar, $M = \frac{1}{m}\sum_{i=1}^{m} g_i^2$ for $g_i \sim \mathcal{N}(0,1)$. In this case, Corollary 2.4.11 says that, with high probability, $M \in [1 - O(\sqrt{1/m}), 1 + O(\sqrt{1/m})]$. But we know from classical probability that we can make much more precise statements about the *shape* of the fluctuations around 1: not only are they of order $\sqrt{1/m}$, but their distribution is roughly $\mathcal{N}(0, 2/m)$ by the CLT. Let us see two kinds of analogous matrix-valued results that we will show soon.

Wigner Semicircle Limit Theorem We first ask, how does the *distribution* of the d eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ look near 1 when $m \gg d$? Perhaps you are tempted to predict that it should look Gaussian, by analogy with the CLT. But this is wrong: Figure 3.1 shows that the typical shape is well-approximated not by the Gaussian bell curve, but by a *semicircle* centered at 1, of radius scaling as $\sim 2\sqrt{d/m}$. Note in particular the key qualitative difference between the semicircle distribution and the Gaussian: the former is *compactly supported*, suggesting that the eigenvalues actually concentrate significantly more strongly than a hypothetical Gaussian limiting distribution would predict.

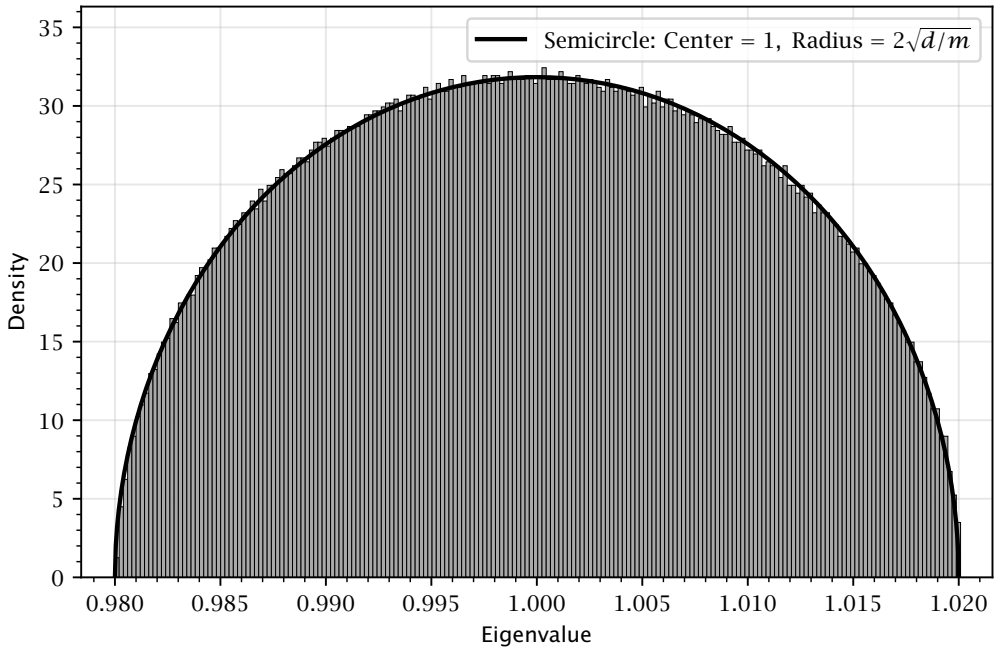


Figure 3.1: A histogram of the eigenvalues of $\boldsymbol{M}$ with $d = 20\,000$ and $m = 200\,000\,000$, using 200 bins, along with the semicircle prediction proposed in the main text.

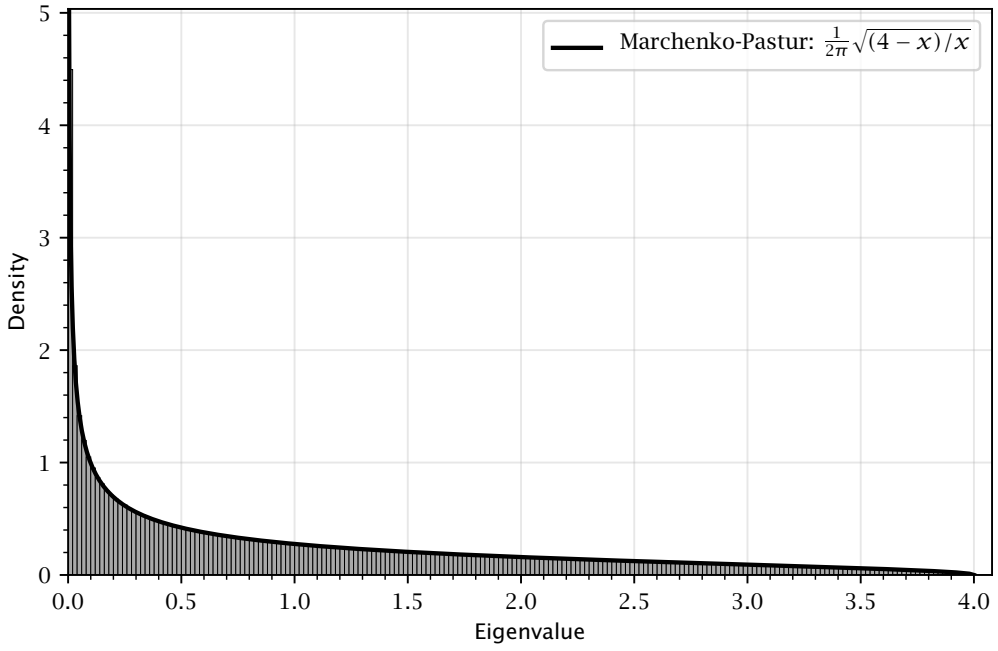


Figure 3.2: A histogram of the eigenvalues of $\boldsymbol{M}$ with $d = m = 20\,000$, using 200 bins, along with the Marchenko-Pastur prediction proposed in the main text.

MARCHENKO-PASTUR LIMIT THEOREM It is also natural to ask the same when $d/m \to c$ for some $c \in (0,\infty)$. In this case, Corollary 2.4.11 is even less informative, only telling us that $\lambda_1,\dots,\lambda_d \in [0, O(1)]$ with high probability, i.e., that the eigenvalues are bounded (*a priori* the lower bound is $-O(1)$, but we also know that $\lambda_i \geq 0$ since $\boldsymbol{M} \succeq \mathbf{0}$). But what is the shape of their distribution and what are their actual typical extreme values? We will see that there is a family of densities depending on c answering this question, called the *Marchenko-Pastur laws* $\mu_{\mathsf{MP}(c)}$. Like the semicircle law, these are totally alien to classical probability, but play a central role in random matrix theory. For instance, as we will prove and as Figure 3.2 illustrates, when $m = d$, then the density of eigenvalues converges as $d \to \infty$ to

$$\frac{d\mu_{\mathsf{MP}(1)}}{dx} = \frac{1}{2\pi}\sqrt{\frac{4-x}{x}}, \tag{3.1.2}$$

supported on $x \in [0,4]$. (As we will see and as you should expect, as $c \to 0$ these densities will approach a semicircle shape, but as this example shows, for larger c they can look qualitatively extremely different.) What is this strange distribution and where does it come from?

Remark 3.1.1 (A word on convergence of random measures)**.** You would be right to wonder what it means to say that a histogram of the $\lambda_1,\dots,\lambda_d$ "looks like" some deterministic curve. The histogram must be renormalized appropriately depending on d, but also and more importantly, the histogram itself (i.e., the collection of bin heights) is random. That is, there are two layers of probability here: the λ_i themselves are random, according to some (as yet mysterious) probability measure, but we are then drawing a picture of their distribution, which is therefore a *random probability measure*. What we are saying, and will say more precisely below (and already you may convince yourself that this holds with numerical experiments), is that (1) the expected height of each bin lies close to the specified curve, and (2) the random height of each bin concentrates around its expectation.

One motivation for studying these questions comes from statistics. Suppose we are trying to perform *covariance estimation*: there is an unknown covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, and we are given $\boldsymbol{g}_1,\dots,\boldsymbol{g}_m \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$. We try to estimate $\boldsymbol{\Sigma}$ by the sample covariance, $\hat{\boldsymbol{\Sigma}} := \frac{1}{m}\sum_{i=1}^m \boldsymbol{g}_i\boldsymbol{g}_i^\top$, which is $\boldsymbol{M}$ above when $\boldsymbol{\Sigma} = \boldsymbol{I}_d$.

In this context, the second limiting behavior alluded to above says something remarkable: even when $\boldsymbol{\Sigma} = \boldsymbol{I}_d$ and we are observing pure white noise, if we do not have enough data points—the number of observations m is only proportional to, not much larger than, the number of features d of each observation—then this direct estimate of $\boldsymbol{\Sigma}$ will be *inconsistent*. In particular, we will be led to believe that there are *spurious correlations* in our dataset: some eigenvalues of $\hat{\boldsymbol{\Sigma}}$ that are larger than others, and therefore particular directions in $\mathbb{R}^d$ in which the $\boldsymbol{g}_i$ are systematically biased. Understanding how $\hat{\boldsymbol{\Sigma}}$ behaves is the first step towards designing revised estimators that avoid these issues, a technique called *eigenvalue shrinkage*. (It should make sense that the right thing to do is to somehow "squeeze" the eigenvalues we think $\hat{\boldsymbol{\Sigma}}$ has closer together, hence this name. If you are familiar with *Stein's paradox* in classical statistics, you might revisit that and observe that this strategy is similar in spirit.)

3.2 CONVERGENCE OF RANDOM MEASURES

Let us first make more precise what it means for random "histograms" or, more formally, *random measures*, to converge to a deterministic measure in a way that captures the numerical experiments presented above. We will present this in a somewhat sophisticated way, but will try to refer back to our concrete experiments and observations about histograms to keep the discussion grounded.

Definition 3.2.1. We write $\mathcal{P}(\mathbb{R})$ for the set of probability measures on $\mathbb{R}$, with respect to the Borel σ-algebra.

3.2.1 Deterministic Weak Convergence

We first propose a more sophisticated way to view the histograms of a finite collection of numbers $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_d) \in \mathbb{R}^d$ (not necessarily random for now). We associate to these their *empirical measure* or *empirical distribution*

$$\mathsf{ed}(\boldsymbol{\lambda}) := \frac{1}{d} \sum_{i=1}^{d} \delta_{\lambda_i} \in \mathcal{P}(\mathbb{R}). \tag{3.2.1}$$

If you are not familiar with measure-theoretic language, just think of this as a function that takes (measurable) sets and outputs the fractions of λ_i that fall in that set:

$$\mathsf{ed}(\boldsymbol{\lambda})(A) = \frac{\#\{i \in [d] : \lambda_i \in A\}}{d}. \tag{3.2.2}$$

In particular, for intervals $A = [a, b]$, this counts the fraction of λ_i falling into that interval. Therefore, $\mathsf{ed}(\boldsymbol{\lambda})$ contains all the information needed to draw a histogram of the λ_i with *any* bin width, making it a more convenient object to work with.

Next, still thinking of deterministic $\boldsymbol{\lambda}$, what does it mean for a sequence of such empirical distributions to converge? We propose a definition for general measures, and then say what this means for empirical distributions.

Definition 3.2.2 (Weak convergence)**.** A sequence $\mu_d \in \mathcal{P}(\mathbb{R})$ *converges weakly* to $\mu \in \mathcal{P}(\mathbb{R})$, denoted $\mu_d \xrightarrow{\text{(w)}} \mu$, if, for all bounded continuous functions $f : \mathbb{R} \to \mathbb{R}$, we have $\int f \, d\mu_d \to \int f \, d\mu$ as $d \to \infty$.

As we will see in a moment, under a mild extra assumption, we are also allowed to consider certain discontinuous f, in particular $f(x) = \mathbb{1}\{x \in [a, b]\}$. Provided we can do this, in the case of $\mu_d = \mathsf{ed}(\boldsymbol{\lambda}^{(d)})$, we learn that, if μ has a continuous density ρ, then

$$\frac{\#\{i \in [d] : \lambda_i^{(d)} \in [a, b]\}}{d} \to \int_a^b \rho(t) \, dt. \tag{3.2.3}$$

Let us check that (3.2.3) describes the pictorial convergence of histograms. In a histogram, we plot the fraction of $\lambda_i^{(d)}$ that fall in bins of some width $\delta > 0$, which is the above situation for the interval $[a, a + \delta]$ for some left bin edge $a \in \mathbb{R}$. Thus if $\mathsf{ed}(\boldsymbol{\lambda}^{(d)}) \xrightarrow{\text{(w)}} \mu$ which has a continuous density ρ, and if the above consequence indeed holds, then for any choice of δ we will have, as $d \to \infty$,

$$\text{height of bin with left edge } a \to \int_a^{a+\delta} \rho(t) \, dt. \tag{3.2.4}$$

In our experiments, though, we observe not just this convergence of bin heights. We also see that, as $\delta \to 0$, the *shape* of the limiting bin heights converges, too. To see this, the bin heights must be renormalized—clearly, if we just take $\delta \to 0$ above, all heights will converge to zero. However, we have

$$\lim_{\delta \to 0} \frac{1}{\delta} \lim_{d \to \infty} (\text{height of bin with left edge } a) = \lim_{\delta \to 0} \frac{1}{\delta} \int_a^{a+\delta} \rho(t) \, dt = \rho(a) \tag{3.2.5}$$

by the Lebesgue differentiation theorem (which just says that, for ρ continuous, averages of its value over shrinking intervals $[a, a+\delta]$ converge to its point value at the point to which the intervals shrink). Thus we see that, upon renormalizing like this, if empirical distributions converge weakly, then histograms of shrinking bin widths will indeed converge to the curve that is the density of the limiting distribution μ.

3.2.2 Equivalent Conditions

As we have alluded to above, there are actually several equivalent definitions of weak convergence, and you might have seen other ones emphasized in your studies. Also, confusingly, there are some conditions that are equivalent for characterizing when a sequence of probability measures converges to another probability measure, but that become distinct when we are concerned with, say, a sequence of probability measures converging to another measure (for instance, describing in what sense the point masses δ_n converge to the zero measure as $n \to \infty$). The following describes many equivalent conditions in the simpler case of probability measures converging weakly to another probability measure, along with some notes.

Definition 3.2.3. Let $\mu \in \mathcal{P}(\mathbb{R})$. Its *cumulative distribution function (cdf)* is $\mathsf{cdf}(\mu) : \mathbb{R} \to \mathbb{R}$ with $\mathsf{cdf}(\mu)(t) := \mathbb{P}_{X\sim\mu}[X \leq t] = \mu((-\infty, t])$.

Theorem 3.2.4 (Extended portmanteau theorem)**.** Let $\mu_d, \mu \in \mathcal{P}(\mathbb{R})$. (We emphasize that we assume these are probability measures on $\mathbb{R}$, without which assumption some of the equivalences below are false.) Then, the following are equivalent:

1. For every continuous, bounded function $f : \mathbb{R} \to \mathbb{R}$, $\int f\, d\mu_d \to \int f\, d\mu$. This is the standard definition of weak convergence in more general settings.
2. For every continuous, compactly supported function $f : \mathbb{R} \to \mathbb{R}$, $\int f\, d\mu_d \to \int f\, d\mu$. This is usually known as *vague convergence* and sometimes denoted $\mu_d \xrightarrow{(\mathrm{v})} \mu$, but is equivalent in our setting.
3. For every smooth, compactly supported function $f : \mathbb{R} \to \mathbb{R}$, $\int f\, d\mu_d \to \int f\, d\mu$.
4. For every $f : \mathbb{R} \to \mathbb{C}$ of the form $f(x) = \exp(itx)$ for $t \in \mathbb{R}$, $\int f\, d\mu_d \to \int f\, d\mu$. These integrals, as a function of the parameter t, are the *characteristic functions* of μ_d and μ, and this result is called *Lévy's continuity theorem*.

The equivalence of the following further conditions to Condition 1 above is usually called the *portmanteau theorem*. To formulate them, we introduce the notion of a *continuity point* x of a measure μ on $\mathbb{R}$: it is a point where the cumulative distribution function $\mathsf{cdf}(\mu)$ is continuous (as we encountered above), or equivalently for which $\mu(\{x\}) = 0$.

5. $\mu_d(A) \to \mu(A)$ for all Borel sets A for which $\mu(\partial A) = 0$ (generalizing continuity points, such sets are sometimes called *continuity sets* of μ).
6. $\mu_d([a,b]) \to \mu([a,b])$ for all $a \leq b$ continuity points of μ.

7. $\mu_d((-\infty, a]) \to \mu((-\infty, a])$ for all a continuity points of μ. The condition is equivalently that $\mathsf{cdf}(\mu_d)(a) \to \mathsf{cdf}(\mu)(a)$ for all such a.

Note that the last three conditions may be viewed as analogous to the first four, where we consider families of indicator functions $f = \mathbb{1}_A$ for various families of sets A.

Note that Condition 6 above is precisely what we proposed earlier in (3.2.3). Thus we see that this is indeed equivalent to the definition we gave of weak convergence, provided the endpoints of the interval are continuity points of μ. That clause is to exclude edge cases of the following kind. Suppose that $\mu_d = \delta_{1/d}$, for instance, the empirical distribution of a few numbers that are all equal to $1/d$. Then, clearly the natural limit we wish this sequence to have is $\mu = \delta_0$. However, if we take $a = b = 0$, then $\mu_d([0,0]) = 0$ for all d, while $\mu([0,0]) = 1$. You may convince yourself that the condition above is a natural condition for excluding all such situations.

3.2.3 Random Weak Convergence

The above discussion concerns deterministic $\lambda_i^{(d)}$ and the empirical measures μ_d of such collections of points. We will be interested in the more complicated situation of random $\lambda_i^{(d)}$, usually $\lambda_i^{(d)} = \lambda_i(\boldsymbol{M}^{(d)})$ for $\boldsymbol{M}^{(d)} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ a random matrix. Then, $\lambda_i^{(d)}$ are themselves random, and $\mu_d = \mathsf{ed}(\boldsymbol{\lambda}^{(d)})$ is a *random measure*.

Essentially the same definition as Definition 3.2.2 is sensible, and will guarantee the same sort of bin-wise convergence of histograms. The only additional rub is that we must specify a *mode of convergence*: unlike before, the quantities $I_d := \int f\, d\mu_d = \frac{1}{d}\sum_{i=1}^d f(\lambda_i^{(d)})$ are now *random*, and we are asking for them to converge to the deterministic number $I := \int f\, d\mu$. You will recall that there are many levels of strength at which to ask that this happens. The ones that we will work with are the following:

1. $I_d \to I$ *in expectation* if $\mathbb{E}[I_d] \to I$.
2. $I_d \to I$ *in probability* if, for all $\epsilon > 0$, $\mathbb{P}[|I_d - I| > \epsilon] \to 0$.
3. $I_d \to I$ *in* L^2 if $\mathbb{E}[(I_d - I)^2] \to 0$, or equivalently (prove it if you do not see why it is true) if $I_d \to I$ in expectation and $\mathsf{Var}[I_d] \to 0$.

We therefore are justified in introducing the following definition.

Definition 3.2.5 (Random weak convergence)**.** Let μ_d be random probability measures on $\mathbb{R}$: random variables taking their values in the set of probability measures on $\mathbb{R}$. We say that μ_d converge *weakly in expectation* (respectively, *weakly in probability* and *weakly in* L^2) to μ, denoted $\mu_d \xrightarrow{(\mathrm{w},\mathbb{E})} \mu$ (respectively, $\mu_d \xrightarrow{(\mathrm{w},\mathbb{P})} \mu$ and $\mu_d \xrightarrow{(\mathrm{w},L^2)} \mu$), if, for all f bounded and continuous, $\int f\, d\mu_d$ converges in expectation (respectively, in probability and in L^2) to $\int f\, d\mu$.

Again, you can view this in the previous context as a statement about the convergence of random histograms, where we ask either for expected bin heights to converge, or for stronger conditions like (for L^2 convergence) the variance of their heights also decreasing.

Remark 3.2.6. Random measures can be slightly tricky to work with in general, but we will only ever study the simpler discrete random measures $\mathsf{ed}(\boldsymbol{\lambda}^{(d)})$ for $\boldsymbol{\lambda}^{(d)}$ a random vector.

It is natural to ask whether versions of the equivalent conditions from Theorem 3.2.4 for the deterministic case remain equivalent in the random case, if we fix a mode of convergence and replace all convergences of numbers in that Theorem with convergence of random variables in that mode. This is indeed the case, but requires some further assumptions: while in the proof of Theorem 3.2.4 one essentially approximates various classes of test functions f by one another and shows that the error terms that arise are small, for random measures these error terms will themselves be random, and one needs to show that they are small in a sense appropriate to the mode of convergence one works with.

3.3 The Combinatorial Method of Random Matrix Theory

How can we actually prove weak convergence of random measures in the above sense? And, how can we deal with specifically the empirical distributions of eigenvalues of random matrices? Let us give a name to these objects: for $\boldsymbol{M} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$, we define the *empirical spectral distribution*

$$\mathsf{esd}(\boldsymbol{M}) := \mathsf{ed}(\boldsymbol{\lambda}(\boldsymbol{M})) = \frac{1}{d}\sum_{i=1}^{d} \delta_{\lambda_i(\boldsymbol{M})} \in \mathcal{P}(\mathbb{R}).$$

Some of the salient issues are already apparent if we just consider trying to prove that $\mathsf{esd}(\boldsymbol{M}^{(d)})$ converges weakly to some μ for *deterministic* matrices $\boldsymbol{M}^{(d)} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$. For instance, the method of characteristic functions and Lévy's continuity theorem (Condition 4 of Theorem 3.2.4) would ask for us to show the convergence of

$$\frac{1}{d}\sum_{j=1}^{d} \exp(it\lambda_j(\boldsymbol{M}^{(d)}))$$

for all $t \in \mathbb{R}$, which will clearly be intractable in all but the simplest situations where we have a very precise understanding of the eigenvalues of $\boldsymbol{M}^{(d)}$ already.

Most often, as in the example of $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d\times m}$ we have studied already, the information we have about $\boldsymbol{M}$ will concern its *entries*. So, we would like a tool for proving weak convergence that allows us to use that information to analyze the eigenvalues. It turns out that one powerful idea for doing this is to study weak convergence through the *moments* of a measure.

We have seen the basic underlying algebraic fact already in Section 2.5.1. In our new language, it is that *the averages of polynomials over* $\mathsf{esd}(\boldsymbol{M})$ *are polynomials in the entries of* $\boldsymbol{M} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$. Concretely,

$$\begin{aligned}\int \lambda^k \, d(\mathsf{esd}(\boldsymbol{M}))(\lambda) &= \frac{1}{d}\sum_{i=1}^{d} \lambda_i(\boldsymbol{M})^k = \frac{1}{d}\,\mathsf{Tr}(\boldsymbol{M}^k) \\ &= \frac{1}{d}\sum_{i_1,\ldots,i_k=1}^{d} M_{i_1 i_2}\cdots M_{i_{k-1} i_k} M_{i_k i_1},\end{aligned}$$

which we have seen is a polynomial of the entries of $\boldsymbol{M}$. In particular, if we are interested in weak convergence in expectation, we wish to establish the convergence of $\mathbb{E}\left[\int f \, d(\mathsf{esd}(\boldsymbol{M}^{(d)}))\right]$ for various functions f. To do this for *polynomials* f, it suffices to consider the monomials $f(\lambda) = \lambda^k$, and the above expectation becomes

$$\mathbb{E}\left[\int f \, d(\mathsf{esd}(\boldsymbol{M}^{(d)}))\right] = \frac{1}{d}\sum_{i_1,\ldots,i_k=1}^{d} \mathbb{E}[M_{i_1 i_2}\cdots M_{i_{k-1} i_k} M_{i_k i_1}].$$

As we will see, we may understand the right-hand side of such an expression *combinatorially*, by considering which $(i_1, \dots, i_k) \in [d]^k$ make the biggest contributions and *how many* of these terms there are.

To give a complete outline, we will prove limit theorems for eigenvalue distributions by proceeding as follows:

1. Estimate $\mathbb{E}\left[\int f \, d(\mathsf{esd}(\boldsymbol{M}^{(d)}))\right]$ for $f(\lambda) = \lambda^k$ for each $k \geq 0$.
2. Establish convergence in expected moments of $\mathsf{esd}(\boldsymbol{M}^{(d)})$ to some μ for polynomial test functions, i.e., that $\mathbb{E}\left[\int f \, d(\mathsf{esd}(\boldsymbol{M}^{(d)}))\right] \to \int f \, d\mu$ for $f(\lambda) = \lambda^k$ for all $k \geq 0$, and thus for all polynomials f.
3. Establish convergence weakly in expectation for larger classes of functions f (in particular $f = \mathbb{1}_{[a,b]}$) by approximating them by polynomials, using tools like the Weierstrass approximation theorem.
4. Establish stronger modes of convergence by showing that, for f polynomials and then more general functions, the random variables $\int f \, d(\mathsf{esd}(\boldsymbol{M}^{(d)}))$ concentrate around their expectations.

We will see that, while they can be technical, Steps 2 through 4 can be performed mechanically for many applications. The art and nuance of this method is in the first step. We therefore call this general approach the **Combinatorial Method** of random matrix theory, because it reduces the matter of understanding eigenvalue distributions (and, as we will see, other related properties like norms) of random matrices to the combinatorics of the above expectations of entrywise monomials.

3.4 Warmup: Central Limit Theorem by Moments

Let us give an example of the Combinatorial Method for the weak convergence of deterministic measures, namely the CLT. The following result gives a general tool for dealing with the latter analytic steps of the method in this case.

Theorem 3.4.1 (Carleman's criterion)**.** Suppose $\mu_n, \mu \in \mathcal{P}(\mathbb{R})$ are all σ^2-subgaussian for some $\sigma^2 > 0$, and that

$$\mathop{\mathbb{E}}_{X \sim \mu_n}[X^k] = \int x^k \, d\mu_n(x) \to \int x^k \, d\mu = \mathop{\mathbb{E}}_{X \sim \mu}[X^k]$$

for each $k \geq 1$. Then, $\mu_n \xrightarrow{(\mathrm{w})} \mu$.

Proof Sketch. By Theorem 3.2.4, it suffices to show that the characteristic functions converge pointwise, i.e., that for all $t \in \mathbb{R}$,

$$\int \exp(itx) \, d\mu_n(x) \to \int \exp(itx) \, d\mu(x).$$

By Taylor's theorem, for any $K \geq 1$, we may estimate

$$\exp(itx) = \sum_{k=0}^{K} \frac{(itx)^k}{k!} + O\left(\frac{|t|^{K+1}|x|^{K+1}}{(K+1)!}\right).$$

Thus,

$$\begin{aligned}
&\left|\int \exp(itx)\, d\mu_n(x) - \int \exp(itx)\, d\mu(x)\right| \\
&\quad \leq \sum_{k=0}^{K} \left|\int \frac{(itx)^k}{k!}\, d\mu_n(x) - \int \frac{(itx)^k}{k!}\, d\mu(x)\right| \\
&\qquad + O\left(\int \frac{|t|^{K+1}|x|^{K+1}}{(K+1)!}\, d\mu_n(x) + \int \frac{|t|^{K+1}|x|^{K+1}}{(K+1)!}\, d\mu(x)\right).
\end{aligned}$$

For any given K, each term in the first sum converges to zero by our assumption. And, estimating subgaussian moments, the last term decays as $K^{-K/2}$. Thus, taking $K \to \infty$ gives the result. □

We will use this to show the following weak version of the CLT. A standard and simple truncation argument lets you extend this to the correct general statement without the subgaussianity assumption.

Theorem 3.4.2 (Subgaussian CLT)**.** Suppose that $X_1, X_2, \ldots$ are i.i.d., σ^2-subgaussian, and have $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[X_i^2] = 1$. Then,

$$\mu_n := \mathsf{Law}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i\right) \xrightarrow{\text{(w)}} \mathcal{N}(0,1) =: \mu.$$

Proof. Standard properties of subgaussian random variables show that μ_n, μ are all σ^2-subgaussian. Thus, by Theorem 3.4.1, it suffices to show the convergence of moments. Consider some $k \geq 1$. First, recall that

$$\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}[Z^k] = \left\{ \begin{array}{ll} 0 & \text{if } k \text{ is odd,} \\ (k-1)!! & \text{if } k \text{ is even} \end{array} \right\}.$$

Further, the right-hand side has the combinatorial interpretation of being the number of *matchings* of the set $[k]$, or the number of partitions into subsets of size exactly 2.

Now, consider the left-hand side of the limit. We have, expanding in a similar spirit to what we have suggested for matrices above,

$$\begin{aligned}
M_k &:= \mathbb{E}\left[\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i\right)^k\right] \\
&= \frac{1}{n^{k/2}} \sum_{i_1,\ldots,i_k=1}^{n} \mathbb{E}[X_{i_1} \cdots X_{i_k}].
\end{aligned}$$

Let us introduce some notation. First, write $\mathsf{Part}([k])$ for the set of *partitions* of the set $[k]$. A partition $\pi \in \mathsf{Part}([k])$ is a set of subsets $\pi = \{A_1, \ldots, A_\ell\}$, where the A_j are disjoint, non-empty, and $A_1 \sqcup \cdots \sqcup A_\ell = [k]$. We call the A_j the *blocks* of the partition. Thus we can use ordinary set-theoretic notation to talk about partitions, writing $|\pi|$ for the number of blocks in π, and $A \in \pi$ to say that A is a block of π.

Next, given $\boldsymbol{i} = (i_1, \ldots, i_k) \in [n]^k$, we write $\mathsf{type}(\boldsymbol{i}) \in \mathsf{Part}([k])$ for the partition where $a, b \in [k]$ are in the same block if and only if $i_a = i_b$. For example, with $k = 9$ and $n = 5$,

$$\mathsf{type}\big((1,5,1,1,5,3,3,4,2)\big) = \{\{1,3,4\},\{2,5\},\{6,7\},\{8\},\{9\}\}.$$

Thus we may group the above sum by type:

$$M_k = \frac{1}{n^{k/2}} \sum_{\pi \in \mathsf{Part}([k])} \sum_{\substack{\boldsymbol{i} \in [n]^k \\ \mathsf{type}(\boldsymbol{i}) = \pi}} \mathbb{E}[X_{i_1} \cdots X_{i_k}]$$

Now, note that if $\mathsf{type}(\boldsymbol{i}) = \pi$ with $\pi = \{A_1, \ldots, A_\ell\}$ and thus $|\pi| = \ell$, then there are some distinct $j_1, \ldots, j_\ell$ such that $X_{i_1} \cdots X_{i_k} = X_{j_1}^{|A_1|} \cdots X_{j_\ell}^{|A_\ell|}$. Since the X_{j_a} are independent, the expectation of such an expression factorizes as $\mathbb{E}[X_{i_1} \cdots X_{i_k}] = \mathbb{E}[X_{j_1}^{|A_1|}] \cdots \mathbb{E}[X_{j_\ell}^{|A_\ell|}]$, and since the X_{j_a} are further i.i.d., we have

$$= \frac{1}{n^{k/2}} \sum_{\pi \in \mathsf{Part}([k])} \sum_{\substack{\boldsymbol{i} \in [n]^k \\ \mathsf{type}(\boldsymbol{i}) = \pi}} \prod_{A \in \pi} \mathbb{E}[X_1^{|A|}]$$

Crucially, this manipulation tells us that all the terms of type π make *the same contribution* to this sum. So, we see a counting problem typical of the Combinatorial Method appear:

$$= \frac{1}{n^{k/2}} \sum_{\pi \in \mathsf{Part}([k])} \left(\prod_{A \in \pi} \mathbb{E}[X_1^{|A|}] \right) \#\{\boldsymbol{i} \in [n]^k : \mathsf{type}(\boldsymbol{i}) = \pi\}$$

Further, the indices of a given type are generated by assigning a distinct value to all positions in each block of π. Therefore,

$$= \frac{1}{n^{k/2}} \sum_{\pi \in \mathsf{Part}([k])} \left(\prod_{A \in \pi} \mathbb{E}[X_1^{|A|}] \right) n(n-1) \cdots (n - |\pi| + 1)$$

To simplify the exposition, let us allow ourselves an approximation here; as an exercise you may show it does not affect the result:

$$\begin{aligned} &\approx \frac{1}{n^{k/2}} \sum_{\pi \in \mathsf{Part}([k])} \left(\prod_{A \in \pi} \mathbb{E}[X_1^{|A|}] \right) n^{|\pi|} \\ &= \sum_{\pi \in \mathsf{Part}([k])} \left(\prod_{A \in \pi} \mathbb{E}[X_1^{|A|}] \right) n^{-k/2 + |\pi|} \end{aligned}$$

Finally, notice a few more properties of this sum. First, every term where π has a block of size 1 has an $\mathbb{E}[X_1] = 0$ factor, and thus does not contribute. So, all non-zero terms involve π having all blocks of size at least 2, and thus $|\pi| \le k/2$. If $|\pi| < k/2$, then the corresponding term is $o(1)$ as $n \to \infty$. And, the number of terms in the entire sum does not depend on n. Therefore, the sum of *all* terms for π with $|\pi| < k/2$ or π having a block of size 1 is $o(1)$. Finally, for π with all blocks of size 2, since $\mathbb{E}[X_1^2] = 1$, the first factor involving expectations equals 1. Thus we have

$$= \#\{\pi \in \mathsf{Part}([k]) : \pi \text{ is a matching}\} + o(1),$$

since the π with $|\pi| = k/2$ and $|A| = 2$ for all $A \in \pi$ are precisely the matchings. □

3.5 Wigner Semicircle Limit Theorem

We will prove in detail using the Combinatorial Method the following theorem, one of the first and most important results of classical random matrix theory.

Definition 3.5.1 (Wigner matrix)**.** Let ν be a probability measure on $\mathbb{R}$. We write $\mathrm{Wig}(d, \nu)$ for the probability measure on $\mathbb{R}^{d\times d}_{\mathsf{sym}}$ where we sample $\boldsymbol{W} \sim \mathrm{Wig}(d, \nu)$ by drawing $W_{ij} = W_{ji} \sim \nu$ i.i.d. for $1 \le i < j \le d$ and setting $W_{ii} := 0$.

Definition 3.5.2 (Semicircle measure)**.** The *(Wigner) semicircle measure* is the probability measure, denoted μ_{SC}, supported on $[-2, 2]$ with density $\frac{1}{2\pi}\sqrt{4 - x^2}$ on that interval.

Theorem 3.5.3 (Wigner's limit theorem)**.** Suppose that ν is a probability measure on $\mathbb{R}$ with $\int x\, d\nu(x) = 0$ and $\int x^2\, d\nu(x) = 1$. Let $\boldsymbol{W}^{(d)} \sim \mathrm{Wig}(d, \nu)$ for each $d \ge 1$. Then,

$$\mathsf{esd}\left(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}\right) \xrightarrow{(\mathrm{w},\mathcal{M})} \mu_{\mathsf{SC}}$$

for each mode of convergence $\mathcal{M} \in \{\mathbb{E}, \mathbb{P}\}$.

Remark 3.5.4 (Relation to rectangular matrices)**.** Recall that we saw the semicircle distribution show up in numerical experiments with the matrices $\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{g}_i\boldsymbol{g}_i^\top$ for $\boldsymbol{g}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ i.i.d., when $m \gg d$. In that case, the semicircle had a radius of $2\sqrt{d/m}$ and was centered at 1, so we would expect to see μ_{SC} appear if we instead considered

$$\sqrt{\frac{m}{d}}\left(\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{g}_i\boldsymbol{g}_i^\top - \boldsymbol{I}_d\right) = \frac{1}{\sqrt{dm}}\sum_{i=1}^{m}(\boldsymbol{g}_i\boldsymbol{g}_i^\top - \boldsymbol{I}_d).$$

Indeed you may check numerically that this is the case. It is also reasonable that this random matrix behaves somewhat like a Wigner matrix, at least if we ignore the diagonal. You may check that the strictly upper-triangular entries of $\boldsymbol{g}_i\boldsymbol{g}_i^\top$ are pairwise uncorrelated and have variance 1. If we took $m \to \infty$ for d fixed, then the above matrix, after setting its diagonal to zero, would converge in distribution to $\frac{1}{\sqrt{d}}\boldsymbol{W}$ for $\boldsymbol{W} \sim \mathrm{Wig}(d, \mathcal{N}(0, 1))$ by the multivariate CLT. Thus it should be plausible as a "softening" of this that the $m \gg d$ limit, where both parameters grow but m grows faster, would give rise to the behavior of a Wigner matrix as well.

Remark 3.5.5 (Universality)**.** One important aspect of this theorem is that it gives a *universal* limit regardless (essentially) of the entrywise distribution. You may think of this as analogous to the CLT, which gives a universal limit for the sums of i.i.d. random variables essentially regardless of their distributions (up to centering and scaling). As stated in Theorem 3.5.3, it suffices to just have $\int x\, d\nu(x) = 0$ and $\int x^2\, d\nu(x) = 1$, but no further moments beyond the second need necessarily be finite. Our moment-method proof will first assume that all moments are finite. A truncation argument allows us to remove that auxiliary assumption.

We will proceed according to the same strategy we outlined before.

3.5.1 Convergence of Expected Moments

As for the CLT, the main ingredient of the proof is the following.

Definition 3.5.6. We say that a sequence of random measures $\mu_d \in \mathcal{P}(\mathbb{R})$ *converges in expected moments* to a deterministic $\mu \in \mathcal{P}(\mathbb{R})$ if, for all $k \geq 1$,

$$\mathbb{E}\left[\int x^k \, d\mu_d(x)\right] \to \int x^k \, d\mu(x).$$

We denote this $\mu_d \xrightarrow{\mathbb{E}\text{ mom.}} \mu$.

Lemma 3.5.7. In the setting of Theorem 3.5.3, if all moments of ν are finite, then

$$\mathsf{esd}\left(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}\right) \xrightarrow{\mathbb{E}\text{ mom.}} \mu_{\mathsf{SC}}.$$

That is, for all $k \geq 1$, we have

$$\lim_{d\to\infty} \mathbb{E}\left[\frac{1}{d}\,\mathsf{Tr}\left(\left(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}\right)^k\right)\right] = \mathop{\mathbb{E}}_{X\sim\mu_{\mathsf{SC}}}[X^k] = \int_{-2}^{2} x^k \cdot \frac{1}{2\pi}\sqrt{4-x^2}\,dx.$$

We separate out the computation of the moments of the semicircle distribution into a separate claim. The result is the following fundamental integer sequence; we will already see several collections of objects that these numbers count below.

Definition 3.5.8 (Catalan numbers)**.** We denote by Cat_ℓ the *Catalan numbers*, defined to satisfy the recurrence

$$\mathsf{Cat}_0 = 1,$$
$$\mathsf{Cat}_{\ell+1} = \sum_{k=0}^{\ell} \mathsf{Cat}_k \mathsf{Cat}_{\ell-k} \text{ for all } \ell \geq 0.$$

The following gives a combinatorial interpretation that is straightforward to prove by showing that the number of non-crossing matchings also satisfies the above recurrence.

Proposition 3.5.9. Cat_ℓ is the number of *non-crossing matchings* of $[2\ell]$, matchings that have no crossings when drawn as a matching of points arranged in a line.

For example, the first two matchings of $[4]$ in Figure 3.3 are non-crossing, while the last has a crossing.

There is also an explicit formula for Cat_ℓ as follows, though we will not need to use it for our purposes.

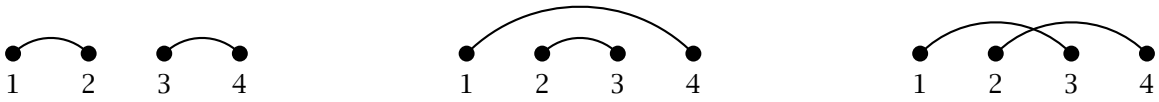


Figure 3.3: The three matchings, two non-crossing and one crossing, of four objects that arise in the computation of the fourth moments of $\mathcal{N}(0,1)$ and μ_{SC}.

k	0	2	4	6	8	10	12	14	16
$\mathbb{E}_{X\sim\mathcal{N}(0,1)}[X^k]$	1	1	3	15	105	945	10 395	135 135	2 027 025
$\mathbb{E}_{X\sim\mu_{\mathsf{SC}}}[X^k]$	1	1	2	5	14	42	132	429	1430

Table 3.1: A comparison of the first few moments of $\mathcal{N}(0,1)$ (the "matching numbers") and of μ_{SC} (the Catalan numbers).

Proposition 3.5.10. For all $\ell \geq 0$, $\mathsf{Cat}_\ell = \frac{1}{\ell+1}\binom{2\ell}{\ell}$.

Two different approaches to the proof are given in Exercise 3.8.1.

Proposition 3.5.11. The moments of the semicircle law are given by

$$\mathop{\mathbb{E}}_{X\sim\mu_{\mathsf{SC}}}[X^k] = \#\{\text{non-crossing matchings of } [k]\} = \left\{ \begin{array}{ll} 0 & \text{if } k \text{ is odd} \\ \mathsf{Cat}_\ell & \text{if } k = 2\ell \text{ is even} \end{array} \right\}. \tag{3.5.1}$$

A guided version of the proof is given in Exercise 3.8.2.

Generally, there are many fewer non-crossing matchings than general matchings, as we can see from the scaling behaviors that, roughly, the total number of matchings is $\approx k^{k/2}$ while the number of non-crossing matchings is $\approx 2^k$ (both by Stirling's formula, omitting lower order terms). The first few of each sequence are given in Table 3.1.

Proof of Lemma 3.5.7. With the above result in hand, we may proceed to the main calculation of the moments of the empirical spectral distributions, which is more complicated than but very much parallels the one from the CLT. The main complication is that the "template" object that we use to group terms by becomes considerably more complicated than a partition. We let $\boldsymbol{W} := \boldsymbol{W}^{(d)}$ to lighten the notation and calculate:

$$\begin{aligned} m_{d,k} &:= \mathbb{E}\left[\frac{1}{d}\,\mathsf{Tr}\left(\left(\frac{1}{\sqrt{d}}\boldsymbol{W}\right)^k\right)\right] \\ &= \frac{1}{d^{k/2+1}}\mathbb{E}[\mathsf{Tr}(\boldsymbol{W}^k)] \end{aligned}$$

and by expanding the matrix multiplication in indices,

$$= \frac{1}{d^{k/2+1}} \sum_{i_1,\ldots,i_k\in[d]} \mathbb{E}[W_{i_1 i_2} W_{i_2 i_3} \cdots W_{i_{k-1} i_k} W_{i_k i_1}].$$

We associate to each such term the following combinatorial object, giving a notion of the "type" of each term above. The terminology below is non-standard but useful for us.

Definition 3.5.12 (Closed walk graph)**.** A directed multigraph G is a *closed walk graph* if it has the following properties:

1. There is a distinguished *root vertex* $\mathsf{root} = \mathsf{root}(G) \in V(G)$.
2. The (directed) edges of G are numbered 1 through $|E(G)|$, and when traversed in this sequence form a closed walk starting and ending at root that visits every vertex.

You may convince yourself that such a graph precisely describes the pattern of equalities among the indices in a term above, where we "forget" the actual values of the i_a. See Figure 3.4 for an example. We consider these graphs up to isomorphism preserving the root and the edge ordering.

We write $V(G)$ and $E(G)$ for the vertex and edge set of a closed walk graph. Also, let us write $UE(G)$ for the multiset of edges of G stripped of the information of their direction (i.e., replacing directed edge (i, j) with undirected edge $\{i, j\}$, but retaining repetitions). And, introduce an independent random variable $W_e \sim \nu$ for each distinct undirected edge e occurring in $UE(G)$. Since $W_{ii} = 0$, all terms making a non-zero contribution will never have $i_a = i_{a+1}$ (with index arithmetic done modulo k), so the associated closed walk graph will never have vertices with self-loops. Thus, we may rewrite

$$m_{d,k} = \frac{1}{d^{k/2+1}} \sum_{\substack{G \text{ closed walk graph} \\ G \text{ loopless} \\ |E(G)|=k}} d(d-1)\cdots(d-|V(G)|+1) \cdot \mathbb{E}\left[\prod_{e \in UE(G)} W_e\right].$$

Further, let $UUE(G)$ be the set of unique undirected edges of G, and let the *multiplicity* $m(e)$ of $e \in UUE(G)$ be the number of times that that edge occurs (in either orientation) in G. We may then factorize as in our CLT argument,

$$= \frac{1}{d^{k/2+1}} \sum_{\substack{G \text{ closed walk graph} \\ G \text{ loopless} \\ |E(G)|=k}} d(d-1)\cdots(d-|V(G)|+1) \cdot \prod_{e \in UUE(G)} \mathop{\mathbb{E}}_{W \sim \nu}[W^{m(e)}]$$

and finally as before since $\mathbb{E}_{W\sim\nu}[W] = 0$, we may restrict

$$= \frac{1}{d^{k/2+1}} \sum_{\substack{G \text{ closed walk graph} \\ G \text{ loopless} \\ |E(G)|=k \\ m(e)\geq 2 \text{ for each } e \in UUE(G)}} d(d-1)\cdots(d-|V(G)|+1) \cdot \prod_{e \in UUE(G)} \mathop{\mathbb{E}}_{W \sim \nu}[W^{m(e)}],$$

which just asks that each edge appear at least twice in G, in either orientation.

Recalling what happened next in the CLT argument, we must find which terms make the largest contribution. Since k is constant, this amounts to understanding for which G (satisfying the conditions in the sum above) we have $|V(G)| = k/2 + 1 = |E(G)|/2 + 1$. Consider $\widetilde{G}$ the graph where each edge of G is stripped of its orientation, and parallel edges are collapsed down to a single edge, so that $\widetilde{G}$ is a simple undirected connected graph. We have $|E(\widetilde{G})| \leq |E(G)|/2$ by the last restriction we made above, while $|V(\widetilde{G})| = |V(G)|$.

Since $\widetilde{G}$ is connected, we have $|V(\widetilde{G})| \leq |E(\widetilde{G})| + 1$, with equality if and only if $\widetilde{G}$ is a tree. (Proof: a simple induction establishes the result with equality for trees, and any connected non-tree

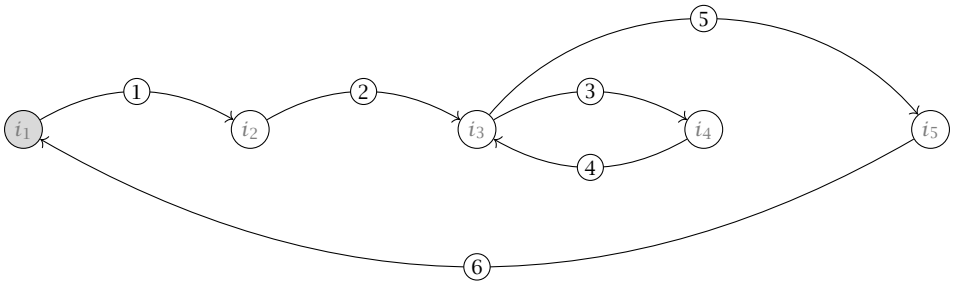


Figure 3.4: The closed walk graph associated to a term of the form $W_{i_1 i_2} W_{i_2 i_3} W_{i_3 i_4} W_{i_4 i_3} W_{i_3 i_5} W_{i_5 i_1}$ in $\mathsf{Tr}(\boldsymbol{W}^6)$, where $i_1, \dots, i_5$ are all distinct. We include these index labels for reference, but we think of a closed walk graph as *not* being decorated with those indices, but rather only the ordering of the directed edges traversing them. We do include the identity of the "root" or starting vertex of the closed walk, which is whatever vertex had the label i_1, which we show as shaded.

has a spanning tree, for which equality holds, as well as further edges.) Thus we have $|V(G)| \leq |E(G)|/2 + 1$, with equality if and only if G is a *(rooted) tree traversal*: a rooted tree where each edge occurs exactly twice, directed in opposite directions, and where the edges are numbered so that, if followed in order, they traverse each directed edge exactly once. When G satisfies this equality, then $m(e) = 2$ for all $e \in UUE(G)$. Thus, we find a counting interpretation of the limiting empirical moments,

$$m_{d,k} = \mathbb{E}\left[\frac{1}{d}\,\mathsf{Tr}\left(\frac{1}{\sqrt{d}}\boldsymbol{W}\right)^k\right] \to \sum_{\substack{G \text{ a rooted tree traversal} \\ |E(G)|=k}} 1$$

$$= \#\{\text{rooted tree traversals with } k \text{ edges}\}.$$

It remains to show that the number of rooted tree traversals with k directed edges (of a tree with $k/2$ undirected edges) is $\mathsf{Cat}_{k/2}$. Clearly this number is zero when k is odd by parity considerations. Suppose then that $k = 2\ell$ is even. We want to show that the number of tree traversals where the underlying tree has ℓ edges is Cat_ℓ.

It is useful to make a few transformations of our description of the Catalan numbers as counting non-crossing matchings (proving which we leave as an exercise). First, Cat_ℓ is the number of *parenthesizations* of length 2ℓ: the number of ways to arrange 2ℓ parentheses that adhere to the usual rules of grammar, where every parenthesis gets closed and you only close a parenthesis that has been opened. For instance, "`(()())`" and "`()(())()`" are valid, while "`())(`" is not, even though the number of open and closed parentheses is equal.

Next, Cat_ℓ is in turn equal to the number of *Dyck paths* of length 2ℓ, the paths in $\mathbb{Z}^2$ of length 2ℓ that start at the origin, return to the x-axis, move right and either up or down with every step, and never go below the x-axis. Finally, a tree traversal gives a Dyck path by recording the distance from the root at each step. Conversely, read a Dyck path from left to right, starting at a root vertex: on each up-step create a new child and move to it, and on each down-step return to the parent. These constructions are inverse because a tree traversal must finish exploring each subtree before returning to its parent, as each edge is crossed only once in each direction.

See Figure 3.5 for examples of all of these bijections. Note that a tree traversal, being a closed walk graph, is supposed to be decorated with information about the order in which edges are traversed, but because of its special structure beyond just a closed walk graph, we can replace the

labels on edges, which are a little redundant, with an ordering on vertices, as we do above—the order in which the vertices are first visited fully determines the order in which the directed edges are traversed. □

3.5.2 Upgrading to Weak Convergence in Probability

We will now use the convergence of moments to sketch the proof of a version of Theorem 3.5.3 involving weak convergence in probability, provided that all moments of ν are finite. It is straightforward to then deduce the version of this for convergence of counts of eigenvalues in an interval, as follows.

Corollary 3.5.13. In the setting of Theorem 3.5.3, for any $a < b$, $\frac{1}{d}\#\{i : \lambda_i(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}) \in [a,b]\} \to \mu_{\mathsf{SC}}([a,b])$ in probability.

Proof. Write $\lambda_i^{(d)} := \lambda_i(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)})$. For any $0 < \epsilon < \frac{b-a}{2}$, we may define functions $f, g : \mathbb{R} \to \mathbb{R}$ as follows:

$$f(t) := \left\{ \begin{array}{ll} 0 & \text{if } t \le a, \\ 1 - \frac{a+\epsilon-t}{\epsilon} & \text{if } a \le t \le a+\epsilon, \\ 1 & \text{if } a+\epsilon \le t \le b-\epsilon, \\ \frac{b-t}{\epsilon} & \text{if } b-\epsilon \le t \le b, \\ 0 & \text{if } t \ge b \end{array} \right\},$$

$$g(t) := \left\{ \begin{array}{ll} 0 & \text{if } t \le a-\epsilon, \\ 1 - \frac{a-t}{\epsilon} & \text{if } a-\epsilon \le t \le a, \\ 1 & \text{if } a \le t \le b, \\ \frac{b+\epsilon-t}{\epsilon} & \text{if } b \le t \le b+\epsilon, \\ 0 & \text{if } t \ge b+\epsilon \end{array} \right\},$$

These are continuous and bounded, and satisfy $f(t) \le \mathbb{1}\{t \in [a,b]\} \le g(t)$, and therefore

$$\frac{1}{d}\sum_{i=1}^{d} f(\lambda_i^{(d)}) \le \frac{1}{d}\#\{i : \lambda_i^{(d)} \in [a,b]\} \le \frac{1}{d}\sum_{i=1}^{d} g(\lambda_i^{(d)}).$$

We also have

$$\begin{aligned} &\left| \int f \, d\mu_{\mathsf{SC}} - \mu_{\mathsf{SC}}([a,b]) \right| \\ &\quad = \left| \int (f - \mathbb{1}_{[a,b]}) \, d\mu_{\mathsf{SC}} \right| \\ &\quad = \left| \int_{-\infty}^{\infty} (f(t) - \mathbb{1}_{[a,b]}(t)) \rho_{\mathsf{SC}}(t) \, dt \right| \\ &\quad = \left| \int_{a}^{a+\epsilon} (f(t) - \mathbb{1}_{[a,b]}(t)) \rho_{\mathsf{SC}}(t) \, dt + \int_{b-\epsilon}^{b} (f(t) - \mathbb{1}_{[a,b]}(t)) \rho_{\mathsf{SC}}(t) \, dt \right| \\ &\quad \le \left| \int_{a}^{a+\epsilon} (f(t) - \mathbb{1}_{[a,b]}(t)) \rho_{\mathsf{SC}}(t) \, dt \right| + \left| \int_{b-\epsilon}^{b} (f(t) - \mathbb{1}_{[a,b]}(t)) \rho_{\mathsf{SC}}(t) \, dt \right| \\ &\quad \le 2 \cdot \epsilon \cdot 2 \cdot \max_{t \in \mathbb{R}} \rho_{\mathsf{SC}}(t) \\ &\quad \le C\epsilon. \end{aligned}$$

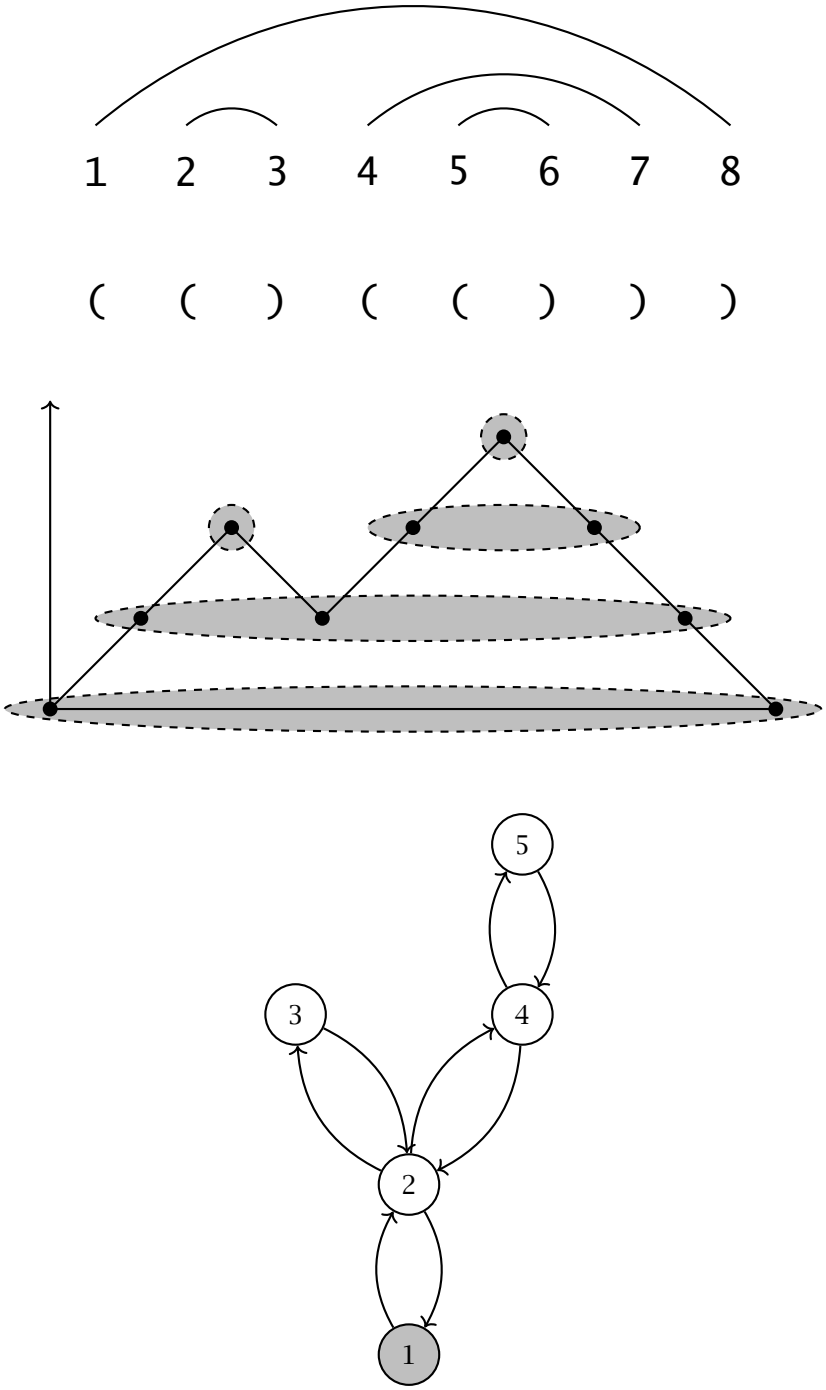


Figure 3.5: An example of the bijection between four kinds of objects that the Catalan numbers (Cat_4 in this case) count, as used in the proof of Lemma 3.5.7: non-crossing matchings, parenthesizations, Dyck paths, and rooted tree traversals. We also show the mapping from Dyck paths to tree traversals by showing which vertices along the Dyck path are identified to form the tree.

for some absolute constant C, and likewise with f replaced by g. In particular, we find that, using Theorem 3.5.3 (with weak convergence in probability) on f and g and combining with the above observation, for any ϵ,

$$\lim_{d\to\infty} \mathbb{P}\left[\mu_{\mathsf{SC}}([a,b]) - 2C\epsilon \le \frac{1}{d}\#\{i : \lambda_i^{(d)} \in [a,b]\} \le \mu_{\mathsf{SC}}([a,b]) + 2C\epsilon\right] = 1.$$

But, taking $\epsilon \to 0$, this is just by definition the convergence in probability given in the statement. □

We now sketch the rest of the proof of the weak convergence in probability statement. Let us write $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{W}}^{(d)} := \frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}$ and $\mu_d := \mathsf{esd}(\widehat{\boldsymbol{W}})$ again; the below technique is sensible in a much wider range of settings. We want to show that, for any bounded continuous f,

$$\int f\, d\mu_d \xrightarrow{\mathbb{P}} \int f\, d\mu_{\mathsf{SC}}.$$

On the other hand, we only know that, extending Lemma 3.5.7 by linearity, for any *polynomial* p, we have

$$\int p\, d\mu_d \xrightarrow{\mathbb{E}} \int p\, d\mu_{\mathsf{SC}}.$$

To get from the latter to the former, we have to handle two differences: the difference between convergence in expectation and convergence in probability, and the difference between polynomials and bounded continuous functions.

To deal with the second issue, the following will be our main tool.

Theorem 3.5.14 (Weierstrass approximation)**.** Let $f : [-B,B] \to \mathbb{R}$ be a continuous function on a compact interval and $\epsilon > 0$. Then, there exists a polynomial $p \in \mathbb{R}[x]$ such that, for all $x \in [-B,B]$, $|f(x) - p(x)| \le \epsilon$.

Now, suppose we are given a bounded continuous f. Since μ_{SC} is supported on $[-2,2]$, we expect that μ_d should not have many outlier eigenvalues far from this interval. So, pick some $B > 2$ and $\epsilon > 0$, and let p be a polynomial approximating f on $[-B,B]$ as above. We may then decompose

$$\begin{aligned}
\left|\int f\, d\mu_d - \int f\, d\mu_{\mathsf{SC}}\right| &= \left|\frac{1}{d}\sum_{i=1}^{d} f(\lambda_i) - \int f\, d\mu_{\mathsf{SC}}\right| \\
&\le \left|\frac{1}{d}\sum_{i=1}^{d} p(\lambda_i) - \mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d} p(\lambda_i)\right]\right| \\
&\quad + \left|\mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d} p(\lambda_i)\right] - \int p\, d\mu_{\mathsf{SC}}\right| \\
&\quad + \left|\frac{1}{d}\sum_{i:|\lambda_i|\le B} (f(\lambda_i) - p(\lambda_i))\right| \\
&\quad + \left|\frac{1}{d}\sum_{i:|\lambda_i|>B} f(\lambda_i)\right| \\
&\quad + \left|\frac{1}{d}\sum_{i:|\lambda_i|>B} p(\lambda_i)\right| \\
&\quad + \left|\int (f-p)\, d\mu_{\mathsf{SC}}\right|
\end{aligned}$$

The second term converges to zero by convergence of expected moments. The third and sixth terms are small (specifically, at most ϵ) by our assumption on p approximating f on $[-B, B]$. The trickier error terms are the first, fourth, and fifth terms, whose treatment you will see in Exercise 3.8.3, showing that

$$\left| \frac{1}{d} \sum_{i=1}^{d} p(\lambda_i) - \mathbb{E}\left[\frac{1}{d} \sum_{i=1}^{d} p(\lambda_i) \right] \right| \xrightarrow{\mathbb{P}} 0,$$
$$\frac{1}{d} \sum_{i:|\lambda_i|>B} |f(\lambda_i)| \xrightarrow{\mathbb{P}} 0,$$
$$\frac{1}{d} \sum_{i:|\lambda_i|>B} |p(\lambda_i)| \xrightarrow{\mathbb{P}} 0.$$

The first statement asks that $\int p \, d\mu_d$ concentrate around its expectation. The second and third both, in slightly different ways, ask to control outlier eigenvalues from the limiting support $[-2, 2]$. Further, in Exercise 3.8.4, you will remove the assumption that all moments of ν are finite and fully prove Theorem 3.5.3 with convergence in probability.

More generally the above are the two obstructions that may arise to extending convergence of expected moments to weak convergence in probability (or stronger modes):

1. The empirical moments of eigenvalues may converge in expectation to a limit, but not concentrate around that limiting value. In this case, you can think roughly speaking that weak convergence fails even for the simple case of polynomial test functions (this is not precisely valid since the definition of weak convergence does not allow polynomial test functions, but is an accurate intuition still).

2. There may be many or very large outlier eigenvalues which make the moments—which are affected by outliers—behave in ways that are not representative of the actual weak limit.

3.5.3 Universality

One of the remarkable things about Theorem 3.5.3 is the *universality* statement that it contains. Namely, after the truncation argument in Exercise 3.8.4, the proof gives the full statement of the theorem: the limiting eigenvalue distribution of matrices drawn from $\mathsf{Wig}(d, \nu)$ as $d \to \infty$ is the same provided only that ν has mean 0 and variance 1. You should think of this as being similar to the universality latent in the central limit theorem, which makes precisely the same kind of statement for sums of i.i.d. random variables distributed according to ν.

As in the case of the central limit theorem, this universality phenomenon actually can be pushed considerably beyond the case of $W_{ij} \sim \nu$ for some ν that does not depend on either d or i, j. Let us give (without proof) two examples of even more general statements; we will mention some more later and plenty can be found throughout the literature.

Sparse Random Matrices First, consider allowing ν to depend on d. We consider a simple example given by "sparsifying" the distribution $\mathsf{Unif}(\{\pm 1\})$, often called the *Rademacher* distribution: for a parameter $p \in (0, 1]$, define the *sparse Rademacher* distribution as the probability measure

$$\mathsf{SparseRad}(p) := \frac{p}{2}\delta_{-1/\sqrt{p}} + (1-p)\delta_0 + \frac{p}{2}\delta_{1/\sqrt{p}}.$$

Note that we have arranged so that, for any p, these measures all have mean 0 and variance 1. If p is a constant, then Theorem 3.5.3 applies and gives convergence of the empirical spectral distribution to the semicircle law of normalizations of $\boldsymbol{W}^{(d)} \sim \mathsf{Wig}(d, \mathsf{SparseRad}(p))$.

But, it turns out to be permissible to take $p = p(d)$ decreasing sufficiently slowly with d. In particular, we have the following.

Theorem 3.5.15. Let $p(d) \in (0,1]$ for each $d \in \mathbb{N}$. Consider Wigner random matrices $\boldsymbol{W}^{(d)} \sim \mathsf{Wig}(d, \mathsf{SparseRad}(p(d)))$ and write $\mu_d := \mathsf{esd}(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)})$. The following hold:

1. If $p(d) \cdot d \to \infty$ as $d \to \infty$, then $\mu_d \xrightarrow{(\mathrm{w},\,\mathbb{P})} \mu_{\mathsf{SC}}$.
2. If $p(d) \cdot d \to c \in (0, \infty)$, then $\mu_d \xrightarrow{(\mathrm{w},\,\mathbb{P})} \gamma_c$ for some probability measure $\gamma_c \neq \mu_{\mathsf{SC}}$.

We note that the condition on $p(d) \cdot d$ is precisely the same as the condition on the central limit theorem holding for such sparse random variables: if $p(d) \cdot d \to \infty$ then the sum of d i.i.d. random variables distributed as $\mathsf{SparseRad}(p(d))$, divided by $\sqrt{d}$, has a standard Gaussian limit. If instead $p(d) \cdot d \to c \in (0, \infty)$, then this normalized sum converges in distribution to $(P_+ - P_-)/\sqrt{c}$, where P_+, P_- are independent $\mathsf{Pois}(c/2)$ random variables.

A simple reason for the breakdown of the semicircle limit theorem when $p(d) \cdot d \to c$ is as follows. The probability that a given column of $\boldsymbol{W}^{(d)}$ is identically zero is roughly $(1 - p(d))^d$. When $p(d) \gg 1/d$, then this converges to zero, but when $p(d) \sim 1/d$, then it converges to a constant. In the latter case, we expect a constant fraction of the columns of $\boldsymbol{W}^{(d)}$ to be zero. Since whenever column i of a matrix is zero then $\boldsymbol{e}_i$ is an eigenvector with eigenvalue zero, this means in the above situation we expect a constant fraction of eigenvalues of $\boldsymbol{W}^{(d)}$ to be zero. Thus, the limit of μ_d would need to include a point mass at zero, which μ_{SC} does not. (Continuing the analogy with sums of i.i.d. random variables, when $p(d) \cdot d \to c$, the probability that all d summands are zero tends to $\exp(-c) > 0$, so the limiting distribution has a point mass at zero.)

INHOMOGENEOUS RANDOM MATRICES We may also allow for the entry distributions of $\boldsymbol{W}^{(d)}$ to be different, provided they have some shared regularity conditions. One simple such setting is the following.

Theorem 3.5.16. Suppose that the symmetric arrays $\sigma_{ji} = \sigma_{ij} = \sigma_{ij}^{(d)} > 0$ satisfy $c \leq \sigma_{ij}^2 \leq C$ for some constants $c, C > 0$ independent of d, for all $d \geq 1$ and $i, j \in [d]$. Suppose also that, for all $d \geq 1$ and $i \in [d]$,

$$\sum_{j=1}^{d} \sigma_{ij}^2 = d.$$

Suppose that $\boldsymbol{W}^{(d)}$ are random matrices with $W_{ji}^{(d)} = W_{ij}^{(d)} \sim \mathcal{N}(0, \sigma_{ij}^2)$ independently and let $\mu_d := \mathsf{esd}(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)})$. Then, $\mu_d \xrightarrow{(\mathrm{w},\,\mathbb{P})} \mu_{\mathsf{SC}}$.

You should think of such results as being akin to the "triangular array" versions of the central limit theorem.

3.6 Combinatorial Method for Extreme Eigenvalues

We also sketch how we can control the norm of a matrix, $\|\boldsymbol{W}\| = \max\{|\lambda_d(\boldsymbol{W})|, |\lambda_1(\boldsymbol{W})|\}$, by similar moment computations, for the case of Wigner matrices. Let us see why our calculations so far do not suffice.

First, note that μ_d converging weakly in any of the modes we have discussed to μ_{SC} is certainly insufficient. Note that, in fact, this weak convergence is not affected at all by the value of $\lambda_1(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)})$. The same is true even for deterministic weak convergence: no point mass with weight $1/d$ affects the weak convergence of probability measures, since it only contributes $O(1/d)$ to the integral of any bounded function.

Still, we may try to reuse our calculations to bound the spectral norm, as follows. We may write the finding of Lemma 3.5.7 in a way more relevant to our setting as follows. Assume for now that all moments of ν are finite. Given the entrywise distribution ν, for each k, there is a $\Delta_{2k} = \Delta_{2k}(\nu)$ such that

$$\mathbb{E}\left[\mathsf{Tr}\left(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}\right)^{2k}\right] \in [d \cdot \mathsf{Cat}_k - \Delta_{2k}, d \cdot \mathsf{Cat}_k + \Delta_{2k}].$$

The point of this for our purposes in Theorem 3.5.3 was that, since Δ_{2k} does not depend on d, upon dividing by d the limit of the above is Cat_k.

For extreme eigenvalues, suppose for example that we want to control the probability

$$\mathbb{P}\left[\left\|\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}\right\| \geq t\right].$$

In particular, since the support of μ_{SC} is $[-2, 2]$, we would like to show that this decays rapidly in d as soon as $t = 2 + \epsilon$. We can try to bound as follows by a Chebyshev-like inequality:

$$\begin{aligned}
\mathbb{P}\left[\left\|\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}\right\| \geq t\right] &= \mathbb{P}\left[\left\|\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}\right\|^{2k} \geq t^{2k}\right] \\
&\leq \mathbb{P}\left[\mathsf{Tr}\left(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}\right)^{2k} \geq t^{2k}\right] \\
&\leq \frac{1}{t^{2k}}\mathbb{E}\left[\mathsf{Tr}\left(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}\right)^{2k}\right] \\
&\leq \frac{d \cdot \mathsf{Cat}_k}{t^{2k}} + \frac{\Delta_{2k}}{t^{2k}} \\
&\leq \left(\frac{2}{t}\right)^{2k} d + \frac{\Delta_{2k}}{t^{2k}}
\end{aligned}$$

We see a few important features.

First, for fixed $t = 2 + \epsilon$, the first term decays if $k = k(d)$ is a sufficiently large multiple of $\log d$, with the constant depending on ϵ. One convenient choice that works for every fixed $\epsilon > 0$ is $k(d) = \lfloor (\log d)^2 \rfloor$. (With k constant, the bound only tends to zero when $t/d^{1/(2k)} \to \infty$, which is far from a tight bound.)

On the other hand, to set $t = 2 + \epsilon$, we need finer control over the error term than we have established. Our analysis thus far only controls it for each fixed k. We need an estimate that remains useful when k grows with d, which requires a more careful count of the walks in Lemma 3.5.7. The following is one version of this.

Lemma 3.6.1. Suppose that ν satisfies the assumptions of Theorem 3.5.3 and in addition has $\mathbb{E}_{w\sim\nu}[|w|^r] \le r^{Cr}$ for some $C > 0$ and every integer $r \ge 1$ (for instance, ν can be any bounded or subgaussian distribution). Then, there are constants $A, B > 0$ depending only on C such that, for all sufficiently large d and integers $1 \le k \le (\log d)^2$,

$$\left| \mathbb{E}\left[\mathsf{Tr}\left(\frac{1}{\sqrt{d}} \boldsymbol{W}^{(d)} \right)^{2k} \right] - d\mathsf{Cat}_k \right| \le Ak^B 2^{2k}.$$

The restriction on k matters: this is not a bound for arbitrarily high moments at fixed d. For the counting argument behind such estimates, see [Tao12, Section 2.3]. Using the lemma in the bound above with $k = \lfloor (\log d)^2 \rfloor$ gives, for every fixed $t > 2$,

$$\mathbb{P}\left[\left\| \frac{1}{\sqrt{d}} \boldsymbol{W}^{(d)} \right\| \ge t \right] \le (d + Ak^B) \left(\frac{2}{t} \right)^{2k} \to 0.$$

Together with the semicircle law, this gives the following limit theorem for the spectral norm and extreme eigenvalues.

Theorem 3.6.2. In the setting of Theorem 3.5.3, under the additional assumption on ν from Lemma 3.6.1, we have the following convergences in probability:

$$\begin{aligned} \left\| \frac{1}{\sqrt{d}} \boldsymbol{W}^{(d)} \right\| &\xrightarrow{\mathbb{P}} 2, \\ \lambda_1\left(\frac{1}{\sqrt{d}} \boldsymbol{W}^{(d)} \right) &\xrightarrow{\mathbb{P}} 2, \\ \lambda_d\left(\frac{1}{\sqrt{d}} \boldsymbol{W}^{(d)} \right) &\xrightarrow{\mathbb{P}} -2. \end{aligned}$$

While our argument above only shows that $\lambda_1 = \lambda_1(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)})$ is not too large with high probability, note that it is easy to derive that λ_1 is also not too small with high probability directly from Theorem 3.5.3: we have

$$\mathbb{P}[\lambda_1 \le 2 - \epsilon] \le \mathbb{P}[\mu_d((2-\epsilon, 2)) = 0],$$

which converges to zero since in fact by Theorem 3.5.3 we know that

$$\mu_d((2-\epsilon, 2)) \xrightarrow{\mathbb{P}} \mu_{\mathsf{SC}}((2-\epsilon, 2)) > 0.$$

Remark 3.6.3 (Weaker conditions)**.** In the same spirit as the universality discussed in Exercise 3.8.4 and Section 3.5.3, it is possible to loosen the moment conditions to only require that the fourth moment is finite. This is a sufficient condition for the edge limit and is stronger than the finite second moment needed for Theorem 3.5.3. There is a choice of entrywise distribution ν with finite second moment but infinite fourth moment such that the eigenvalues of $\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}$ converge weakly to the semicircle law while the spectral norm diverges: $\|\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}\| = \omega(1)$ with high probability. Exercise 3.8.5, also Homework 3, Problem 4, asks you to construct such ν.

3.7 Marchenko-Pastur Limit Theorem

We give a brief discussion of the statement and proof by the moment method of the second limit theorem we alluded to before, for the eigenvalues of $\boldsymbol{G}\boldsymbol{G}^\top$ for a rectangular Gaussian random

matrix $\boldsymbol{G}$. (The non-zero singular values of $\boldsymbol{G}$ are the square roots of the non-zero eigenvalues of $\boldsymbol{G}\boldsymbol{G}^\top$.)

Definition 3.7.1. For $c \in (0,\infty)$, define the measure $\mu_{\mathsf{MP}(c)}$ by

$$d\mu_{\mathsf{MP}(c)}(x) = \max\left\{0, 1 - \frac{1}{c}\right\} d\delta_0(x) + \rho_c(x)\,dx,$$

where

$$\rho_c(x) = \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi c x} \text{ for } \lambda_- < x < \lambda_+,$$
$$\lambda_\pm = \lambda_\pm(c) = (1 \pm \sqrt{c})^2 = 1 + c \pm 2\sqrt{c}.$$

The density is defined to be zero outside (λ_-, λ_+).

Theorem 3.7.2. Suppose that $d = d_n \to \infty$ and $m = m_n \to \infty$ with $d_n/m_n \to c \in (0,\infty)$. Suppose $\boldsymbol{G} = \boldsymbol{G}^{(n)} \sim \mathcal{N}(0,1)^{\otimes d \times m}$ and let $\boldsymbol{M} = \boldsymbol{M}^{(n)} := \frac{1}{m}\boldsymbol{G}\boldsymbol{G}^\top \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$. Then, $\mathsf{esd}(\boldsymbol{M}^{(n)}) \xrightarrow{(\mathrm{w},\,\mathcal{M})} \mu_{\mathsf{MP}(c)}$ as $n \to \infty$ for any $\mathcal{M} \in \{\mathbb{E}, \mathbb{P}, L^2\}$.

To interpret the result, consider two cases. First, if $c < 1$, then eventually $d < m$, so $\boldsymbol{M}$ is full rank almost surely and none of its eigenvalues equal zero. That corresponds to the delta mass δ_0 being absent from the expression for $\mu_{\mathsf{MP}(c)}$. In this case, the shape of the limiting distribution is supported on the interval centered at $1 + c$ and of radius $2\sqrt{c}$. Its shape will be a general "blob" for specific finite c, but as $c \to 0$, after rescaling it will resemble a semicircle—recall that a semicircle of width $2\sqrt{d/m}$ is precisely what we observed earlier in this setting in Figure 3.1. At the boundary $c = 1$, the support is $[0,4]$ and there is still no atom at zero. In this case d may be larger or smaller than m, but the fraction of zero eigenvalues is $\max\{0, 1 - m/d\} \to 0$.

Second, if $c > 1$, then eventually $d > m$, so $\boldsymbol{M}$ has rank m almost surely and a fraction $\frac{d-m}{d} = 1 - \frac{m}{d} \to 1 - \frac{1}{c}$ of its eigenvalues equal to zero, which explains the presence and coefficient of δ_0 in $\mu_{\mathsf{MP}(c)}$. The remaining density has total mass $\frac{1}{c}$, and its support is again an interval centered at $c + 1$ (where now c is the "dominant" part) and of radius $2\sqrt{c}$. It may be surprising at first that, as $c \to \infty$, the limiting shape (with a different rescaling) will *also* be a semicircle. But that is just because the spectrum of $\boldsymbol{G}^\top\boldsymbol{G}$ is the same as that of $\boldsymbol{G}\boldsymbol{G}^\top$ up to padding with zeros, so there is a general symmetry between the continuous parts of $\mu_{\mathsf{MP}(c)}$ and $\mu_{\mathsf{MP}(1/c)}$ for all c (which is slightly obscured because the natural normalizations divide $\boldsymbol{G}\boldsymbol{G}^\top$ by m and $\boldsymbol{G}^\top\boldsymbol{G}$ by d).

The proof by moments is a more complicated version of our argument for Theorem 3.5.3. Write $c_n = d_n/m_n$ for the finite-dimensional aspect ratio. We see the similarities and differences upon taking the first few steps:

$$\begin{aligned}
\frac{1}{d}\mathsf{Tr}(\boldsymbol{M}^k) &= \frac{1}{dm^k}\mathsf{Tr}((\boldsymbol{G}\boldsymbol{G}^\top)^k) \\
&= \frac{c_n^k}{d^{k+1}} \sum_{i_1,\dots,i_k \in [d]} (\boldsymbol{G}\boldsymbol{G}^\top)_{i_1 i_2} \cdots (\boldsymbol{G}\boldsymbol{G}^\top)_{i_k i_1} \\
&= \frac{c_n^k}{d^{k+1}} \sum_{\substack{i_1,\dots,i_k \in [d] \\ j_1,\dots,j_k \in [m]}} G_{i_1 j_1} G_{i_2 j_1} G_{i_2 j_2} G_{i_3 j_2} \cdots G_{i_k j_k} G_{i_1 j_k}.
\end{aligned}$$

Thus we will end up counting a kind of bipartite version of the closed walk structures that appeared for Theorem 3.5.3. The parameter c will appear as a weight given to these objects: a graph with a distinct row vertices and b distinct column vertices has $(d)_a(m)_b$ labelings, asymptotic to $d^a m^b = d^{a+b}/c_n^b$ for fixed a, b. Still, it is possible to carry out these calculations and prove Theorem 3.7.2 with essentially the same tools we have seen already. The objects now playing the role of the Catalan numbers—the limiting moments in this calculation—are similar objects called the *Narayana polynomials* (which are now not just numbers but functions because of the dependence on the scalar parameter c).

3.8 Exercises

Exercise 3.8.1 (Explicit formula for the Catalan numbers)**.** This exercise gives two different approaches to proving Proposition 3.5.10, which states that, for every integer $\ell \geq 0$,

$$\mathsf{Cat}_\ell = \frac{1}{\ell+1}\binom{2\ell}{\ell}.$$

1. Recall that Cat_ℓ counts Dyck paths of length 2ℓ, as shown in the proof of Lemma 3.5.7. Prove the formula using this interpretation by counting all paths from $(0,0)$ to $(2\ell,0)$ with steps $(1,1)$ and $(1,-1)$. To count those that go below the x-axis, reflect the initial segment up to the first visit to height -1 across the line $y = -1$.

2. Prove the formula instead using the formal generating function

$$C(z) := \sum_{\ell \geq 0} \mathsf{Cat}_\ell z^\ell.$$

   Use the defining recurrence to derive a quadratic equation for $C(z)$. Solve it, choosing the solution with $C(0) = 1$, and expand using the generalized binomial theorem.

Exercise 3.8.2 (Moments of the semicircle law)**.** Let $X \sim \mu_{\mathsf{SC}}$ and write $\rho(x) = \frac{1}{2\pi}\sqrt{4-x^2}$ for its density on $[-2,2]$. We will prove Proposition 3.5.11 by integration by parts.

1. Show that, for every polynomial f,

$$\mathbb{E}[Xf(X)] = \frac{1}{3}\mathbb{E}[(4-X^2)f'(X)].$$

   (**Hint:** Differentiate $(4-x^2)\rho(x)$ and integrate by parts. Explain why the boundary term vanishes.)

2. Write $m_k := \mathbb{E}[X^k]$ for $k \geq 0$. Show that $m_0 = 1$ and $m_k = 0$ for odd k, and use Part 1 to show that, for every integer $\ell \geq 1$,

$$m_{2\ell} = \frac{2(2\ell-1)}{\ell+1} m_{2\ell-2}.$$

3. Deduce that, for every integer $\ell \geq 0$,

$$m_{2\ell} = \frac{1}{\ell+1}\binom{2\ell}{\ell} = \mathsf{Cat}_\ell,$$

   using Proposition 3.5.10 for the last equality.

Exercise 3.8.3 (Finite moments, convergence in probability)**.** We showed in Lemma 3.5.7 that, if ν has mean 0, variance 1, and all moments finite, then $\boldsymbol{W}^{(d)} \sim \mathsf{Wig}(d, \nu)$ have the random measures $\mu_d := \mathsf{esd}(\widehat{\boldsymbol{W}}^{(d)})$ converging in expected moments to the semicircle distribution μ_{SC}, where $\widehat{\boldsymbol{W}}^{(d)} := \frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}$. That is, for all $k \geq 1$,

$$\lim_{d\to\infty} \mathbb{E}\left[\int x^k \, d\mu_d(x)\right] = \int x^k \, d\mu_{\mathsf{SC}}(x).$$

The goal of this problem is to finish the proof of Theorem 3.5.3 for weak convergence in probability, as outlined in Section 3.5.2. Parts 2 and 3 are also Homework 3, Problem 2.

1. Show that the above convergence, for any given $k \geq 1$, also holds in L^2. That is, show that, for all $k \geq 1$,

   $$\lim_{d\to\infty} \mathsf{Var}\left[\int x^k \, d\mu_d(x)\right] = 0.$$

   Conclude that, for any polynomial $p(x)$,

   $$\int p \, d\mu_d - \mathbb{E}\left[\int p \, d\mu_d\right] \xrightarrow{\mathbb{P}} 0.$$

   (**Hint:** The (random) integral in the variance is equal to $I_k = \frac{1}{d^{k/2+1}} \mathsf{Tr}((\boldsymbol{W}^{(d)})^k)$. Expand $\mathsf{Var}[I_k] = \mathbb{E}[I_k^2] - (\mathbb{E}[I_k])^2$. The convergence of expected moments tells you the asymptotics of the second term. For the first term, adapt our combinatorial argument for the expected moments. When you expand out the square of the trace completely, associate each term with a graph whose edges are given by *two* closed walks (it might help to say that this graph has k "red" edges and k "blue" edges). Show that for odd k this expression is $o(1)$, while for even k its leading order term is the same as that of $(\mathbb{E}[I_k])^2$, so that we still have $\mathsf{Var}[I_k] = o(1)$.)

2. Let $N(\epsilon) := \#\{i : |\lambda_i(\widehat{\boldsymbol{W}}^{(d)})| \geq 2 + \epsilon\}$, the number of eigenvalues of $\widehat{\boldsymbol{W}}^{(d)}$ that lie outside of the support of μ_{SC} by ϵ or more. Show that, for any $\ell \geq 1$,

   $$N(\epsilon) \leq \frac{1}{(2+\epsilon)^{2\ell}} \mathsf{Tr}((\widehat{\boldsymbol{W}}^{(d)})^{2\ell}).$$

   Conclude that

   $$\frac{1}{d} N(\epsilon) \xrightarrow{\mathbb{P}} 0$$

   as $d \to \infty$ for any fixed $\epsilon > 0$, and therefore that also

   $$\frac{1}{d} \sum_{i:|\lambda_i(\widehat{\boldsymbol{W}}^{(d)})|>2+\epsilon} |f(\lambda_i(\widehat{\boldsymbol{W}}^{(d)}))| \xrightarrow{\mathbb{P}} 0$$

   for any fixed bounded function f and $\epsilon > 0$.

   (**Hint:** Use Markov's inequality on $\mathbb{P}[N(\epsilon) \geq \delta d]$ for any constant $\delta > 0$, then use the convergence of expected moments and a suitable bound on the Catalan numbers. First take $d \to \infty$ with ℓ fixed, then take $\ell \to \infty$.)

3. Adapt the argument from Part 2 to show that the latter conclusion also holds for polynomials (even though they are not bounded): show that for any polynomial p,

   $$\frac{1}{d} \sum_{i:|\lambda_i(\widehat{\boldsymbol{W}}^{(d)})|>2+\epsilon} |p(\lambda_i(\widehat{\boldsymbol{W}}^{(d)}))| \xrightarrow{\mathbb{P}} 0.$$

(**Hint:** Reduce to the case of monomials $p(x) = x^k$. Then, bound $\frac{1}{d}\sum_{i:|\lambda_i|>2+\epsilon} |\lambda_i|^k \le \frac{1}{d(2+\epsilon)^{2\ell-k}} \sum_{i=1}^d \lambda_i^{2\ell}$ for $2\ell > k$. Apply Markov's inequality again and take the limits in the same order as in Part 2.)

4. Following the sketch from Section 3.5.2, put together Parts 1, 2, and 3 to prove a version of the semicircle limit theorem with weak convergence in probability: show that, for any bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,
$$\int f \, d\mu_d \xrightarrow{\mathbb{P}} \int f \, d\mu_{\mathsf{SC}}.$$
You may use the Weierstrass approximation theorem (Theorem 3.5.14).

Exercise 3.8.4 (Removing finite moments assumption)**.** The above Exercise 3.8.3 showed that, if ν has mean zero, variance 1, and all moments finite, then $\boldsymbol{W}^{(d)} \sim \mathsf{Wig}(d, \nu)$ have $\mathsf{esd}(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)})$ converging weakly in probability to μ_{SC}. That is, for any $f : \mathbb{R} \to \mathbb{R}$ smooth and of compact support, $\frac{1}{d}\sum_{i=1}^d f(\lambda_i(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)})) \to \int f \, d\mu_{\mathsf{SC}}$ in probability (we considered more general f, but only worry about convergence in probability for these f for this problem). In this problem, you will probe what conditions on ν are really necessary for what kinds of limit theorems. You should consult Homework 2, Problem 3, on the Hoffman–Wielandt inequality before continuing with this problem. Part 2 is also Homework 3, Problem 3.

1. Let f be a smooth and compactly supported function. Show that there is a constant $K = K(f)$ depending only on f such that, for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$,
$$\left| \frac{1}{d}\sum_{i=1}^d f(\lambda_i(\boldsymbol{A})) - \frac{1}{d}\sum_{i=1}^d f(\lambda_i(\boldsymbol{B})) \right| \le \frac{K}{\sqrt{d}} \|\boldsymbol{A} - \boldsymbol{B}\|_F.$$

2. Prove that Wigner's semicircle limit theorem (convergence in probability of averages of smooth and compactly supported functions, as stated above) holds assuming only that ν has mean 0 and variance 1.

   (**Hint:** For a sufficiently large constant C, set $T_{ij} = W_{ij}\mathbb{1}\{|W_{ij}| \le C\}$ for $i < j$ and replace W_{ij} by the centered, normalized truncation $(T_{ij} - \mathbb{E}[T_{ij}])/\sqrt{\mathsf{Var}[T_{ij}]}$. Keep the matrix symmetric and its diagonal zero. Apply the all-moments theorem to this matrix and use Part 1 to compare its spectral averages with those of the original matrix. First take $d \to \infty$ with C fixed, then take $C \to \infty$.)

3. Extend Part 2 to every bounded continuous f, completing the proof of Theorem 3.5.3.

   (**Hint:** First use uniform approximation to pass from smooth compactly supported functions to continuous compactly supported functions. Then use a continuous cutoff and the bound $\mathbb{E}[\frac{1}{d}\sum_i \lambda_i^2] = (d-1)/d \le 1$ to control the fraction of eigenvalues outside a large compact interval.)

Exercise 3.8.5 (Limitations of edge limit theorem)**.** Find a choice of ν that has mean 0 and variance 1 but such that, for $\boldsymbol{W}^{(d)} \sim \mathsf{Wig}(d, \nu)$, we have $\lim_{d\to\infty} \mathbb{P}[\|\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}\| \ge C_d] = 1$ for some diverging sequence $C_d \to \infty$. Consequently, the Wigner edge or norm limit theorem (Theorem 3.6.2) does require further moment assumptions on ν.

(**HINT:** Prove and use that $\|\boldsymbol{W}\| \geq \max_{i,j\in[d]} |W_{ij}|$. As we have mentioned, it is known that the Wigner edge limit theorem *does* hold provided that the fourth moment of ν is finite, so your choice must not have that property.)

4 | Free Probability

Our next goal is to develop a systematic way to answer the following kinds of questions:

1. Suppose that $\boldsymbol{W} \sim \mathsf{Wig}(d, \mathcal{N}(0,1))$ and $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d \times m}$ are drawn independently. What do the eigenvalues of the random matrix
$$\frac{1}{\sqrt{d}}\boldsymbol{W} + \frac{1}{m}\boldsymbol{G}\boldsymbol{G}^\top \tag{4.0.1}$$
look like?

2. Suppose $\boldsymbol{G}, \boldsymbol{H} \sim \mathcal{N}(0,1)^{\otimes d \times m}$ are independent. What do the singular values of the (square but asymmetric) random matrix
$$\boldsymbol{G}\boldsymbol{H}^\top \tag{4.0.2}$$
look like?

The results in the previous chapter characterize some related eigenvalue and singular value distributions, namely the eigenvalue distribution of the individual summands in (4.0.1) and the individual factors in (4.0.2). So, we are asking: how does adding or multiplying independent random matrices affect the spectrum? In fact, while these specific example may look artificial, we will see that later in this chapter that similar random matrices arise in several interesting applications.

Let us first note that the situation is not entirely hopeless if we are guided by scalar probability. Indeed, if $X \sim \mu$ and $Y \sim \nu$ are independent scalar random varaibles, then

$$\mathsf{Law}(X+Y) = \mu * \nu$$

is given by the *convolution*. The convolution can be characterized in various ways:

- **Densities:** If μ and ν are absolutely continuous with respect to Lebesgue measure and have respective densities ρ_μ and ρ_ν, then the density of the convolution is
$$\rho_{\mu*\nu}(t) = \int_{-\infty}^{\infty} \rho_\mu(s)\rho_\nu(t-s)\,ds.$$

- **Generating functions / transforms:** Write $\phi_\mu(t) := \mathbb{E}_{X\sim\mu} \exp(tX)$ and $\psi_\mu(t) := \log \phi_\mu(t)$ for the *moment* and *cumulant generating functions* of a probability measure μ. Then,
$$\phi_{\mu*\nu} = \phi_\mu \phi_\nu, \tag{4.0.3}$$
$$\psi_{\mu*\nu} = \psi_\mu + \psi_\nu. \tag{4.0.4}$$
The first identity is also satisfied by the *characteristic function*, which is just the moment generating function evaluated at imaginary arguments.

- **Moments:** Suppose that for each $k \geq 1$ we have

$$\mathbb{E} X^k = m_k, \tag{4.0.5}$$
$$\mathbb{E} Y^k = n_k. \tag{4.0.6}$$

A simple calculation with the binomial theorem gives that the moments of the convolution are

$$\mathop{\mathbb{E}}_{S \sim \mu * \nu} S^k = \mathop{\mathbb{E}}_{\substack{X \sim \mu \\ Y \sim \nu}} (X+Y)^k = \sum_{a=0}^{k} \binom{k}{a} m_a n_{k-a}. \tag{4.0.7}$$

Provided a condition like that in Theorem 3.4.1 holds, these moments determine the convolution. Further, this method generalizes very simply to the product of X and Y, as we also have

$$\mathop{\mathbb{E}}_{\substack{X \sim \mu \\ Y \sim \nu}} (XY)^k = m_k n_k. \tag{4.0.8}$$

As we have seen, the moment or combinatorial method is one from scalar probability that extends especially nicely to random matrix theory. So, we will now explore how to adapt the third characterization above of the scalar convolution to certain sums and products of random matrices.

4.1 Computing Joint Tracial Moments

As we will be working with moment calculations very extensively, a bit of extra notation will be useful going forward.

Definition 4.1.1. For $\boldsymbol{X} \in \mathbb{R}^{d \times d}_{\mathsf{sym}}$, write $\mathsf{tr}(\boldsymbol{X}) := \frac{1}{d} \mathsf{Tr}(\boldsymbol{X})$.

The following will also turn out to be convenient.

Definition 4.1.2. For μ a measure, we write $c \cdot \mu := \mathsf{Law}(cX)$ where $X \sim \mu$.

As we have seen, the combinatorial method analyzes the distributions of eigenvalues and singular values by looking at traces of powers of matrices (or expectations of those in the case of random matrices). We refer to these quantities as the *tracial moments* of a matrix. We have seen in Chapter 3 that a large part of the work of determining the limiting eigenvalue distribution of a sequence of random matrices $\boldsymbol{M} = \boldsymbol{M}^{(d)}$ involves determining these tracial moments of $\boldsymbol{M}$,

$$\lim_{d \to \infty} \mathbb{E}\, \mathsf{tr}(\boldsymbol{M}^k).$$

In either of the two examples above, $\boldsymbol{M}$ is itself a polynomial of two random matrices (either the sum or the product, in our case). So, in either case, upon expanding the power above, we will have a linear combination of *joint tracial moments*, the expected values of traces of products of two random matrices. This is analogous to a mixed moment $\mathbb{E} X^a Y^b$ of two scalar random variables X, Y, except that since matrices do not commute in general, we can have expressions such as $\mathbb{E}\, \mathsf{tr}\, \boldsymbol{X}\boldsymbol{Y}\boldsymbol{X}^2\boldsymbol{Y}$ that cannot in general be reduced any further. Let us see how such expressions arise in the case of a sum of two matrices, and work towards a sufficient condition under which they may be analyzed.

4.1.1 First Steps with Word Expansion

With our normalized trace notation notation, convergence of expected moments of two sequences of random matrices $\boldsymbol{X} = \boldsymbol{X}^{(d)}, \boldsymbol{Y} = \boldsymbol{Y}^{(d)} \in \mathbb{R}_{\mathsf{sym}}^{d\times d}$ can be written as

$$\lim_{d\to\infty} \mathbb{E}\,\mathsf{tr}\,\boldsymbol{X}^k = m_k, \tag{4.1.1}$$

$$\lim_{d\to\infty} \mathbb{E}\,\mathsf{tr}\,\boldsymbol{Y}^k = n_k \tag{4.1.2}$$

for some associated moment sequences m_k, n_k. We will now explore under what circumstances we can say anything about the limits

$$\lim_{d\to\infty} \mathbb{E}\,\mathsf{tr}(\boldsymbol{X}+\boldsymbol{Y})^k =: s_k = ???$$

in terms of m_k, n_k. (It is not even obvious under what circumstances the limits defining s_k exist, but let us suppose they do for the moment.)

Following the scalar case, we will assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent (for each d). We will also assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are symmetric-orthogonally invariant, because, as we will see in a moment, the following property will be useful in our calculations.

Proposition 4.1.3. Suppose $\boldsymbol{X} \in \mathbb{R}_{\mathsf{sym}}^{d\times d}$ is a symmetric–orthogonally invariant random matrix. Then,

$$\mathbb{E}\boldsymbol{X}^k = \left(\mathbb{E}\,\mathsf{tr}\,\boldsymbol{X}^k\right)\boldsymbol{I}_d.$$

Proof. The matrix $\boldsymbol{X}^k$ is also symmetric–orthogonally invariant, so without loss of generality we may assume $k = 1$. We have that $\mathbb{E}\boldsymbol{X} = \mathbb{E}\boldsymbol{Q}\boldsymbol{X}\boldsymbol{Q}^\top = \boldsymbol{Q}(\mathbb{E}\boldsymbol{X})\boldsymbol{Q}^\top$ for any $\boldsymbol{Q} \in \mathcal{O}(d)$, and the only matrices that have this property are multiples of the identity. So, we must have $\mathbb{E}\boldsymbol{X} = c\boldsymbol{I}_d$. Taking normalized traces then gives $c = \mathbb{E}\,\mathsf{tr}\,\boldsymbol{X}$. □

To compute the s_k, we will try to mimic the binomial expansion leading to the formula for scalar random variables in (4.0.7). We will run into the issue that, since $\boldsymbol{X}$ and $\boldsymbol{Y}$ are matrices and not numbers now, they do not necessarily commute, and so we cannot always group all $\boldsymbol{X}$ and $\boldsymbol{Y}$ together. We introduce a few definitions to work with the resulting quantities.

Definition 4.1.4 (Word)**.** A *word* in matrices $\boldsymbol{X}, \boldsymbol{Y}$ of length k is $w \in \{\boldsymbol{X}, \boldsymbol{Y}\}^k$. We may also identify this combinatorial string w with a polynomial $w(\boldsymbol{X}, \boldsymbol{Y})$ in two non-commuting formal variables, which contains just one monomial.

We will freely move back and forth between the "string interpretation" and the "polynomial interpretation" of what a word is. For instance, words include $\boldsymbol{X}\boldsymbol{X}\boldsymbol{X}\boldsymbol{X} = \boldsymbol{X}^4$, $\boldsymbol{X}\boldsymbol{X}\boldsymbol{Y}\boldsymbol{Y} = \boldsymbol{X}^2\boldsymbol{Y}^2$, and $\boldsymbol{X}\boldsymbol{Y}\boldsymbol{X}\boldsymbol{Y}$. Our point about non-commutativity above is just saying that the last two examples are different words.

Definition 4.1.5 (Limiting joint tracial moments)**.** For $w \in \{\boldsymbol{X}, \boldsymbol{Y}\}^k$ a word, we define

$$t(w) := \lim_{d\to\infty} \mathbb{E}\,\mathsf{tr}\,w(\boldsymbol{X}, \boldsymbol{Y}).$$

Note that each such quantity depends on the distribution of the ambient sequence of pairs

$\boldsymbol{X} = \boldsymbol{X}^{(d)}, \boldsymbol{Y} = \boldsymbol{Y}^{(d)}$, but for the sake of brevity we do not indicate this in the notation.

These are precisely the joint tracial moments we alluded to before. Just as the m_k are the moments of the limiting spectral distribution of $\boldsymbol{X}$ (and similarly for the n_k and $\boldsymbol{Y}$), the $t(w)$ should be viewed as the moments of the limiting "joint spectral distribution" of the distribution of the pair $(\boldsymbol{X}, \boldsymbol{Y})$.

Remark 4.1.6. One may make this idea more precise by appealing to invariant theory. It is an algebraic fact that the polynomials $p(\boldsymbol{X})$ that are orthogonally invariant, i.e., satisfying $p(\boldsymbol{X}) = p(\boldsymbol{Q}\boldsymbol{X}\boldsymbol{Q}^\top)$ for all $\boldsymbol{Q} \in \mathcal{O}(d)$, are generated algebraically by the polynomials $\mathrm{tr}(\boldsymbol{X}^k)$ over $k \geq 1$. Likewise, the polynomials $p(\boldsymbol{X}, \boldsymbol{Y})$ with $p(\boldsymbol{X}, \boldsymbol{Y}) = p(\boldsymbol{Q}\boldsymbol{X}\boldsymbol{Q}^\top, \boldsymbol{Q}\boldsymbol{Y}\boldsymbol{Q}^\top)$ for all $\boldsymbol{Q} \in \mathcal{O}(d)$ are generated algebraically by the polynomials $\mathrm{tr}(w(\boldsymbol{X}, \boldsymbol{Y}))$ over all words w. Thus the "joint spectral distribution" above may be viewed as the collection of values of all invariant polynomials of the pair $(\boldsymbol{X}, \boldsymbol{Y})$.

Again, let us ignore the issue of whether this limit exists for now; we will see that we eventually end up working under a condition that ensures that it always does. The following result is the outcome of a binomial-type expansion of $(\boldsymbol{X} + \boldsymbol{Y})^k$, except where we are not allowed to commute $\boldsymbol{X}$ and $\boldsymbol{Y}$:

Proposition 4.1.7 (Word expansion)**.** Provided all limits involved exist, we have

$$s_k = \lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}(\boldsymbol{X} + \boldsymbol{Y})^k = \sum_{w \in \{\boldsymbol{X}, \boldsymbol{Y}\}^k} t(w).$$

For a few special kinds of words, we can evaluate $t(w)$ just using the observations above. First, note that when w only involves $\boldsymbol{X}$ or only involves $\boldsymbol{Y}$, then, by definition, the moment convergence assumptions (4.1.1) and (4.1.2) respectively determine $t(w)$:

$$t(\boldsymbol{X}^k) = t(\boldsymbol{X} \cdots \boldsymbol{X}) = m_k,$$
$$t(\boldsymbol{Y}^k) = t(\boldsymbol{Y} \cdots \boldsymbol{Y}) = n_k.$$

Second, if w involves both $\boldsymbol{X}$ and $\boldsymbol{Y}$ but in two separate factors, then we may use independence and orthogonal invariance to determine $t(w)$:

$$\begin{aligned}
t(\boldsymbol{X}^a\boldsymbol{Y}^b) &= \lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}\,\boldsymbol{X}^a\boldsymbol{Y}^b \\
&= \lim_{d\to\infty} \mathrm{tr}(\mathbb{E}\boldsymbol{X}^a)(\mathbb{E}\boldsymbol{Y}^b) && \text{(independence)} \\
&= \lim_{d\to\infty} \mathrm{tr}\left((\mathbb{E}\,\mathrm{tr}\,\boldsymbol{X}^a)\boldsymbol{I}_d \cdot (\mathbb{E}\,\mathrm{tr}\,\boldsymbol{Y}^b)\boldsymbol{I}_d\right) && \text{(Proposition 4.1.3)} \\
&= \lim_{d\to\infty} (\mathbb{E}\,\mathrm{tr}\,\boldsymbol{X}^a)(\mathbb{E}\,\mathrm{tr}\,\boldsymbol{Y}^b)\,\mathrm{tr}\,\boldsymbol{I}_d \\
&= m_a n_b.
\end{aligned}$$

Lastly, $t(w)$ inherits the following property from the corresponding property of the trace. For two words w_1, w_2, we write $w_1 w_2$ for their concatenation, which is compatible with our polynomial interpretation since the polynomial associated to $w_1 w_2$ is $w_1(\boldsymbol{X}, \boldsymbol{Y})w_2(\boldsymbol{X}, \boldsymbol{Y})$.

Proposition 4.1.8 (Cyclic property)**.** For $w_1 \in \{\boldsymbol{X}, \boldsymbol{Y}\}^a$ and $w_2 \in \{\boldsymbol{X}, \boldsymbol{Y}\}^b$, $t(w_1 w_2) = t(w_2 w_1)$.

For instance, this lets us calculate

$$t(\boldsymbol{X}^a \boldsymbol{Y}^b \boldsymbol{X}^c) = t(\boldsymbol{X}^{a+c} \boldsymbol{Y}^b) = m_{a+c} n_b,$$

and similarly for words with the roles of $\boldsymbol{X}$ and $\boldsymbol{Y}$ reversed.

Let us summarize what we have learned above in a formal statement. The words for which we can so far compute $t(w)$ explicitly in terms of the m_a and n_b assuming only that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent and orthogonally invariant are of the following kind.

Definition 4.1.9 (Tangled and simple words)**.** Call a word $w \in \{\boldsymbol{X}, \boldsymbol{Y}\}^k$ *simple* if by a cyclic permutation (writing $w = w_1 w_2$ and replacing w with $w_2 w_1$) w can be brought to the form $\boldsymbol{X}^a \boldsymbol{Y}^b$ (where possibly $a = 0$ or $b = 0$). Call a word *tangled* if it is not simple.

Lemma 4.1.10 (Simple word values)**.** Let w be a simple word with a occurrences of $\boldsymbol{X}$ and b occurrences of $\boldsymbol{Y}$, and let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be independent, symmetric-orthogonally invariant, and satisfy (4.1.1) and (4.1.2). Then, $t(w) = m_a n_b$.

We may in fact compute the first few s_k using just these quite basic observations. The key observation, which you may verify by hand, is that all words of length at most 3 are simple.

Corollary 4.1.11 (Summation rules for $k = 1, 2, 3$)**.** If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, symmetric-orthogonally invariant, and satisfy (4.1.1) and (4.1.2), and $s_k := \lim_{d \to \infty} \mathbb{E} \operatorname{tr}(\boldsymbol{X} + \boldsymbol{Y})^k$, then

$$\begin{aligned} s_1 &= m_1 + n_1, \\ s_2 &= m_2 + 2m_1 n_1 + n_2, \\ s_3 &= m_3 + 3m_2 n_1 + 3m_1 n_2 + n_3. \end{aligned}$$

Proof. We use the word expansion of Proposition 4.1.7, the evaluation of $t(w)$ for simple words of Lemma 4.1.10, and the fact that all words of length 1, 2, and 3 are simple. As an example, let us write out the derivation in detail for $k = 2$:

$$\begin{aligned} s_2 &= t(\boldsymbol{X}\boldsymbol{X}) + t(\boldsymbol{X}\boldsymbol{Y}) + t(\boldsymbol{Y}\boldsymbol{X}) + t(\boldsymbol{Y}\boldsymbol{Y}) && \text{(Proposition 4.1.7)} \\ &= m_2 + m_1 n_1 + m_1 n_1 + n_2 && \text{(Lemma 4.1.10)} \\ &= m_2 + 2m_1 n_1 + n_2, \end{aligned}$$

giving the result. □

We remark that these are identical to the formulas in (4.0.7) from the case of scalar random variables; that is, for these low-degree moments, we do not yet see the (generally quite large, as we will see) effect of the non-commutativity of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

As another remark, it is possible to relax these conditions slightly and only assume that *one* of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is orthogonally invariant.

Corollary 4.1.12. The result of Corollary 4.1.11 holds if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, satisfy (4.1.1) and (4.1.2), and $\boldsymbol{Y}$ is symmetric-orthogonally invariant.

Proof. Let us introduce an additional random matrix $\boldsymbol{Q} = \boldsymbol{Q}^{(d)} \sim \mathsf{Haar}(\mathcal{O}(d))$. Then, since $\mathsf{tr}(\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^\top) = \mathsf{tr}(\boldsymbol{A})$ and $\boldsymbol{Q}\boldsymbol{A}^k\boldsymbol{Q}^\top = (\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^\top)^k$ for any $\boldsymbol{A} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{Y}} \mathsf{tr}(\boldsymbol{X}+\boldsymbol{Y})^k &= \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Q}} \mathsf{tr}\,\boldsymbol{Q}(\boldsymbol{X}+\boldsymbol{Y})^k\boldsymbol{Q}^\top \\ &= \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Q}} \mathsf{tr}(\boldsymbol{Q}\boldsymbol{X}\boldsymbol{Q}^\top + \boldsymbol{Q}\boldsymbol{Y}\boldsymbol{Q}^\top)^k\end{aligned}$$

and now, since $\mathsf{Law}(\boldsymbol{Q}\boldsymbol{Y}\boldsymbol{Q}^\top) = \mathsf{Law}(\boldsymbol{Y})$ for any fixed $\boldsymbol{Q}$ by the invariance of $\boldsymbol{Y}$,

$$= \mathop{\mathbb{E}}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Q}} \mathsf{tr}(\boldsymbol{Q}\boldsymbol{X}\boldsymbol{Q}^\top + \boldsymbol{Y})^k.$$

Letting $\widetilde{\boldsymbol{X}}^{(d)} := \boldsymbol{Q}\boldsymbol{X}^{(d)}\boldsymbol{Q}^\top$, the sequence of pairs $(\widetilde{\boldsymbol{X}}^{(d)}, \boldsymbol{Y})$ still satisfy (4.1.1), and are symmetric-orthogonally invariant and independent, so the result follows by applying Corollary 4.1.11 to these pairs. □

Unfortunately, if we try to continue the same approach to compute s_4, we get stuck because we encounter our first tangled words: you may check that

$$\begin{aligned}s_4 &= m_4 + 4m_3n_1 + 4m_2n_2 + 4m_1n_3 + n_4 + t(\boldsymbol{X}\boldsymbol{Y}\boldsymbol{X}\boldsymbol{Y}) + t(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{Y}\boldsymbol{X}) \\ &= m_4 + 4m_3n_1 + 4m_2n_2 + 4m_1n_3 + n_4 + 2t(\boldsymbol{X}\boldsymbol{Y}\boldsymbol{X}\boldsymbol{Y}),\end{aligned}$$

where we have reduced slightly by using the cyclic property of $t(\cdot)$. Note that these two tangled words precisely correspond to the two "missing" terms of m_2n_2 accounting for the difference between the above and the binomial formula (4.0.7).

4.1.2 Gaussian Orthogonal Ensemble Calculations

To understand more about how tangled words behave, let us work out some cases for a simple example, where we can also compute the values of s_k exactly using other special properties of the random matrices involved.

Definition 4.1.13. Consider $\boldsymbol{W} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ with independent entries on and above the diagonal, distributed as $W_{ij} = W_{ji} \sim \mathcal{N}(0,\sigma^2)$ for $i < j$ and $W_{ii} \sim \mathcal{N}(0, 2\sigma^2)$. We write $\mathsf{GOE}(d,\sigma^2)$ for the law of such a random matrix, which we call the *Gaussian orthogonal ensemble (GOE)*.

The GOE is one of the nicest and most widely studied random matrices, because it enjoys several useful properties beyond just the Gaussianity of its entries (that distinguish it from, for instance, the somewhat but less nice Gaussian random matrix $\mathsf{Wig}(d, \mathcal{N}(0,\sigma^2))$ that has the same off-diagonal entries but equals zero on the diagonal). The main special property is the following.

Proposition 4.1.14. $\boldsymbol{W} \sim \mathsf{GOE}(d,\sigma^2)$ is a symmetric–orthogonally invariant random matrix; that is, $\mathsf{Law}(\boldsymbol{Q}\boldsymbol{W}\boldsymbol{Q}^\top) = \mathsf{Law}(\boldsymbol{W})$ for all $\boldsymbol{Q} \in \mathcal{O}(d)$.

The following is also inherited from a simple variant of Theorem 3.5.3:

Corollary 4.1.15. If $\boldsymbol{X}^{(d)} \sim \mathsf{GOE}(d, \frac{1}{d})$, then $\mathsf{esd}(\boldsymbol{X}^{(d)}) \xrightarrow{(\mathbb{E}\text{ mom.})} \mu_{\mathsf{SC}}$ and $\mathsf{esd}(\boldsymbol{X}^{(d)}) \xrightarrow{(\mathrm{w},\,\mathbb{P})} \mu_{\mathsf{SC}}$.

Finally, for our purposes, the following property will be very useful: if $\boldsymbol{X}^{(d)}, \boldsymbol{Y}^{(d)} \sim \mathsf{GOE}(d, \frac{1}{d})$, then

$$\mathsf{Law}(\boldsymbol{X}^{(d)} + \boldsymbol{Y}^{(d)}) = \mathsf{GOE}\left(d, \frac{2}{d}\right) = \mathsf{Law}(\sqrt{2} \cdot \boldsymbol{X}^{(d)}). \tag{4.1.3}$$

With these ideas in hand, consider our question of computing

$$\lim_{d\to\infty} \mathbb{E}\,\mathsf{tr}(\boldsymbol{X} + \boldsymbol{Y})^k =: s_k$$

in this special case where $\boldsymbol{X} = \boldsymbol{X}^{(d)}, \boldsymbol{Y} = \boldsymbol{Y}^{(d)} \sim \mathsf{GOE}(d, \frac{1}{d})$ are drawn independently. (Note that this satisfies the assumptions we worked under previously: $\boldsymbol{X}$ and $\boldsymbol{Y}$ are indeed independent and both are symmetric-orthogonally invariant.) By Lemma 3.5.7 from the previous chapter, we know that each sequence has converging moments, which in this case are in fact the same, given by

$$\begin{aligned}
\lim_{d\to\infty} \mathbb{E}\,\mathsf{tr}\,\boldsymbol{X}^k = m_k &= \left\{ \begin{array}{ll} 0 & \text{if } k \text{ is odd,} \\ \mathsf{Cat}_k & \text{if } k \text{ is even} \end{array} \right\}, \\
\lim_{d\to\infty} \mathbb{E}\,\mathsf{tr}\,\boldsymbol{Y}^k &= n_k = m_k.
\end{aligned}$$

And, the relation (4.1.3) also gives the explicit answer

$$s_k = \lim_{d\to\infty} \mathbb{E}\,\mathsf{tr}(\boldsymbol{X} + \boldsymbol{Y})^k = \lim_{d\to\infty} \mathbb{E}\,\mathsf{tr}(\sqrt{2}\boldsymbol{X})^k = 2^{k/2} m_k \tag{4.1.4}$$

for the limiting moments of the sum.

In particular, we may write out the first few values:

$$\begin{aligned}
m_1 = n_1 &= 0, \\
s_1 = 2^{1/2} m_1 &= 0, \\
m_2 = n_2 &= 1, \\
s_2 = 2^{2/2} m_2 &= 2, \\
m_3 = n_3 &= 0, \\
s_3 = s^{3/2} m_3 &= 0, \\
m_4 = n_4 &= 2, \\
s_4 = s^{2/2} m_4 &= 8.
\end{aligned}$$

As a sanity check, we may verify that the formulas derived in the previous section for s_1, s_2, and s_3 indeed hold:

$$\begin{aligned}
0 &= s_1 \\
&= m_1 + n_1 \\
&= 0 + 0, \\
2 &= s_2 \\
&= m_2 + 2m_1 n_1 + n_2 \\
&= 1 + 2 \cdot 0 \cdot 0 + 1, \\
0 &= s_3 \\
&= m_3 + 3m_2 n_1 + 3m_1 n_2 + n_3 \\
&= 0 + 3 \cdot 0 \cdot 0 + 3 \cdot 0 \cdot 0 + 0.
\end{aligned}$$

Similarly, we may deduce the value of $t(\boldsymbol{XYXY})$ in this case. In fact, it is possible to do so in two ways, using the above type of expansion or using a more direct combinatorial analysis of the trace defining this value. Let us show both arguments.

Proposition 4.1.16. If $\boldsymbol{X}, \boldsymbol{Y} \sim \mathsf{GOE}(d, \frac{1}{d})$ are independent, then

$$t(\boldsymbol{XYXY}) = 0.$$

Proof 1. We compute using the previous observations,

$$\begin{aligned} 8 &= s_4 \\ &= m_4 + 4m_3n_1 + 4m_2n_2 + 4m_1n_3 + n_4 + 2t(\boldsymbol{XYXY}) \\ &= 2 + 4 \cdot 0 \cdot 0 + 4 \cdot 1 \cdot 1 + 4 \cdot 0 \cdot 0 + 2 + 2t(\boldsymbol{XYXY}) \\ &= 8 + 2t(\boldsymbol{XYXY}), \end{aligned}$$

and rearranging gives the result. □

Proof 2. Let $\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Y}} \sim \mathsf{GOE}(d, 1)$, so that we can take $\boldsymbol{X} = \frac{1}{\sqrt{d}}\widetilde{\boldsymbol{X}}$ and $\boldsymbol{Y} = \frac{1}{\sqrt{d}}\widetilde{\boldsymbol{Y}}$. Expanding the definition of the normalized trace, we find

$$\begin{aligned} \mathbb{E}\,\mathrm{tr}\,\boldsymbol{XYXY} &= \frac{1}{d^3} \sum_{i,j,k,\ell=1}^{d} \mathbb{E}\,\widetilde{X}_{ij}\widetilde{Y}_{jk}\widetilde{X}_{k\ell}\widetilde{Y}_{\ell i} \\ &= \frac{1}{d^3} \sum_{i,j,k,\ell=1}^{d} \mathbb{E}[\widetilde{X}_{ij}\widetilde{X}_{k\ell}] \cdot \mathbb{E}[\widetilde{Y}_{jk}\widetilde{Y}_{\ell i}] \end{aligned}$$

where the inner expectation is non-zero only when $i = k$ and $j = \ell$. There are only $O(d^2)$ such terms, so in total we have

$$\begin{aligned} &= \frac{1}{d^3} \cdot O(d^2) \\ &= O\left(\frac{1}{d}\right), \end{aligned}$$

giving the result. □

We might hope that, at least in this simple example and perhaps in other cases, *all* tangled words w have $t(w) = 0$. Then, in the word expansion $s_k = \sum_{w \in \{\boldsymbol{X}, \boldsymbol{Y}\}^k} t(w)$ we could restrict the summation to only simple words, whose values are determined by Lemma 4.1.10. Unfortunately, the next example shows that that is too much to hope for even for independent GOE matrices.

Proposition 4.1.17. If $\boldsymbol{X}, \boldsymbol{Y} \sim \mathsf{GOE}(d, \frac{1}{d})$ are independent, then

$$t(\boldsymbol{X}^2\boldsymbol{Y}^2\boldsymbol{X}^2\boldsymbol{Y}^2) = 3 > 0.$$

Proof. Let us just give the proof that the value is positive; we will discuss the idea of how to calculate the exact value below. Calculating as before,

$$\begin{aligned}\mathbb{E}\,\mathsf{tr}\,\boldsymbol{X}\boldsymbol{X}\boldsymbol{Y}\boldsymbol{Y}\boldsymbol{X}\boldsymbol{X}\boldsymbol{Y}\boldsymbol{Y} &= \frac{1}{d^5}\sum_{i,j,k,\ell,m,n,p,q=1}^{d} \mathbb{E}\,\widetilde{X}_{ij}\widetilde{X}_{jk}\widetilde{Y}_{k\ell}\widetilde{Y}_{\ell m}\widetilde{X}_{mn}\widetilde{X}_{np}\widetilde{Y}_{pq}\widetilde{Y}_{qi} \\ &= \frac{1}{d^5}\sum_{i,j,k,\ell,m,n,p,q=1}^{d} \mathbb{E}[\widetilde{X}_{ij}\widetilde{X}_{jk}\widetilde{X}_{mn}\widetilde{X}_{np}] \cdot \mathbb{E}[\widetilde{Y}_{k\ell}\widetilde{Y}_{\ell m}\widetilde{Y}_{pq}\widetilde{Y}_{qi}]\end{aligned}$$

Here, all expectations are certain Gaussian moments and thus are non-negative for every term. And, there is a positive contribution from the terms where $i = k = m = p$, of which there are $\Omega(d^5)$. So, we get

$$\begin{aligned}&\geq \frac{1}{d^5} \cdot \Omega(d^5) \\ &= \Omega(1),\end{aligned}$$

giving the result. □

Another way to intuit that this trace should be biased towards being positive is to note that we may rewrite

$$\mathsf{tr}\,\boldsymbol{X}^2\boldsymbol{Y}^2\boldsymbol{X}^2\boldsymbol{Y}^2 = \mathsf{tr}\,\boldsymbol{X}\boldsymbol{Y}^2\boldsymbol{X}^2\boldsymbol{Y}^2\boldsymbol{X} = \mathsf{tr}((\boldsymbol{X}\boldsymbol{Y}^2\boldsymbol{X})^2),$$

the trace of a positive semidefinite matrix.

Remark 4.1.18 (Graphical method for joint tracial moments)**.** Using this direct combinatorial method, you may also convince yourself that Propositions 4.1.16 and 4.1.17 also hold for $\boldsymbol{X}, \boldsymbol{Y} \sim \mathsf{Wig}(d, \frac{1}{\sqrt{d}} \cdot \nu)$ for any ν with mean zero and variance 1.

In general, the proof technique from Lemma 3.5.7 using closed walk graphs to "keep the books" of the expansion of a trace can be generalized to compute $t(w)$ for any word w when $\boldsymbol{X}, \boldsymbol{Y}$ are independent Wigner matrices with the above scaling. You may check that the right thing to do is to consider closed walk graphs where the edges are ordered and have two "colors" corresponding to the entries of $\boldsymbol{X}$ (say blue) and the entries of $\boldsymbol{Y}$ (say red). The word w specifies the order in which the edges are colored as the closed walk is traversed; for instance, if we consider the word $w = \boldsymbol{X}^2\boldsymbol{Y}^2\boldsymbol{X}^2\boldsymbol{Y}^2$ studied above, we have the edge sequence colored blue, blue, red, red, blue, blue, red, red. Only the closed walk graphs where each colored edge occurs with multiplicity at least 2 (in either direction) make a non-zero contribution, and only those graphs with the maximum possible number of vertices subject to that constraint make a leading order contribution. As in Lemma 3.5.7, these will again be the doubled trees, but now we must count doubled trees with a suitable edge coloring that respects the order of colors prescribed by the word w. See Figure 4.1 for an example corresponding to the proof of Proposition 4.1.17.

Thus we seem to have reached a difficult position: in the GOE case, *some* tangled words are negligible, while others are not. To make progress, we may gather from the above observation on positive semidefiniteness that the issue making $t(\boldsymbol{X}^2\boldsymbol{Y}^2\boldsymbol{X}^2\boldsymbol{Y}^2) > 0$ is that the polynomials $\boldsymbol{X}^2$ and $\boldsymbol{Y}^2$ are "uncentered," in particular having a non-zero and indeed positive semidefinite mean. Namely, we have by Proposition 4.1.3 that $\mathbb{E}\boldsymbol{X}^2 = \mathbb{E}\boldsymbol{Y}^2 = (\mathbb{E}\,\mathsf{tr}\,\boldsymbol{X}^2)\boldsymbol{I}_d \approx m_2\boldsymbol{I}_d = \boldsymbol{I}_d$ for large d. That makes the following result giving a "fix" to this issue not surprising.

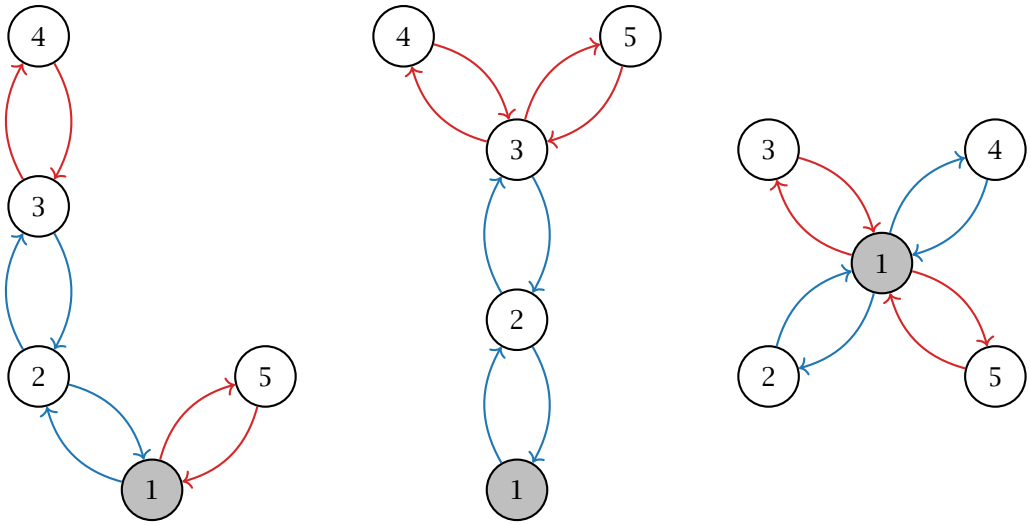


Figure 4.1: The graphical description of the three "types" of terms or colored closed walk graphs that give the value $t(\boldsymbol{X}^2\boldsymbol{Y}^2\boldsymbol{X}^2\boldsymbol{Y}^2) = 3$ in Proposition 4.1.17, as described in Remark 4.1.18. As in Figure 3.5, we give the ordering in which the vertices of each tree are visited, but really the information included in a closed walk graph is the ordering in which the edges are traversed; for a tree these are equivalent collections of information.

Proposition 4.1.19. If $\boldsymbol{X}, \boldsymbol{Y} \sim \mathsf{GOE}(d, \frac{1}{d})$ are independent, then

$$\begin{aligned} 0 &= \lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}\left((\boldsymbol{X}^2 - \boldsymbol{I}_d)(\boldsymbol{Y}^2 - \boldsymbol{I}_d)(\boldsymbol{X}^2 - \boldsymbol{I}_d)(\boldsymbol{Y}^2 - \boldsymbol{I}_d)\right) \\ &= t(\boldsymbol{X}^2\boldsymbol{Y}^2\boldsymbol{X}^2\boldsymbol{Y}^2) - 2t(\boldsymbol{X}^4\boldsymbol{Y}^2) - 2t(\boldsymbol{X}^2\boldsymbol{Y}^4) + 4t(\boldsymbol{X}^2\boldsymbol{Y}^2) \\ &\qquad + t(\boldsymbol{X}^4) + t(\boldsymbol{Y}^4) - 2t(\boldsymbol{X}^2) - 2t(\boldsymbol{Y}^2) + 1. \end{aligned}$$

Proof. Expanding the product and using the linearity of trace and expectation lets us write the given limit as a linear combination of the $t(w)$: after simplifying, we find

$$\begin{aligned} &\lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}\left((\boldsymbol{X}^2 - \boldsymbol{I}_d)(\boldsymbol{Y}^2 - \boldsymbol{I}_d)(\boldsymbol{X}^2 - \boldsymbol{I}_d)(\boldsymbol{Y}^2 - \boldsymbol{I}_d)\right) \\ &\quad = t(\boldsymbol{X}^2\boldsymbol{Y}^2\boldsymbol{X}^2\boldsymbol{Y}^2) - 2t(\boldsymbol{X}^4\boldsymbol{Y}^2) - 2t(\boldsymbol{X}^2\boldsymbol{Y}^4) + 4t(\boldsymbol{X}^2\boldsymbol{Y}^2) \\ &\qquad + t(\boldsymbol{X}^4) + t(\boldsymbol{Y}^4) - 2t(\boldsymbol{X}^2) - 2t(\boldsymbol{Y}^2) + 1 \end{aligned}$$

Using Lemma 4.1.10 and Proposition 4.1.17 to evaluate all terms,

$$\begin{aligned} &= t(\boldsymbol{X}^2\boldsymbol{Y}^2\boldsymbol{X}^2\boldsymbol{Y}^2) - 2m_4n_2 - 2m_2n_4 + 4m_2n_2 + m_4 + n_4 - 2m_2 - 2n_2 + 1 \\ &= 3 - 2\cdot 2\cdot 1 - 2\cdot 1\cdot 2 + 4\cdot 1\cdot 1 + 2 + 2 - 2\cdot 1 - 2\cdot 1 + 1 \\ &= 0, \end{aligned}$$

as claimed. □

The above observation is at the heart of the crucial definition that we present next. Hopefully from this sketch you can see how, by such combinatorial experiments, you could have arrived at this definition yourself.

4.1.3 Asymptotic Freeness

Continuing to experiment, you would see that the centering of monomials performed in Proposition 4.1.19 indeed makes all such "recentered words" have limiting normalized trace equal to zero. This turns out to be a property that simultaneously has very powerful consequences in determining the empirical spectral distribution of $\boldsymbol{X} + \boldsymbol{Y}$, and holds for many sequences of pairs $(\boldsymbol{X} = \boldsymbol{X}^{(d)}, \boldsymbol{Y} = \boldsymbol{Y}^{(d)})$. Thus, we give it its own name and take it as the fundamental assumption of the rest of our study of free probability theory.

Definition 4.1.20 (Asymptotic freeness)**.** Let $\boldsymbol{X} = \boldsymbol{X}^{(d)}, \boldsymbol{Y} = \boldsymbol{Y}^{(d)} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ be a sequence of pairs of random matrices satisfying the convergence of moments

$$\lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}\,\boldsymbol{X}^k = m_k, \tag{4.1.5}$$

$$\lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}\,\boldsymbol{Y}^k = n_k. \tag{4.1.6}$$

We say that this sequence of pairs $(\boldsymbol{X}, \boldsymbol{Y})$ is *asymptotically free* if, for all $t \geq 1$ and all $k_1, \ldots, k_t, \ell_1, \ldots, \ell_t \geq 1$, we have

$$\lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}\left((\boldsymbol{X}^{k_1} - m_{k_1}\boldsymbol{I}_d)(\boldsymbol{Y}^{\ell_1} - n_{\ell_1}\boldsymbol{I}_d)\cdots(\boldsymbol{X}^{k_t} - m_{k_t}\boldsymbol{I}_d)(\boldsymbol{Y}^{\ell_t} - n_{\ell_t}\boldsymbol{I}_d)\right) = 0. \tag{4.1.7}$$

For the sake of brevity, below we assume that asymptotic freeness also includes in the definition the convergence of moments.

One way to understand the definition is that each individual factor has limiting normalized trace zero:

$$\lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}^{k_a} - m_{k_a}\boldsymbol{I}_d) = \lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}(\boldsymbol{Y}^{\ell_b} - n_{\ell_b}\boldsymbol{I}_d) = 0.$$

Asymptotic freeness is the condition that you cannot interleave such factors and somehow generate an "alignment" that makes the trace become substantial in the limit.

You should think of the typical situation where asymptotic freeness breaks down as that where $\boldsymbol{X}$ and $\boldsymbol{Y}$ commute. In that case, we could rewrite, e.g.,

$$\begin{aligned}
&\lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}\left((\boldsymbol{X}^k - m_k\boldsymbol{I}_d)(\boldsymbol{Y}^\ell - n_\ell\boldsymbol{I}_d)(\boldsymbol{X}^k - m_k\boldsymbol{I}_d)(\boldsymbol{Y}^\ell - n_\ell\boldsymbol{I}_d)\right) \\
&\quad = \lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}\left((\boldsymbol{X}^k - m_k\boldsymbol{I}_d)^2(\boldsymbol{Y}^\ell - n_\ell\boldsymbol{I}_d)^2\right) \\
&\quad = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\langle(\boldsymbol{X}^k - m_k\boldsymbol{I}_d)^2, (\boldsymbol{Y}^\ell - n_\ell\boldsymbol{I}_d)^2\rangle,
\end{aligned}$$

where now the two matrices whose inner product we are taking are positive semidefinite and typically have eigenvalues of order $O(1)$, whereby this limit will typically be non-zero (for instance, you could engineer a situation where both of $(\boldsymbol{X}^k - m_k\boldsymbol{I}_d)^2$ and $(\boldsymbol{Y}^\ell - m_\ell\boldsymbol{I}_d)^2$ have all eigenvalues at least $\epsilon > 0$). Generally, you should hold the intuition that

"freeness is maximal non-commutativity."

Let us now see how the traces of words $t(w)$ behave under asymptotic freeness. The key observation is that, upon expanding the products in these conditions, we obtain a linear system in the $t(w)$ for various words w. For instance, you may check that the case $t = 2$ and $k_1 = k_2 = \ell_1 = \ell_2 = 1$, which gives the condition

$$0 = \lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}\left((\boldsymbol{X} - m_1\boldsymbol{I}_d)(\boldsymbol{Y} - n_1\boldsymbol{I}_d)(\boldsymbol{X} - m_1\boldsymbol{I}_d)(\boldsymbol{Y} - n_1\boldsymbol{I}_d)\right)$$

gives, upon expanding and rearranging by linearity, the equation

$$\begin{aligned}
0 &= t(\boldsymbol{X}\boldsymbol{Y}\boldsymbol{X}\boldsymbol{Y}) - n_1 t(\boldsymbol{X}\boldsymbol{Y}\boldsymbol{X}) - m_1 t(\boldsymbol{X}\boldsymbol{Y}^2) + m_1 n_1 t(\boldsymbol{X}\boldsymbol{Y}) \\
&\quad - n_1 t(\boldsymbol{X}^2\boldsymbol{Y}) + n_1^2 t(\boldsymbol{X}^2) + m_1 n_1 t(\boldsymbol{X}\boldsymbol{Y}) - m_1 n_1^2 t(\boldsymbol{X}) \\
&\quad - m_1 t(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{Y}) + m_1 n_1 t(\boldsymbol{Y}\boldsymbol{X}) + m_1^2 t(\boldsymbol{Y}^2) - m_1^2 n_1 t(\boldsymbol{Y}) \\
&\quad + m_1 n_1 t(\boldsymbol{X}\boldsymbol{Y}) - m_1 n_1^2 t(\boldsymbol{X}) - m_1^2 n_1 t(\boldsymbol{Y}) + m_1^2 n_1^2 \\
&= t(\boldsymbol{X}\boldsymbol{Y}\boldsymbol{X}\boldsymbol{Y}) - 2n_1 t(\boldsymbol{X}^2\boldsymbol{Y}) - 2m_1 t(\boldsymbol{X}\boldsymbol{Y}^2) \\
&\quad + 4m_1 n_1 t(\boldsymbol{X}\boldsymbol{Y}) + n_1^2 t(\boldsymbol{X}^2) + m_1^2 t(\boldsymbol{Y}^2) \\
&\quad - 2m_1 n_1^2 t(\boldsymbol{X}) - 2m_1^2 n_1 t(\boldsymbol{Y}) + m_1^2 n_1^2,
\end{aligned}$$

where in the second formulation we condense a little bit using the cyclic property. Moreover, this linear system is "triangular," in the sense that in the constraint associated to a centered version of the word w, $t(w)$ occurs with coefficient 1, and all other words appearing have length strictly smaller than the length of w. Thus, this linear system may be solved recursively, and we obtain the following.

Theorem 4.1.21. Suppose that $\boldsymbol{X} = \boldsymbol{X}^{(d)}, \boldsymbol{Y} = \boldsymbol{Y}^{(d)} \in \mathbb{R}_{\mathsf{sym}}^{d\times d}$ are asymptotically free. Then, for every $w \in \{\boldsymbol{X}, \boldsymbol{Y}\}^k$, there exists a polynomial T_w on $2k$ inputs depending only on the word w such that

$$t(w) = T_w(m_1, \ldots, m_k, n_1, \ldots, n_k).$$

Further, for all $k \geq 1$, there exists a polynomial S_k on $2k$ inputs depending only on k such that

$$\lim_{d\to\infty} \mathbb{E}\,\mathrm{tr}(\boldsymbol{X} + \boldsymbol{Y})^k = S_k(m_1, \ldots, m_k, n_1, \ldots, n_k).$$

Proof. The first claim follows from the argument given above. The second claim follows from the word expansion in Proposition 4.1.7, and indeed that gives the explicit description

$$S_k = \sum_{w\in\{\boldsymbol{X},\boldsymbol{Y}\}^k} T_w$$

relating the two families of polynomials. □

As we have mentioned before, you should think of the $t(w)$ as the joint tracial moments of $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the limit $d \to \infty$, describing a "joint spectral distribution" in this limit. Then, the Proposition says that, under asymptotic freeness, this joint spectral distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is fully determined by the individual limiting spectral distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ (the individual limiting tracial moments m_k and n_k). See Remark 4.1.6 above for a more precise discussion of these notions.

In the scalar case, the simple condition that gives the same behavior is independence: if scalar random variables X and Y are independent, then their individual distributions determine their joint distribution. Thus, asymptotic freeness is in some ways analogous to independence of scalar random variables (for this reason it is also sometimes called asymptotic *free independence* instead).

Bibliography

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

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

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

[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.

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

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

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

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

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

[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.

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

[Tao12] Terence Tao. *Topics in random matrix theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, 2012.