Lecture Notes on Random Matrix Theory in Data Science and Statistics

Dmitriy (Tim) Kunisky

Fall 2024 (Last Updated: November 24, 2024)

Contents

1 | First Steps With Rectangular Matrices

Often in statistical and data science applications, we are given the vague task of "understanding" a large and high dimensional dataset. Say we are given points $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^m$, for both m and n large. For instance, each $\boldsymbol{y}_i$ might represent a sample drawn from a population, and each coordinate $(\boldsymbol{y}_i)_j \in \mathbb{R}$ a feature or property of that sample (say, we have a sample of n people living in a given city, each having m numerical properties, like height, weight, age, income, duration of residence, answers to survey questions, and so forth).

At the beginning of an investigation like this, we do not always have a concrete question to ask of the data; rather, we are doing *exploratory data analysis*, looking for useful summaries and easily parseable descriptions of this dataset. We want to know: What is typical of the people who live in this city? Are they divided into a few natural subpopulations? What are the important correlations between the quantities we measured? Et cetera. What should we calculate?

If $m = 1$ and we only measure, say, the heights of a random sample of $n = 10\,000$ people from Baltimore, we can draw a histogram of the distribution of these values, which will capture essentially[1] all of the same information as our dataset itself. We will be able to see various important features like how much this distribution is concentrated, whether it has one or two or more modes, how its right and left tails compare, whether there are outliers, and so on. To accompany this visualization, we can also compute numerical summary statistics like the mean, median, quartiles, and standard deviation.

If $m = 2$ and we measure, say, the height and weight, we can still draw a picture of the entire dataset in a scatter plot, drawing each $\boldsymbol{y}_i$ as a point in $\mathbb{R}^2$. We can also repeat the $m = 1$ analysis for each feature individually. In addition to the individual means and standard deviations, you are probably aware that it is important to assess the correlation between the two features you draw in this way. You can calculate the covariance or the correlation coefficient, but it is also important to look at the scatter plot, because not all kinds of dependence are captured by these summary statistics.

If $m = 10$, already this methodology faces challenges. We can repeat the $m = 1$ analysis for each feature, and the $m = 2$ analysis for each pair of features (there are $\binom{10}{2} = 45$, which already sounds a bit unpleasant). But, again and moreso, there are many kinds of structures of dependence among these 10 variables that this approach can miss. Early statisticians, prominently John Tukey among them, spent time developing methods for dealing with this

[1] Setting aside the issue of choosing histogram bin widths and positions.

kind of situation.[2]

But now, what if $m = 1000$? Already looking at every pair of variables, that is $\binom{1000}{2} = 499\,500$ many pairs, is prohibitively expensive. Clearly we need some automated method of directing our attention to the "most important" variables or structural features of the dataset that is faster and more rigorous than just visual inspection. This is the first set of ideas that we will discuss.

Random matrices can actually be seen as playing two different roles in what we have discussed. On the one hand, when we calculate things like empirical covariances, we are implicitly viewing the $\boldsymbol{y}_i$ as being organized into a matrix $\boldsymbol{Y} = [\, \boldsymbol{y}_1 \cdots, \boldsymbol{y}_n \,] \in \mathbb{R}^{m \times n}$ and looking at various structural properties of this matrix (see the first section below). When the $\boldsymbol{y}_i$ are drawn at random from some ambient distribution (say, from a statistical population), then $\boldsymbol{Y}$ is itself already a random matrix. However, we will begin to approach random matrix theory from a different direction. It turns out that simpler random matrices—other matrices $\boldsymbol{G}$ with simple structures like having independent standard Gaussian entries—are also useful for us to construct and use as algorithmic and analytical tools upon $\boldsymbol{Y}$. We will get an initial handle on some of the flavor of how random matrices behave by looking at these methods.

1.1 Warmup: Singular Values and Principal Components

Before discussing the use of random matrices and the associated random mappings of a dataset, let us review one of the classical deterministic ways of finding a low-dimensional summary of $\boldsymbol{Y}$. There are numerous such methods, but we will focus on a simple one that is closely related to the eigenvalues and singular values of matrices, which will play a crucial role throughout.

1.1.1 Preliminaries and Singular Value Decomposition

Let us first recall the singular value decomposition (SVD) theorem.

Definition 1.1.1 (Orthogonal matrices)**.** *$\mathcal{O}(m)$ denotes the set of $m \times m$ matrices $\boldsymbol{U}$ that are* orthogonal*, i.e., that have $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_m$.*

One interpretation is that the columns of an orthogonal matrix give an orthonormal basis of $\mathbb{R}^m$ (it is slightly confusing; perhaps $\mathcal{O}(m)$ should really have been called the orthonormal matrices), and then $\mathcal{O}(m)$ is the set of all orthonormal bases. That is fine and true, but the more mature interpretation is that $\mathcal{O}(m)$ is really the set of orthogonal *transformations* of $\mathbb{R}^m$ (those linear maps that preserve both the origin and angles and distances). These transformations correspond to bases by $\boldsymbol{U} \in \mathcal{O}(m)$ corresponding to that basis to which it sends to the standard basis $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$. Reflect on this for a little while if it does not sound obvious to you. For geometric intuition, you may rely on the fact that orthogonal matrices

[2] One of my favorite ideas to come from this early statistics in moderately high dimension: Herman Chernoff, the namesake of the Chernoff bound, devised the method of "Chernoff faces" to plot each $\boldsymbol{y}_i$ as a cartoonish human face.

are generated by (i.e., each is a product of) rotations in different two-dimensional subspaces and reflections across different hyperplanes (look up "Householder reflections" and "Givens rotations" if you want more details).

Theorem 1.1.2 (Singular value decomposition)**.** *For any $\boldsymbol{Y} \in \mathbb{R}^{m\times n}$, there exist $\boldsymbol{U} \in \mathcal{O}(m)$, $\boldsymbol{V} \in \mathcal{O}(n)$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ such that $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ and such that $\boldsymbol{\Sigma}$ satisfies:*

1. *$\Sigma_{ij} = 0$ unless $i = j$.*
2. *$\sigma_i := \Sigma_{ii} \geq 0$.*
3. *$\sigma_1 \geq \cdots \geq \sigma_{\min\{m,n\}}$.*

Another way to say this is that, if $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m \in \mathbb{R}^m$ are the columns of $\boldsymbol{U}$ (an orthonormal basis, per the above discussion), and the $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^n$ are those of $\boldsymbol{V}$, then $\boldsymbol{Y} = \sum_{i=1}^{\min\{m,n\}} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$. The σ_i are called the singular values *of $\boldsymbol{Y}$, and the $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ the* left *and* right singular vectors*, respectively.*

Moreover, the σ_i (under the above conditions) are uniquely determined by $\boldsymbol{Y}$. We write $\boldsymbol{\sigma} = \boldsymbol{\sigma}(\boldsymbol{Y})$ and $\sigma_i = \sigma_i(\boldsymbol{Y})$ for this mapping. If a given σ_i occurs only once, then $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are also uniquely determined, up to the sign flip $(\boldsymbol{u}_i, \boldsymbol{v}_i) \mapsto (-\boldsymbol{u}_i, -\boldsymbol{v}_i)$.[3]

There are a few ways to think about the SVD. On the one hand, it gives a *structure theorem* for the linear maps described by arbitrary matrices: we have $\boldsymbol{Y}\boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i$, so the SVD theorem says that *any* matrix can be described as mapping some orthonormal basis to another orthonormal basis and then applying a rescaling. In particular, the singular values of $\boldsymbol{\sigma}(\boldsymbol{Y})$ are the part of this characterization that are independent of choices of orthonormal basis for either the domain $\mathbb{R}^n$ or the range $\mathbb{R}^m$, and so are the correct geometric summary of the "shape" of how a matrix acts. We will return to this interpretation and the importance of the singular values later.

On the other hand, and what will concern us for now, the SVD can be viewed as giving a decomposition of any matrix as a sum of matrices $\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ of *rank one*. Most simply, you can think of rank one matrices as ones that are easy to visualize or understand: instead of the full mn many numbers needed to express $\boldsymbol{Y}$, to express $\boldsymbol{u}_i \boldsymbol{v}_i^\top$ requires only $m+n$ many numbers, and moreover we can plot the coordinates of two or three of the $\boldsymbol{u}_i$ or $\boldsymbol{v}_i$ against one another easily. More precisely, a rank one matrix is a "one-dimensional object," in the sense that, as a matrix, it picks out just the $\boldsymbol{v}_i$ direction of $\mathbb{R}^n$ and acts non-trivially on it; $\boldsymbol{u}_i \boldsymbol{v}_i^\top$ maps every vector $\boldsymbol{x}$ that is orthogonal to $\boldsymbol{v}_i$ to zero.

1.1.2 SVD as Dimensionality Reduction

The second interpretation above leads, on further investigation, to a "variational" description of the SVD, i.e., to a description of the SVD as solving an optimization problem, in this case expressing that truncations of the SVD yield the best possible low-rank approximations of a matrix.

We need a few definitions and basic properties of the operator norm.

[3] If σ_i occurs several times, then it is only the *subspaces* spanned by the corresponding columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ that are unique.

Definition 1.1.3. *The* operator norm *of a matrix* $\boldsymbol{Y}$ *is* $\|\boldsymbol{Y}\| := \sigma_1(\boldsymbol{Y})$.

Proposition 1.1.4. *The operator norm is indeed a norm; that is, the following properties hold:*

1. *(Linearity) For any* $c \in \mathbb{R}$ *and* $\boldsymbol{Y} \in \mathbb{R}^{m\times n}$, $\|c\boldsymbol{Y}\| = |c| \cdot \|\boldsymbol{Y}\|$.
2. *(Triangle inequality) For any* $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{m\times n}$, $\|\boldsymbol{X} + \boldsymbol{Y}\| \le \|\boldsymbol{X}\| + \|\boldsymbol{Y}\|$.
3. *(Positivity) For* $\boldsymbol{Y} \in \mathbb{R}^{m\times n}$, $\|\boldsymbol{Y}\| = 0$ *if and only if* $\boldsymbol{Y} = \boldsymbol{0}$.

Proposition 1.1.5 (Variational form of operator nom)**.** $\|\boldsymbol{Y}\| = \sup_{\boldsymbol{v}\neq\boldsymbol{0}} \|\boldsymbol{Y}\boldsymbol{v}\|/\|\boldsymbol{v}\|$.

Theorem 1.1.6 (Eckart-Young-Mirsky for operator norm)**.** *For any* $\boldsymbol{Y} \in \mathbb{R}^{m\times n}$ *and* $1 \le d < m$,

$$\left\{ \begin{array}{ll} \textit{minimize} & \|\boldsymbol{Y} - \boldsymbol{Z}\| \\ \textit{subject to} & \boldsymbol{Z} \in \mathbb{R}^{m\times n}, \\ & \mathsf{rank}(\boldsymbol{Z}) \le d \end{array} \right\} = \sigma_{d+1}(\boldsymbol{Y}), \tag{1.1.1}$$

and a minimizer is $\boldsymbol{Z}^\star = \sum_{i=1}^{d} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ *where* $\sigma_i, \boldsymbol{u}_i, \boldsymbol{v}_i$ *are as in the discussion of the SVD above. If no* σ_i *are repeated, then this is the unique minimizer.*

Proof. First, observe that $\boldsymbol{Z}^\star$ indeed achieves the claimed objective value:

$$\|\boldsymbol{Y} - \boldsymbol{Z}\| = \left\| \sum_{i=d+1}^{\min\{m,n\}} \sigma_i(\boldsymbol{Y}) \boldsymbol{u}_i \boldsymbol{v}_i^\top \right\| = \sigma_{d+1}, \tag{1.1.2}$$

since the remaining matrix inside is given in its singular value decomposition and the largest remaining singular value is σ_{d+1}.

Next we must show that $\|\boldsymbol{Y} - \boldsymbol{Z}\| \ge \sigma_{d+1}$ whenever $\mathsf{rank}(\boldsymbol{Z}) \le d$. Note that $\dim(\ker(\boldsymbol{Z})) = n - \mathsf{rank}(\boldsymbol{Z}) \ge n - d$. Thus there must be some $\boldsymbol{v} \in \mathsf{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{d+1})$ such that $\boldsymbol{v} \neq \boldsymbol{0}$ and $\boldsymbol{Z}\boldsymbol{v} = \boldsymbol{0}$. Suppose we may expand $\boldsymbol{v} = \sum_{i=1}^{d+1} \alpha_i \boldsymbol{v}_i$, in which case $\|\boldsymbol{v}\|^2 = \sum_{i=1}^{d+1} \alpha_i^2$. We have:

$$\begin{aligned} \|(\boldsymbol{Y} - \boldsymbol{Z})\boldsymbol{v}\| = \|\boldsymbol{Y}\boldsymbol{v}\| &= \left\| \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \right) \left(\sum_{j=1}^{d+1} \alpha_j \boldsymbol{v}_j \right) \right\| \\ &= \left\| \sum_{i=1}^{\min\{m,n\}} \sum_{j=1}^{d+1} \sigma_i \alpha_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle \boldsymbol{u}_i \right\| \\ &= \left\| \sum_{j=1}^{d+1} \sigma_j \alpha_j \boldsymbol{u}_j \right\| && \text{(orthonormality of } \boldsymbol{v}_j\text{)} \\ &= \sqrt{\sum_{j=1}^{d+1} \sigma_j^2 \alpha_j^2} && \text{(orthonormality of } \boldsymbol{u}_j\text{)} \\ &\ge \sigma_{d+1} \sqrt{\sum_{j=1}^{d+1} \alpha_j^2} \\ &= \sigma_{d+1} \|\boldsymbol{v}\|, \end{aligned} \tag{1.1.3}$$

and thus $\|\boldsymbol{Y} - \boldsymbol{Z}\| \ge \sigma_{d+1}$ by Proposition 1.1.5. □

Actually, the operator norm here is not very special, and other ways of measuring the quality of a low-rank approximation work as well. Here is one important one which we will take this opportunity to introduce.

Definition 1.1.7. *The* Frobenius norm *of a matrix $\boldsymbol{Y}$ is $\|\boldsymbol{Y}\|_F := \sqrt{\mathsf{Tr}(\boldsymbol{Y}^\top \boldsymbol{Y})} = \sqrt{\sum_{i,j} Y_{ij}^2}$. (It is just the standard ℓ^2 norm when we ignore the matrix structure and view $\boldsymbol{Y}$ as a big vector.)*

Proposition 1.1.8. *The Frobenius norm is indeed a norm.*

Theorem 1.1.9 (Eckart-Young-Mirsky for Frobenius norm)**.** *For any $\boldsymbol{Y} \in \mathbb{R}^{m\times n}$ and $1 \le d < m$,*

$$\left\{ \begin{array}{ll} \textit{minimize} & \|\boldsymbol{Y} - \boldsymbol{Z}\|_F \\ \textit{subject to} & \boldsymbol{Z} \in \mathbb{R}^{m\times n}, \\ & \mathsf{rank}(\boldsymbol{Z}) \le d \end{array} \right\} = \left(\sum_{i=d+1}^{\min\{m,n\}} \sigma_i(\boldsymbol{Y})^2 \right)^{1/2}, \tag{1.1.4}$$

and a minimizer is again $\boldsymbol{Z}^\star = \sum_{i=1}^d \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$.

Thus in various senses truncating the SVD to the top (i.e., highest singular value) d components gives the best possible rank-d approximation of a matrix. Let us see how such an approximation can be useful in applications; we will first give some more "plain matrix" applications, and then return to the statistical setting from before.

1.1.3 Application: Compressing Images

One straightforward application of this method of dimensionality reduction is a naive, yet surprisingly effective, approach to image compression. We may encode a grayscale image as a matrix $\boldsymbol{Y} \in [0,1]^{m\times n}$ with the value of an entry corresponding to the intensity of a pixel. The only issue with directly approximating $\boldsymbol{Y}$ by $\sum_{i=1}^d \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ is that the entries of this latter matrix are not necessarily in $[0,1]$. Crudely resolving this issue by "clipping" the entries—replacing them with 0 if they are smaller than 0 or 1 if they are greater than 1—is fine for our purposes.

Figure 1.1 shows two examples of this approach. It is of course no match to more advanced approaches like JPEG, but is perhaps surprisingly effective. You can see, however, the tendency even at fairly high rank for the approximated image to look "blocky" or "grainy" with poorly rendered solid regions having a solid background but a curved boundary. You might consider what kinds of images a rank 1 matrix can represent to get some intuition for this phenomenon.

1.1.4 Application: Drawing Graphs

Another, perhaps less straightforward application is to finding informative drawings of graphs. We may encode a graph G on n vertices by a matrix through the *adjacency matrix* $\boldsymbol{A} \in \mathbb{R}^{n\times n}_{\mathsf{sym}}$. This is a symmetric matrix, whereby a more natural approach is to look at the *spectral decomposition* $\boldsymbol{A} = \sum_{i=1}^n \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$ for orthonormal eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_1$ and $\lambda_1 \ge \cdots \ge \lambda_n$. Note that the eigenvalues are not necessarily positive, and indeed we must have $0 = \mathsf{Tr}(\boldsymbol{A}) = \sum_{i=1}^n \lambda_i$, so some eigenvalues must be negative.

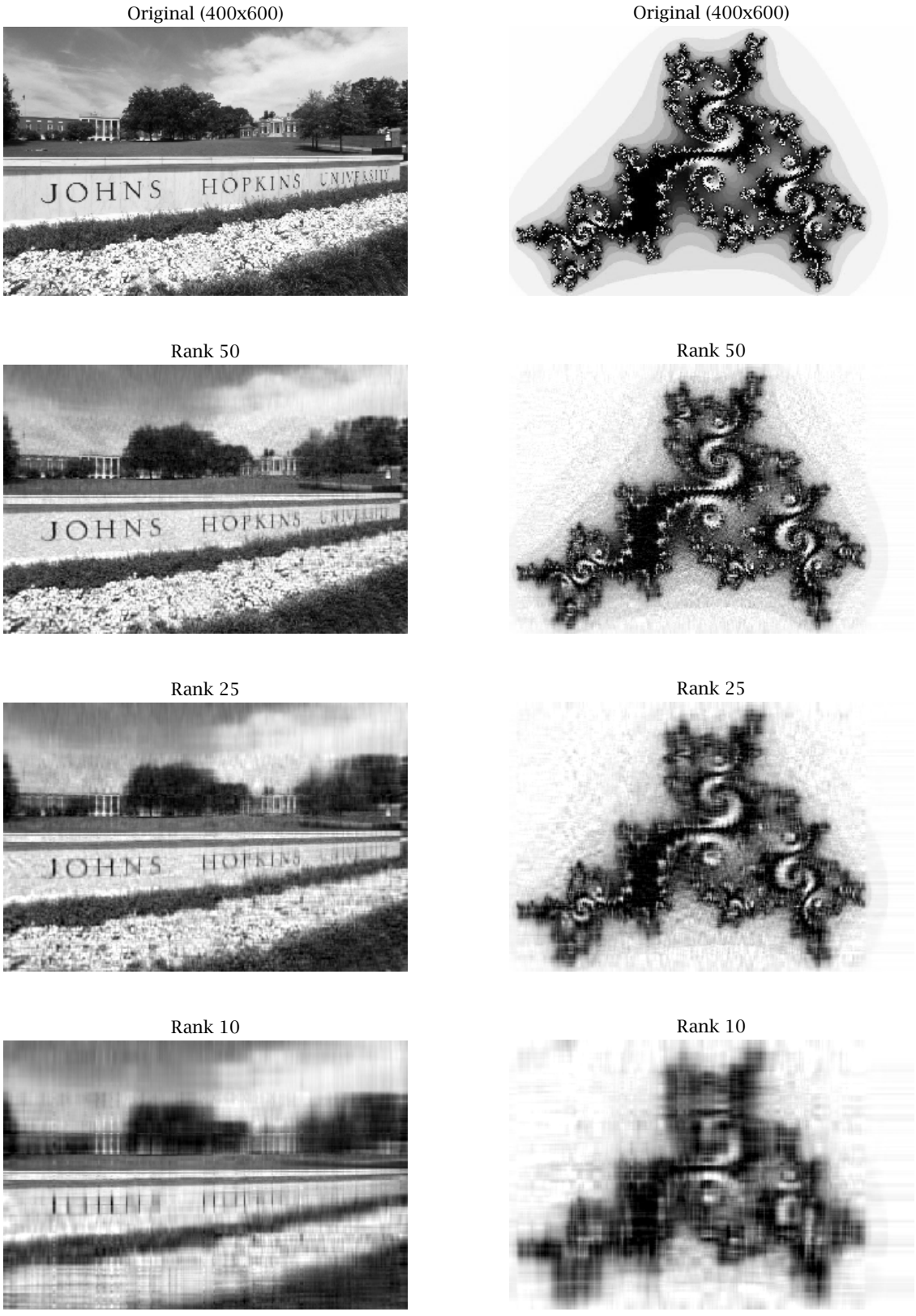


Figure 1.1: Two examples of image compression using truncated singular value decomposition, showing the deterioration of the approximation as the rank decreases.

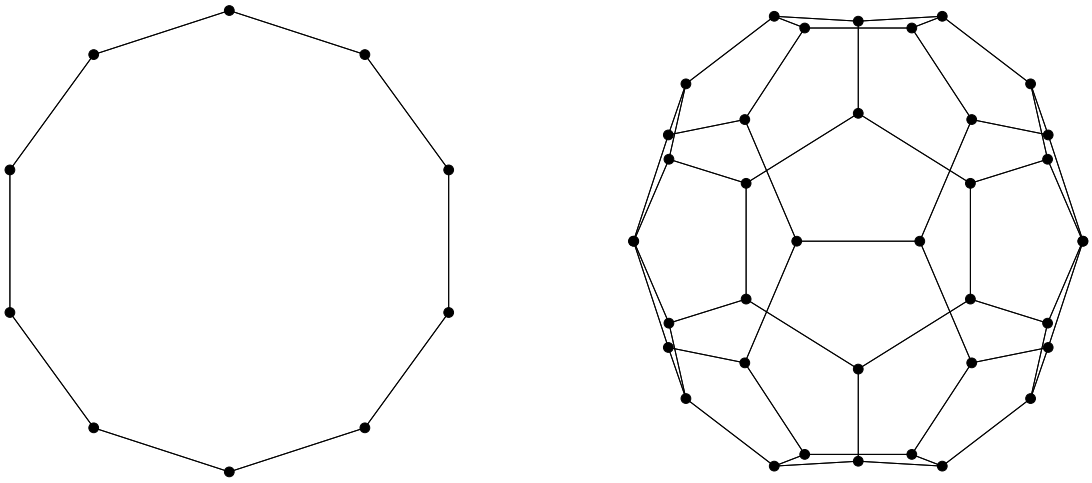


Figure 1.2: Two examples of two-dimensional graph drawings obtained by embedding using eigenvectors. The first is a cycle, while the second is a so-called *fullerene*, a graph describing a three-dimensional structure that can be formed by carbon atoms.

By the Perron-Frobenius theorem (which you might have encountered in the context of Markov chain theory), the top eigenvector $\boldsymbol{v}_1$ has non-negative entries. At a high level, you can think of its entries as related to the proportional degrees of vertices in G. For example, if G is *d-regular*, i.e., every vertex has degree d, then the top eigenvector is the (normalized) all-ones vector $\boldsymbol{v}_1 = \frac{1}{\sqrt{n}}\mathbf{1}$. This is usually not very informative, so the usual approach to obtain a good two-dimensional drawing of a graph using the eigenvectors is to plot the vertices according to their coordinates in $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$.

We show two examples in Figure 1.2. As a sanity check, we find that this method recovers the "normal" drawing of a cycle graph. More remarkably, it seems to be fairly faithful to the structure of a *fullerene* graph that describes a three-dimensional object—even though the eigenvectors have no reason to "know" about this secret dimensionality.

1.1.5 Application: Summarizing Point Clouds

Finally, let us consider again the issue of "summarizing" a collection of points $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, say given as the results of a statistical experiment. We think geometrically of trying to approximate the "cloud" of points in $\mathbb{R}^m$ formed by the $\boldsymbol{y}_i$. Let us look for a low-dimensional subspace W that comes close to interpolating these points. We parametrize this subspace, say d-dimensional for $1 \leq d \leq m$, by a spanning set $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d$. Then, we want each $\boldsymbol{y}_i$ to be close to W, so that there exist $x_{i1}, \ldots, x_{id}$ such that

$$\boldsymbol{y}_i \approx \sum_{j=1}^{d} x_{ij}\boldsymbol{w}_j = \boldsymbol{W}\boldsymbol{x}_i \text{ for each } i \in [n], \tag{1.1.5}$$

where we introduce the matrix $\boldsymbol{W} = [\, \boldsymbol{w}_1 \cdots \boldsymbol{w}_d \,] \in \mathbb{R}^{m \times d}$ and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id}) \in \mathbb{R}^d$. To fully use the powerful matrix notation, if we want this to hold for each $\boldsymbol{y}_i$, we can just ask

that we have the approximate matrix decomposition

$$\boldsymbol{Y} \approx \boldsymbol{W}\boldsymbol{X}, \tag{1.1.6}$$

for $\boldsymbol{X} = [\, \boldsymbol{x}_1 \, \cdots \, \boldsymbol{x}_n \,] \in \mathbb{R}^{d \times n}$.

In fact, using the SVD we can see that *any* matrix of rank at most d can be written as $\boldsymbol{W}\boldsymbol{X}$ like this. So, if we seek to minimize the natural objective

$$\|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{X}\|_F^2 = \sum_{i=1}^{n} \left\| \boldsymbol{y}_i - \sum_{j=1}^{d} x_{ij} \boldsymbol{w}_j \right\|^2, \tag{1.1.7}$$

then we will end up with precisely one of the variational problems that leads to truncating the SVD, per Theorem 1.1.9 (the Eckart-Young-Mirsky theorem for the Frobenius norm). Applying the Theorem, we find that the best subspace to project to in this sense for dimensionality reduction of a point cloud is that spanned by the first d left singular vectors, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$.

1.2 Multiplication by a Gaussian Random Matrix

The toolkit based on SVD discussed above, while powerful, has some downsides. First, the SVD takes a relatively long time—about $O(n^3)$ for $m \asymp n$ for the whole thing, or $O(dn^2)$ to extract the top d singular values and vectors—to compute.

Second, the SVD requires us to first gather the entire matrix and store it in memory before computing an approximation. If our goal is statistical analysis, maybe this is fine; we would run our experiment or survey, gather the data, and then experiment with dimensionality reduction. But, if our bottleneck is in the storage of data itself, or if we seek to apply compression immediately, or if we want to run analytics on a reduced dataset quickly in a *streaming* fashion as data arrive, then this requirement is inconvenient.

Finally, while the SVD promises to be the "best" low-rank approximation in the sense of distances between matrices, this guarantee is not always relevant. Consider, for example, very small numbers $\epsilon_1 > \cdots > \epsilon_m$, and suppose $m = n$ and we are given the point cloud of $\boldsymbol{y}_i = (1+\epsilon_i)\boldsymbol{e}_i$ for $i \in [n]$. These are very close to just being the standard basis of $\mathbb{R}^m$, and $\boldsymbol{Y}$ is close to the identity matrix $\boldsymbol{I}_m$. For ϵ_i very small, all points are "comparably important" to the geometric configuration of this set. Yet, you can check that the SVD truncated to D components will just throw away all but the first d of the data points. Clearly a more even-handed treatment of the data points themsleves would often be valuable.

To address all these issues, we will now explore our first application of random matrices, which will simply entail reducing the dimension of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ by multiplying them by a random matrix $\boldsymbol{G} \in \mathbb{R}^{d \times m}$, i.e., by applying a random linear mapping to them. Moreover, we will take $\boldsymbol{G}$ to be what we will see to be the simplest and most canonical of random matrices, having i.i.d. standard Gaussian entries. That is, we will have $G_{ij} \sim \mathcal{N}(0,1)$ independently. We abbreviate this $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d \times m}$ (this notation, alluding with the tensor product symbol "$\otimes$" to $\boldsymbol{G}$ having a *product measure* for its law, is occasionally used in the literature but is not entirely standard).

Let us start by trying to gain some intuition about what $\boldsymbol{G}$ does to a single $\boldsymbol{y} \in \mathbb{R}^n$. We immediately run into the reason why working with Gaussian entries specifically is useful, which is the following fact.

Definition 1.2.1. *A* Gaussian random vector *is a random vector $(v_1, \dots, v_d)$ with a density $\det(2\pi\boldsymbol{\Sigma})^{-1/2} \exp(-\frac{1}{2}(\boldsymbol{v}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{v}-\boldsymbol{\mu}))$ for $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ strictly positive definite. We write $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.*

Proposition 1.2.2. *If $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a d-dimensional Gaussian random vector and $\boldsymbol{A} \in \mathbb{R}^{k\times d}$, then*

$$\mathsf{Law}(\boldsymbol{A}\boldsymbol{v}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}^\top \boldsymbol{\Sigma} \boldsymbol{A}). \tag{1.2.1}$$

That is, the linear image of a Gaussian random vector is another Gaussian random vector, and in particular is determined by its mean and covariance, or its first two moments.

Thus to identify the law of $\boldsymbol{G}\boldsymbol{y}$ it suffices to compute its first two moments. You may verify that:

$$\mathbb{E}[\boldsymbol{G}\boldsymbol{y}] = \mathbb{E}[\boldsymbol{G}]\boldsymbol{y} = \boldsymbol{0}, \tag{1.2.2}$$

$$\begin{aligned} \mathsf{Cov}((\boldsymbol{G}\boldsymbol{y})_i, (\boldsymbol{G}\boldsymbol{y})_j) &= \mathbb{E}(\boldsymbol{G}\boldsymbol{y})_i (\boldsymbol{G}\boldsymbol{y})_j \\ &= \mathbb{E}\left(\sum_{a=1}^n G_{ia}\mathcal{y}_a\right)\left(\sum_{b=1}^n G_{jb}\mathcal{y}_b\right) \\ &= \left\{ \begin{array}{ll} 0 & \text{if } i \neq j, \\ \|\boldsymbol{y}\|^2 & \text{if } i = j \end{array} \right\}. \end{aligned} \tag{1.2.3}$$

In short, we have

$$\mathsf{Law}(\boldsymbol{G}\boldsymbol{y}) = \mathcal{N}(\boldsymbol{0}, \|\boldsymbol{y}\|^2 \boldsymbol{I}_d). \tag{1.2.4}$$

From this we may, for example, compute what multiplication by $\boldsymbol{G}$ does to expected lengths:

$$\mathbb{E}\|\boldsymbol{G}\boldsymbol{y}\|^2 = \mathsf{Tr}\,\mathsf{Cov}(\boldsymbol{G}\boldsymbol{y}) = d\|\boldsymbol{y}\|^2. \tag{1.2.5}$$

This suggests that, if we want to produce embeddings preserving geometry, we should instead work with the normalization

$$\widehat{\boldsymbol{G}} := \frac{1}{\sqrt{d}}\boldsymbol{G}. \tag{1.2.6}$$

This, at the very least, will preserve the lengths of vectors in expectation:

$$\mathbb{E}\|\widehat{\boldsymbol{G}}\boldsymbol{y}\|^2 = \|\boldsymbol{y}\|^2. \tag{1.2.7}$$

What about angles? By a similar calculation you can find

$$\mathbb{E}\langle \widehat{\boldsymbol{G}}\boldsymbol{y}_1, \widehat{\boldsymbol{G}}\boldsymbol{y}_2\rangle = \langle \boldsymbol{y}_1, \boldsymbol{y}_2\rangle, \tag{1.2.8}$$

that is, that multiplication by $\widehat{\boldsymbol{G}}$ also preserves angles in expectation. (Actually, this also follows directly from the preservation of 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, $\widehat{\boldsymbol{G}}$ preserves 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 as the columns of $\boldsymbol{Y}$, we have

$$\mathbb{E}(\widehat{\boldsymbol{G}}\boldsymbol{Y})^\top(\widehat{\boldsymbol{G}}\boldsymbol{Y}) = \boldsymbol{Y}^\top\boldsymbol{Y}. \tag{1.2.9}$$

Remark 1.2.3. *One way to view these preliminary calculations is as studying 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 $\boldsymbol{y} \in \mathbb{R}^m$. The calculations of $\mathbb{E}\boldsymbol{G}\boldsymbol{y}$ and $\mathbb{E}(\boldsymbol{G}\boldsymbol{y})(\boldsymbol{G}\boldsymbol{y}')^\top$ (similar to but more general than what we have done above) are then just calculating the mean and covariance of this field, which, as it is Gaussian, characterize it completely.*

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 $\boldsymbol{Q} \in \mathcal{O}(n)$ 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.

So, this is telling us that, in expectation, $\widehat{\boldsymbol{G}}$ preserves the relative geometry of any given point cloud. Yet, this is too good to be true in any reasonable "hard" sense (say, a bound on how much $\widehat{\boldsymbol{G}}$ changes lengths and angles) for an *arbitrary* point cloud once we fix $\widehat{\boldsymbol{G}}$: in fact $\widehat{\boldsymbol{G}}$, being low rank, sends a whole $(n-d)$-dimensional subspace of $\mathbb{R}^n$ to zero, so there are many point clouds whose structure it completely destroys! It is important, rather, that we are thinking of a point cloud chosen *before* $\widehat{\boldsymbol{G}}$ is drawn, so that the randomness of $\widehat{\boldsymbol{G}}$ acts in our favor, the random subspace that is the kernel of $\widehat{\boldsymbol{G}}$ (the vectors it sends to zero) avoiding the point cloud. Even with a point cloud "oblivious" to $\widehat{\boldsymbol{G}}$ in this fashion, we cannot always make such a guarantee: if we choose a dense enough grid of points in $\mathbb{R}^n$, many points will come close to any subspace and some aspects of their relative geometry must be "flattened" by $\widehat{\boldsymbol{G}}$, so the point cloud also cannot be too dense.

1.3 Application: Johnson-Lindenstrauss Lemma [JL82]

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

Definition 1.3.1. *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[4] *on the $\boldsymbol{y}_i$ if, for all $i, j \in [n]$,*

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

Theorem 1.3.2 (Johnson-Lindenstrauss [JL82, IM98])**.** *There is an absolute constant $C > 0$ such that the following holds. Let $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \in \mathbb{R}^m$ be arbitrary and*

$$d := C\frac{\log n}{\epsilon^2}. \tag{1.3.1}$$

[4] My own non-standard terminology.

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

In fact, as the proof will make clear, by increasing the constant C, you can achieve any smaller polynomial failure rate $O(1/n^K)$ for any $K > 0$.

1.3.1 Concentration of Gaussian Vector Norms

For the proof we will need the following result, which is itself a fundamental one expressing a crucial aspect of high-dimensional geometry.

Lemma 1.3.3. *For $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and any $t \geq 0$,*

$$\mathbb{P}\left[\left|\|\boldsymbol{g}\|^2 - d\right| \geq t\right] = \mathbb{P}\left[\left|\sum_{i=1}^{d} g_i^2 - d\right| \geq t\right] \leq 2\left\{\begin{array}{ll} \exp\left(-\frac{t^2}{8d}\right) & \text{if } t \leq d, \\ \exp\left(-\frac{t}{8}\right) & \text{if } t \geq d \end{array}\right\}. \tag{1.3.2}$$

Proof. The proof uses the venerable Chernoff bound. I will only deal with one of the tails; the other follows similarly. Note that $\mathbb{E}g_i^2 = 1$, so, defining $x_i := g_i^2 - 1$, we have

$$\begin{aligned} \mathbb{P}\left[\sum_{i=1}^{d} g_i^2 - d \geq t\right] &= \mathbb{P}\left[\sum_{i=1}^{d} x_i \geq t\right] \\ &= \mathbb{P}\left[\exp\left(\lambda \sum_{i=1}^{d} x_i\right) \geq \exp(\lambda t)\right] \\ &\leq \frac{\mathbb{E}\exp\left(\lambda \sum_{i=1}^{d} x_i\right)}{\exp(\lambda t)} && \text{(Markov inequality)} \\ &= \frac{(\mathbb{E}\exp(\lambda x_1))^d}{\exp(\lambda t)} && \text{(independence)} \\ &= \exp\left(-\lambda t + d \log \mathbb{E}\exp(\lambda x_1)\right). \end{aligned}$$

Let us write

$$\psi(\lambda) := \log \mathbb{E}\exp(\lambda x_1) = \log \mathbb{E}\exp(\lambda(g_1^2 - 1)), \tag{1.3.3}$$

the *moment generating function* of the random variable $g_1^2 - 1$ (where we retain the -1 so that the expectation is zero. You may check as a calculus exercise that the remaining expectation is finite if and only if $\lambda < \frac{1}{2}$, and in this case

$$\mathbb{E}\exp(\lambda g_1^2) = \frac{1}{\sqrt{1 - 2\lambda}}, \tag{1.3.4}$$

whereby

$$\psi(\lambda) = -\lambda + \frac{1}{2}\log\left(\frac{1}{1 - 2\lambda}\right). \tag{1.3.5}$$

Taylor expansion shows that $\psi(\lambda) = \lambda^2 + O(\lambda)$. Turning this into a concrete bound, $\psi(\lambda) \leq 2\lambda^2$ for all $\lambda \leq \frac{1}{4}$.

Thus we have

$$\mathbb{P}\left[\sum_{i=1}^{d} g_i^2 - d \geq t\right] \leq \exp\left(-\lambda t + d\psi(\lambda)\right)$$
$$\leq \exp\left(-\lambda t + 2d\lambda^2\right)$$

and now we may compute the optimal choice $\lambda := t/4d$, which gives

$$\leq \exp\left(-\frac{t^2}{8d}\right),$$

completing the proof of this side of the inequality with $c = 1/8$, provided that $\lambda = t/4d \leq 1/4$, or $t \leq d$.

In the other case $t \geq d$, we obtain the result by taking $\lambda = \frac{1}{4}$. □

Remark 1.3.4. *The heart of the calculation, you can convince yourself, is that $\psi(\lambda) = O(\lambda^2)$ for λ smaller than some constant. Note that the bound on λ is required, since this expectation becomes infinite for $\lambda \geq \frac{1}{2}$ in our case. This property, with the bound on λ, is called a random variable's being* subexponential*, a weaker version of the property of being* subgaussian *that you might have encountered before. The result of the Lemma is characteristic of sums of i.i.d. subexponential random variables: they have "Gaussian tails" scaling as $\exp(-ct^2/d)$ up to a certain cutoff of $t \sim d$, beyond which they only have "exponential tails" scaling as $\exp(-ct)$.* Bernstein's inequality *is the general tool expressing this behavior; see Chapter 1 of [RH17] for more on that. An rough intuitive explanation for this is that, when x_i themselves have exponential tails, then large deviations of $\sum_{i=1}^{d} x_i$ of order $t \ll d$ are driven by the x_i each being slightly unusually large, while large deviations of order $t \gg d$ are driven by the largest of x_i being unusually large, whereby the tail behavior becomes the same as that of an individual x_i.*

Informally, the result says that $\|\boldsymbol{g}\|^2 = d + O(\sqrt{d})$ with high probability. Taking square roots, we see that $\|\boldsymbol{g}\| = \sqrt{d} + O(1)$, so a random standard Gaussian vector usually falls close to the spherical shell of width $O(1)$ around the sphere of radius $\sqrt{d}$. This is quite counterintuitive if you have not seen it before: we think of a one- or two-dimensional Gaussian as having its "typical set" being a solid blob around the origin. But, a high-dimensional Gaussian actually has a *non-convex* typical set of a hollow spherical shell!

The general intuition you should have about high dimensional Gaussians is the following approximate equivalence of laws:

$$\text{“ } \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d) \approx \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d})) \text{ ”} \tag{1.3.6}$$

For example, the following is another instance of this idea:

Theorem 1.3.5 (Borel's limit theorem)**.** *Let $\boldsymbol{x} \sim \mathsf{Unif}(\mathbb{S}^{d-1}(1))$. Then, $\sqrt{d}\, x_1$ converges weakly in law, as $d \to \infty$, to $\mathcal{N}(0,1)$. In fact for any fixed $k \geq 1$, $(\sqrt{d}\, x_1, \ldots, \sqrt{d}\, x_k)$ converges weakly in law to $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$.*

1.3.2 Proof of Theorem 1.3.2

We are now ready to prove the main Johnson-Lindenstrauss result.

Proof of Theorem 1.3.2. Note first that we may assume ϵ is sufficiently small without loss of generality, since if ϵ is larger than some constant we can just perform the analysis for a smaller ϵ and absorb the difference into the constant C.

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

Consider an arbitrary $\boldsymbol{y} \in \mathbb{R}^m$. Recall from the discussion in Section 1.2 that $\mathsf{Law}(\widehat{\boldsymbol{G}}\boldsymbol{y}) = \mathcal{N}(0, \frac{1}{d}\|\boldsymbol{y}\|^2)$. We then have

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

and for ϵ sufficiently small by Lemma 1.3.3 we have

$$\leq 2\exp(-c\epsilon^2 d). \tag{1.3.7}$$

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

$$\mathop{\mathbb{P}}_{\hat{G}\sim\mathcal{N}(0,\frac{1}{d})^{\otimes d\times m}}\left[(1-\epsilon)\|\boldsymbol{y}\| \leq \|\widehat{\boldsymbol{G}}\boldsymbol{y}\| \leq (1+\epsilon)\|\boldsymbol{y}\|\right] \geq 1 - 4\exp(-c\epsilon^2 d). \tag{1.3.8}$$

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

$$\begin{aligned}
\mathbb{P}[\text{some } E_{ij} \text{ does not occur}] &\leq \binom{n}{2}\cdot 2\exp(-c\epsilon^2 d) \\
&\leq n^2\exp(-c\epsilon^2 d) \\
&= \exp(2\log n - c\epsilon^2 d),
\end{aligned} \tag{1.3.9}$$

and the result follows since, if $d \geq \frac{3}{c}\frac{\log n}{\epsilon^2}$, then this is at most $1/n$. □

Remark 1.3.6. *It is easy to extend the proof to show that we also have* $(1-\epsilon)\|\boldsymbol{y}_i\| \leq \|f(\boldsymbol{y}_i)\| \leq (1+\epsilon)\|\boldsymbol{y}_i\|$ *with the same high probability up to constants (say, by increasing* n *by one and adding* $\boldsymbol{0}$ *to the* $\boldsymbol{y}_i$*, or by including this directly in the union bound above). As mentioned before, it is also possible to get an error bound of* $1/n^K$ *for any* $K > 0$ *by setting a suitable* $C = C(K)$ *in the statement of the Theorem.*

1.3.3 Fast Johnson-Lindenstrauss Transform [AC09]

We originally complained that the SVD was too slow to compute, and reduced our computation to a matrix-vector product, which takes time for each $\boldsymbol{y}_i$ of $O(m \cdot \frac{\log n}{\epsilon^2}) = O(m \log n)$ for constant ϵ (the same as the number of entries of $\widehat{\boldsymbol{G}}$). While the above tells us that we cannot reduce the actual size of $\widehat{\boldsymbol{G}}$ to improve this, we can try to give it an additional structure that allows us to compute the matrix-vector product faster. There have been various interesting ideas to this effect, which are surveyed nicely in the introduction of [AC09], whose main idea we will sketch here.

First, various prior work showed that the particular Gaussian distribution of $\widehat{\boldsymbol{G}}$ is not essential. Indeed, the proof we presented is not that of Johnson and Lindenstrauss [JL82], who used $\widehat{\boldsymbol{G}}$ having random orthonormal rows, but rather a simplification due to Indyk and Motwani [IM98]. Other work also considered replacing $\mathcal{N}(0,1)$ with $\mathsf{Unif}(\{\pm 1\})$, making the projecting matrix simpler to sample [Ach01]. That work also began to consider obtaining (modest) speedups by replacing $\widehat{\boldsymbol{G}}$ with a *sparse* random matrix. If this is so, then of course the matrix-vector product can be computed faster since all zero entries can be ignored.

Attempt 1: Sparse Projection Specifically, we may consider fixing a small $q \in (0,1)$ and using the sparser random projection matrix with i.i.d. entries drawn as

$$P_{ij} \sim \begin{cases} \mathcal{N}(0, \frac{1}{qd}) & \text{with probability } q, \\ 0 & \text{with probability } 1-q \end{cases}. \tag{1.3.10}$$

We may multiply by such P_{ij} in time $O(qm \log n)$, which if we take $q = o(1)$ will give an asymptotic improvement. It is easy to check that $\mathbb{E}P_{ij} = \mathbb{E}\widehat{G}_{ij} = 0$ and $\mathbb{E}P_{ij}^2 = \mathbb{E}\widehat{G}_{ij}^2 = \frac{1}{d}$, so we again have $\mathbb{E}\|\boldsymbol{P}\boldsymbol{y}\|^2 = \|\boldsymbol{y}\|^2$ for any $\boldsymbol{y} \in \mathbb{R}^m$. However, there is an issue with the *concentration* of this quantity for $\boldsymbol{y}$ sparse. Indeed, consider the sparsest possible $\boldsymbol{y}$, $\boldsymbol{y} = \boldsymbol{e}_1$. We have $\boldsymbol{P}\boldsymbol{e}_1 = [\, P_{11} \cdots P_{d1} \,]$, and we calculate

$$\begin{aligned} \mathsf{Var}[\|\boldsymbol{P}\boldsymbol{e}_1\|^2] &= \mathsf{Var}\left[\sum_{i=1}^{d} P_{1d}^2\right] \\ &= d\,\mathsf{Var}[P_{11}^2] && \text{(independence)} \\ &= d(\mathbb{E}P_{11}^4 - (\mathbb{E}P_{11}^2)^2) \\ &= d\left(\frac{3}{q^2d^2} - \frac{1}{d^2}\right) \\ &= \frac{1}{d}\left(\frac{3}{q^2} - 1\right). \end{aligned} \tag{1.3.11}$$

We see in particular that this diverges as $q \to 0$, showing that indeed the sparser $\boldsymbol{P}$ is, the less concentrated $\|\boldsymbol{P}\boldsymbol{e}_1\|^2$. You may check that this causes substantial issues, making the Johnson-Lindenstrauss argument break down for $d \sim \log n$.

Attempt 2: Fourier Transform Preconditioning The next idea is to *precondition* by using as a projection matrix $\boldsymbol{P}\boldsymbol{H}$ for some $\boldsymbol{H} \in \mathbb{R}^{m \times m}$. We have a few desiderata for $\boldsymbol{H}$:

(1) it must itself not distort vector norms very much, (2) it must be possible to multiply by $\boldsymbol{H}$ much faster than the brute force $O(m^2)$, and (3) $\boldsymbol{H}$ must map sparse vectors to dense vectors, resolving the issue above. The beautiful idea of [AC09] is to use a *Fourier transform*, namely the *Walsh-Hadamard transform*, which is the Fourier transform with respect to the abelian group $(\mathbb{Z}/2\mathbb{Z})^k$ for $m = 2^k$. We will not go into the details here, but this choice satisfies all of the conditions above. First, $\boldsymbol{H} \in \mathcal{O}(m)$ and thus preserves norms exactly. Second, we may multiply by $\boldsymbol{H}$ in time $O(m \log m)$ using the fast Fourier transform. Finally, *uncertainty principles* (various mathematical facts in the spirit of the physical Heisenberg principle of quantum mechanics) guarantee that both $\boldsymbol{y}$ and $\boldsymbol{Hy}$ cannot be too sparse. Still, an issue remains: while $\boldsymbol{Hy}$ will never be sparse for sparse $\boldsymbol{y}$, $\boldsymbol{Hy}$ will still be sparse for some *dense* $\boldsymbol{y}$, as indeed is unavoidable for $\boldsymbol{H}$ as above since it is invertible.

ATTEMPT 3: SIGN FLIP PRECONDITIONING The actual construction of [AC09] adds a final layer of random preconditioning, using instead the matrix $\boldsymbol{PHD}$[5] where $\boldsymbol{D} \in \mathbb{R}^{m \times m}$ is diagonal with $D_{ii} \sim \mathsf{Unif}(\{\pm 1\})$ independently. We again have $\boldsymbol{D} \in \mathcal{O}(m)$, and multiplication by $\boldsymbol{D}$ may be performed in time $O(m)$. And, as the main Lemma of [AC09] states, with high probability, so long as $d = \Omega(\log n)$, we have $\boldsymbol{HD}\boldsymbol{y}_1, \dots, \boldsymbol{HD}\boldsymbol{y}_n$ are all not very sparse. The idea here is again elegant and subtle: if $\boldsymbol{y}$ is sparse, then $\boldsymbol{Dy}$ is sparse as well, and $\boldsymbol{HDy}$ is dense. If $\boldsymbol{y}$ is dense, say with all entries roughly equal, then $\boldsymbol{Dy}$ is uniformly random among 2^m far apart vectors in $\mathbb{R}^m$. And, most of these cannot map under $\boldsymbol{H}$ to sparse vectors by a volumetric argument: "there are fewer sparse vectors than dense vectors," an idea which we will leave vague for now but will return to when we discuss compressed sensing.

It turns out that the best scaling of q that still lets the Johnson-Lindenstrauss argument work is

$$q = \Theta\left(\frac{\log^2 n}{m}\right). \tag{1.3.12}$$

With this, we end up with a projection matrix $\boldsymbol{PHD}$ which achieves the same guarantee as in Theorem 1.3.2 (with constant ϵ), multiplication by which requires time

$$O\Big(\underbrace{\frac{\log^2 n}{m} \cdot m \log n}_{\text{for } \boldsymbol{P}} + \underbrace{m \log m}_{\text{for } \boldsymbol{H}} + \underbrace{m}_{\text{for } \boldsymbol{D}}\Big) = O\left(\log^3 n + m \log m\right), \tag{1.3.13}$$

which is much faster than the original projection when $n = \exp(m^\delta)$ for $0 < \delta < 1/2$.

1.3.4 EMBEDDING ARBITRARY METRIC SPACES

Another way to think about the definition of being pairwise ϵ-faithful is to view the $\boldsymbol{y}_i$ and their pairwise distances as defining a finite metric space on n abstract points. Then, a pairwise ϵ-faithful embedding is one that takes this metric space and embeds it with small *distortion* into a lower-dimensional Euclidean metric space. Distortion is a related notion

[5] The pun, I am told, was intended.

to our definition of faithfulness, which you can take to be, given f and $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ (or some other finite metric space on which f acts), defined as

$$\max_{i,j \in [n]} \frac{\big| \|f(\boldsymbol{y}_i) - f(\boldsymbol{y}_j)\| - \|\boldsymbol{y}_i - \boldsymbol{y}_j\| \big|}{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|}. \tag{1.3.14}$$

The Johnson-Lindenstrauss lemma says that any finite metric space on n points in any Euclidean space can be embedded in Euclidean space of dimension $d = O(\frac{\log n}{\epsilon^2})$ with distortion at most ϵ.

From this point of view it is also reasonable to ask about the distortion achievable when embedding other metric spaces, that did not start out Euclidean, into Euclidean space. Fairly strong guarantees are possible in that case also: by *Bourgain's embedding theorem* [Bou85], it is possible to embed *any* finite metric space on n points into some Euclidean space with distortion $O(\log n)$. Combined with the Johnson-Lindenstrauss result, this implies that in fact any finite metric space on n points can be embedded in $\mathbb{R}^{O(\log n)}$ with distortion $O(\log n)$. These abstract-sounding statements actually hahave some quite down-to-earth applications, such as in the analysis of the approximation ratio achieved by a linear programming relaxation of the sparsest cut problem (see, e.g., discussion in the seminal paper [ARV09]).

1.3.5 Nearest Neighbors Algorithms

Finally, let us make a digression to explain the use of the Johnson-Lindenstrauss transform and its variants in applications. Perhaps the main application is to *nearest neighbors (NN)* problems. Here, we are given a set of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^m$ and want to build a data structure that lets us, given a further $\boldsymbol{x} \in \mathbb{R}^m$, output the closest or the k closest $\boldsymbol{x}_i$ to that $\boldsymbol{x}$. It is often acceptable to allow errors, in the sense that we may output any k of the $\boldsymbol{x}_i$ whose distance is within a factor of $1 + \epsilon$ of the kth closest point. This is called an *approximate nearest neighbors (ANN)* problem. We will not go into the specific construction of data structures for NN or ANN, but it is clear that the dimension m is a source of costliness for such data structures. Johnson-Lindenstrauss and its relatives may then be used in a black-box fashion to reduce the dimensionality before applying other techniques, for ANN in particular where small metric distortion is acceptable.

We sketch here how NN and ANN can be used for regression problems. See [Sha13] for the perspective that we will take.

Recall that, in linear regression, we are given $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)$, with $\boldsymbol{x}_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$ (the y_i can also be vectors, but let us assume they are numbers for simplicity). We then want to construct a *predictor* $f : \mathbb{R}^m \to \mathbb{R}$ such that $y_i \approx f(\boldsymbol{x}_i)$. In linear regression, we use $f(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x} \rangle$ for some $\boldsymbol{a} \in \mathbb{R}^m$. (Often one also allows for a constant, $f(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x} \rangle + b$, but we again omit this for simplicity's sake.) We then seek to minimize the ℓ^2 loss:

$$\hat{\boldsymbol{a}} := \arg\min_{\boldsymbol{a}} \sum_{i=1}^{n} (\langle \boldsymbol{a}, \boldsymbol{x}_i \rangle - y_i)^2. \tag{1.3.15}$$

A simple calculation taking the derivatives in each a_i of this shows that the solution is, for $\widehat{\boldsymbol{C}} := \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$, given by

$$\hat{\boldsymbol{a}} = \widehat{\boldsymbol{C}}^{-1} \sum_{i=1}^{n} y_i \boldsymbol{x}_i. \tag{1.3.16}$$

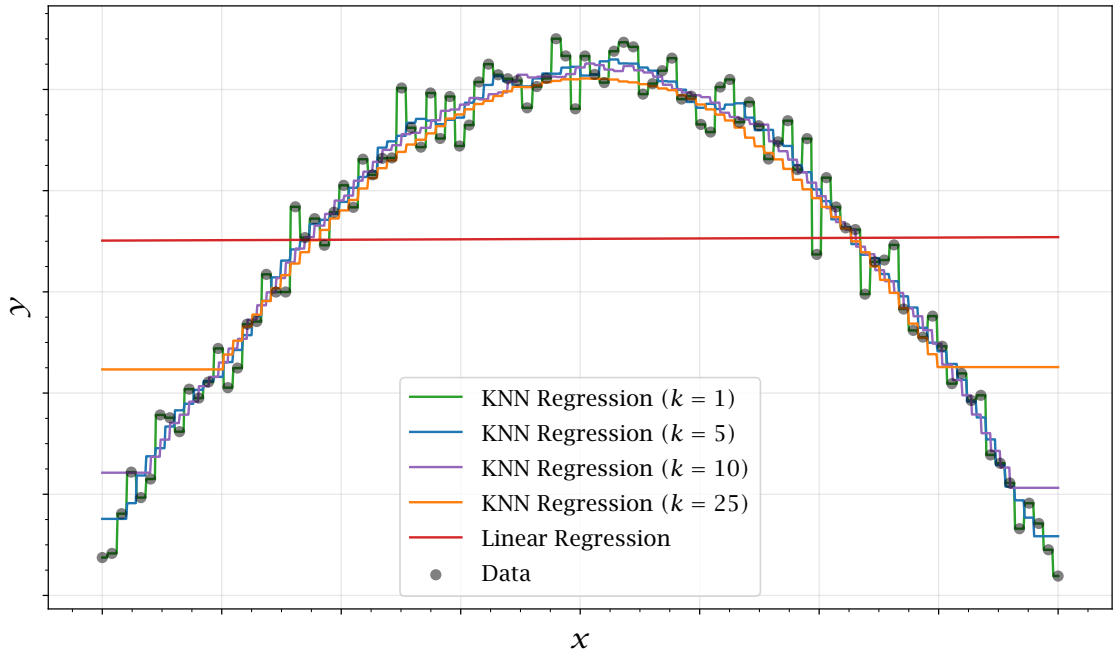


Figure 1.3: An example of linear regression and nearest neighbors regression fits to evaluations of a quadratic function with noise added.

Let us write down what the predictor does with this choice of $\hat{\boldsymbol{a}}$, rewriting slightly:

$$f(\boldsymbol{x}) = \langle \hat{\boldsymbol{a}}, \boldsymbol{x} \rangle = \sum_{i=1}^{n} \langle \widehat{\boldsymbol{C}}^{-1/2} \boldsymbol{x}_i, \widehat{\boldsymbol{C}}^{-1/2} \boldsymbol{x} \rangle y_i. \tag{1.3.17}$$

This predictor may be viewed as taking a weighted sum of the training outcomes y_i to obtain the prediction, where the weight of each is

$$\widehat{w}(\boldsymbol{x}_i, \boldsymbol{x}) := \langle \widehat{\boldsymbol{C}}^{-1/2} \boldsymbol{x}_i, \widehat{\boldsymbol{C}}^{-1/2} \boldsymbol{x} \rangle. \tag{1.3.18}$$

This has some advantages: it is linear, and has the sensible interpretation of giving the "similarity" of $\boldsymbol{x}_i$ to $\boldsymbol{x}$ after "whitening" by multiplying by $\boldsymbol{C}^{-1/2}$ (interpreting $\boldsymbol{C}$ as the sample covariance of the training inputs $\boldsymbol{x}_i$, provided they are centered). On the other hand, some of the pitfalls of linear regression may be viewed as originating from $\widehat{w}(\boldsymbol{x}_i, \boldsymbol{x})$ being symmetric for points near $\boldsymbol{x}$ to those antipodal from $\boldsymbol{x}$ across the origin, since $\widehat{w}(-\boldsymbol{x}_i, \boldsymbol{x}) = -\widehat{w}(\boldsymbol{x}_i, \boldsymbol{x})$. We see in Figure 1.3, for instance, that for this reason the linear function best fitting a parabola is nearly flat.

This perspective (sometimes called one of viewing linear regression as a *smoothing* scheme) suggests that, if we do not much care about the derivation of linear regression, we might consider other schemes of choosing $\widehat{w}(\boldsymbol{x}_i, \boldsymbol{x})$. *Nearest neighbors regression* corresponds to taking

$$\widehat{w}(\boldsymbol{x}_i, \boldsymbol{x}) := \frac{1}{k} \mathbb{1}\{\boldsymbol{x}_i \text{ one of the } k \text{ closest training points to } \boldsymbol{x}\}. \tag{1.3.19}$$

With this choice, $f(\boldsymbol{x})$ will give the weighted average of the outcomes y_i of the k points nearest to $\boldsymbol{x}$. The parameter k controls the amount of smoothing: if $k = n$, then $f(\boldsymbol{x})$ is just a constant and the mean of the observed y_i; if $k = 1$, then $f(\boldsymbol{x})$ is the value of the outcome for the single nearest $\boldsymbol{x}_i$ to $\boldsymbol{x}$, which is a very "jumpy" and irregular function, again as shown in Figure 1.3.

Remark 1.3.7 (Kernel regression)**.** *One may further extend this idea to give the $\boldsymbol{x}_i$ a smoother range of effects on $f(\boldsymbol{x})$, by taking a choice like $\widehat{w}(\boldsymbol{x}_i, \boldsymbol{x}) = \rho(\|\boldsymbol{x}_i - \boldsymbol{x}\|)$ for a* kernel function *ρ. There are numerous ideas to speed up working with related models quite similar in spirit to the Johnson-Lindenstrauss transform; see, e.g., the idea of* random features *or* random kitchen sinks *of [RR07] and the* Fastfood transform *of [LSS13].*

1.4 Spectral Analysis of Wide Gaussian Matrices

We next revisit the Johnson-Lindenstrauss analysis with an eye towards extracting some more mathematical insight. Let us change notation $\boldsymbol{G} := \widehat{\boldsymbol{G}}$.

As we have mentioned, that we are considering the particular pairwise difference vectors $\boldsymbol{y}_i - \boldsymbol{y}_j$ is not essential at all to the claim. The same argument as before in fact gives the following generalization.

Theorem 1.4.1. *There is a constant $C > 0$ such that the following holds. For any $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^m$, if $d \geq C\frac{\log N}{\epsilon^2}$ and $\boldsymbol{G} \sim \mathcal{N}(0, \frac{1}{d})^{\otimes d \times m}$,*

$$\mathbb{P}\left[\, \left| \|\boldsymbol{G}\boldsymbol{x}_i\|^2 - \|\boldsymbol{x}_i\|^2 \right| \leq \epsilon \|\boldsymbol{x}_i\|^2 \text{ for all } i \in [N] \right] \geq 1 - \frac{1}{N}. \tag{1.4.1}$$

In our application we took $N = \binom{n}{2}$ and the $\boldsymbol{x}_i$ to be the $\boldsymbol{y}_j - \boldsymbol{y}_k$, but the same exact arguments works for the above as well.

We may reframe this by writing

$$\left| \|\boldsymbol{G}\boldsymbol{x}_i\|^2 - \|\boldsymbol{x}_i\|^2 \right| = \left| \boldsymbol{x}_i^\top (\boldsymbol{G}^\top \boldsymbol{G}) \boldsymbol{x}_i - \boldsymbol{x}_i^\top \boldsymbol{I}_m \boldsymbol{x}_i \right|. \tag{1.4.2}$$

Thus it seems that, from the "point of view" of a small number of deterministic quadratic form evaluations, the random matrix $\boldsymbol{G}^\top \boldsymbol{G}$ behaves like the identity. This of course cannot be true without restricting the number of quadratic form evaluations, since the former matrix has rank at most $d \ll m$. We will next look for a more sophisticated and "more spectral" explanation of this phenomenon than what we saw before.

To do that, we will undertake an analysis of the eigendecomposition of the matrix $\boldsymbol{G}^\top \boldsymbol{G} \in \mathbb{R}^{m \times m}_{\mathsf{sym}}$, seeking to make claims about the distribution of its eigenvectors and eigenvalues. 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 $\lambda_{d+1} = \cdots = \lambda_n = 0$. The remaining eigenvalues are related to the singular values of $\boldsymbol{G}$ by $\lambda_i = \sigma_i(\boldsymbol{G})^2$.

1.4.1 Eigenvectors

Geometric considerations about the set of matrices with repeated singular values give the following.

Lemma 1.4.2. *Almost surely (with probability 1),* $\lambda_1 > \cdots > \lambda_d > 0$.

Proof. We will use that $\lambda_i = \lambda_i(\boldsymbol{G}^\top \boldsymbol{G}) = \lambda_i(\boldsymbol{G}\boldsymbol{G}^\top)$, where the latter matrix is $d \times d$. To show that $\lambda_d > 0$, it suffices to check that $p(\boldsymbol{G}) = \mathsf{det}(\boldsymbol{G}\boldsymbol{G}^\top)$ is a non-zero polynomial (which you may do by exhibiting a single $\boldsymbol{G}$ for which $p(\boldsymbol{G}) \neq 0$). Then, since $\boldsymbol{G}$ has a smooth multivariate density and $p(\boldsymbol{G})$ is a smooth function, $p(\boldsymbol{G})$ also has a density (i.e., its law is absolutely continuous to Lebesgue measure). In particular,

$$\mathbb{P}[\lambda_d = 0] = \mathbb{P}[\mathsf{det}(\boldsymbol{G}\boldsymbol{G}^\top) = 0] = \mathbb{P}[p(\boldsymbol{G}) = 0] = 0. \tag{1.4.3}$$

For the other claim, we will proceed similarly but must construct a subtler polynomial. Recall the characteristic polynomial of a matrix:

$$\mathsf{det}(t\boldsymbol{I}_d - \boldsymbol{G}\boldsymbol{G}^\top) = \sum_{k=0}^{d} (-1)^{d-k} s_{d-k}(\boldsymbol{G}) t^k, \tag{1.4.4}$$

where on the one hand the coefficients $s_{d-k}(\boldsymbol{G})$ are polynomials in the entries of $\boldsymbol{G}$, but on the other hand are given by the elementary symmetric polynomials in λ_i:

$$s_k(\boldsymbol{G}) = \sum_{1 \leq i_1 < \cdots < i_k \leq d} \lambda_{i_1} \cdots \lambda_{i_k}. \tag{1.4.5}$$

For instance, $s_0(\boldsymbol{G}) = 1$, $s_1(\boldsymbol{G}) = \mathsf{Tr}(\boldsymbol{G}\boldsymbol{G}^\top) = \sum_{i=1}^d \lambda_i$, and $s_d(\boldsymbol{G}) = \mathsf{det}(\boldsymbol{G}\boldsymbol{G}^\top) = \prod_{i=1}^d \lambda_i$.

Now, consider the following quantity:

$$p(\boldsymbol{G}) := \prod_{1 \leq i < j \leq d} (\lambda_i - \lambda_j)^2. \tag{1.4.6}$$

Visibly $p(\boldsymbol{G})$ is a symmetric polynomial in the λ_i. Therefore, by the fundamental theorem of symmetric polynomials, $p(\boldsymbol{G})$ is a polynomial of the $s_0(\boldsymbol{G}), \ldots, s_d(\boldsymbol{G})$, and therefore is itself a polynomial in the entries of $\boldsymbol{G}$. On the other hand, $p(\boldsymbol{G}) = 0$ if and only if two of the λ_i are the same. Thus, by the same argument as before,

$$\mathbb{P}[\lambda_i = \lambda_j \text{ for some } i \neq j] = \mathbb{P}[p(\boldsymbol{G}) = 0] = 0, \tag{1.4.7}$$

completing the proof. □

Thus we may speak of the spans of the unit eigenvectors, $L_i := \mathsf{span}(\{\boldsymbol{v}_i\}) \subset \mathbb{R}^m$, which are uniquely associated to the eigenvalues $\lambda_1, \ldots, \lambda_d$ in the spectral decomposition

$$\boldsymbol{G}^\top \boldsymbol{G} = \sum_{i=1}^{d} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top. \tag{1.4.8}$$

The $\boldsymbol{v}_i$ themselves, unfortunately, are not uniquely determined by λ_i, since $\boldsymbol{v}_i$ can be replaced by $-\boldsymbol{v}_i$ in the spectral decomposition without affecting it. Let us also write $\boldsymbol{L}(\boldsymbol{G}) = (L_1(\boldsymbol{G}), \ldots, L_d(\boldsymbol{G}))$. You might find this to be an unusual object: an ordered tuple of lines in $\mathbb{R}^m$. Think of it this way: there is a more conventional object called the *Stiefel manifold*, given by

$$\mathsf{Stief}(m, d) := \{\boldsymbol{V} \in \mathbb{R}^{m \times d} : \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_d\}, \tag{1.4.9}$$

which is just a subset of $m \times d$ matrices, those whose columns are orthonormal. Note that

$$\mathsf{Stief}(m, m) = \mathcal{O}(m). \tag{1.4.10}$$

A collection of orthogonal lines can be viewed as an equivalence class of 2^d elements in $\mathsf{Stief}(m, d)$, where the equivalence class of $\boldsymbol{V} = [\, \boldsymbol{v}_1 \cdots \boldsymbol{v}_d \,]$ consists of all $[\, \pm\boldsymbol{v}_1 \cdots \pm \boldsymbol{v}_d \,]$. We may speak (as is intuitively obvious but maybe mathematically a bit obscure) of applying an orthogonal matrix to such a collection of lines, where $\boldsymbol{Q}$ maps the equivalence class of $\boldsymbol{V}$ to that of $\boldsymbol{QV}$, as we will use below.

Straightforward Gaussian calculations give the following.

Proposition 1.4.3. *If $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_m)$ and $\boldsymbol{Q} \in \mathcal{O}(m)$, then*

$$\mathsf{Law}(\boldsymbol{Qg}) = \mathsf{Law}(\boldsymbol{g}) = \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_m). \tag{1.4.11}$$

Proof. By Proposition 1.2.2 or direct calculation of the first two moments. □

Corollary 1.4.4. *If $\boldsymbol{G} \sim \mathcal{N}(0, \sigma^2)^{\otimes d \times m}$ and $\boldsymbol{Q} \in \mathcal{O}(m)$, then*

$$\mathsf{Law}(\boldsymbol{GQ}) = \mathsf{Law}(\boldsymbol{G}) = \mathcal{N}(0, \sigma^2)^{\otimes d \times m}. \tag{1.4.12}$$

Proof. Use that $\boldsymbol{Q}$ acts separately on the independent rows of $\boldsymbol{G}$ in forming $\boldsymbol{GQ}$. □

Next we have a deterministic claim about frames of lines given by eigendirections of a matrix.

Proposition 1.4.5. *$\boldsymbol{L}(\boldsymbol{GQ}) = \boldsymbol{Q}^\top \boldsymbol{L}(\boldsymbol{G})$, where $\boldsymbol{Q}$ acts on a frame of lines in the sense discussed above.*

Proof. Use that $(\boldsymbol{GQ})^\top (\boldsymbol{GQ}) = \boldsymbol{Q}^\top \boldsymbol{G}^\top \boldsymbol{GQ} = \sum_{i=1}^d \lambda_i (\boldsymbol{Q}^\top \boldsymbol{v}_i)(\boldsymbol{Q}^\top \boldsymbol{v}_i)^\top$. □

Putting the pieces together, we find:

Corollary 1.4.6. *For all $\boldsymbol{Q} \in \mathcal{O}(m)$, $\mathsf{Law}(\boldsymbol{QL}(\boldsymbol{G})) = \mathsf{Law}(\boldsymbol{L}(\boldsymbol{G}))$.*

Proof. Using the above results, $\mathsf{Law}(\boldsymbol{L}(\boldsymbol{G})) = \mathsf{Law}(\boldsymbol{L}(\boldsymbol{GQ}^\top)) = \mathsf{Law}(\boldsymbol{QL}(\boldsymbol{G}))$. □

What the above establishes is that the collection of lines $\boldsymbol{L}(\boldsymbol{G})$ has a orthogonally invariant distribution (one unchanged by the action of any orthogonal matrix). By choosing uniformly at random from the equivalence class of $\mathsf{Stief}(m, d)$ associated to $\boldsymbol{L}(\boldsymbol{G})$, you may lift this up to a orthogonally invariant distribution on $\mathsf{Stief}(m, d)$ (a distribution on eigenvectors of $\boldsymbol{G}^\top \boldsymbol{G}$, where we handle the sign ambiguity by choosing either $\boldsymbol{v}_i$ or $-\boldsymbol{v}_i$ as the eigenvector of λ_i each with probability $1/2$). Now comes the key further result, saying that this *completely* determines the law of $\boldsymbol{L}(\boldsymbol{G})$.

Theorem 1.4.7 (Haar)**.** *There is a unique orthogonally invariant probability measure on each $\mathsf{Stief}(m, d)$, called the* Haar measure *and which we will denote $\mathsf{Haar}(\mathsf{Stief}(m, d))$. Either of the two procedures below yields a sample from this measure:*

1. *Draw $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_d \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$ independently and perform the Gram-Schmidt procedure on them to obtain $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ orthonormal, forming $\boldsymbol{V} = [\, \boldsymbol{v}_1 \cdots \boldsymbol{v}_d \,] \in \mathsf{Stief}(m, d)$.*

2. *Draw* $\boldsymbol{v}_1 \sim \mathsf{Unif}(\mathbb{S}^{m-1}(1))$. *Then, for* $i = 2, \dots, d$, *draw* $\boldsymbol{v}_i \sim \mathsf{Unif}(\mathbb{S}^{m-1}(1) \cap \{\boldsymbol{v} : \langle \boldsymbol{v}_1, \boldsymbol{v} \rangle = \cdots = \langle \boldsymbol{v}_{i-1}, \boldsymbol{v} \rangle = 0\})$.

Proof. You may check that the latter two procedures both yield a matrix with the same orthogonally invariant law, so it suffices to prove uniqueness. We will give the proof for $\mathsf{Stief}(m, m) = \mathcal{O}(m)$, from which the case of general $d \le m$ follows without too much trouble. Suppose μ and ν are two orthogonally invariant probability measures on $\mathcal{O}(m)$. Then, $\frac{1}{2}\mu + \frac{1}{2}\nu$ (i.e., assigning measure $\frac{1}{2}\mu(A) + \frac{1}{2}\nu(A)$ to each measurable set A) is another orthogonally invariant probability measure. Moreover, μ is absolutely continuous with respect to this measure, and thus has a density f with respect to it. That is:

$$\begin{aligned}\int_{\mathcal{O}(m)} g(\boldsymbol{Q}) d\mu(\boldsymbol{Q}) &= \int_{\mathcal{O}(m)} f(\boldsymbol{Q}) g(\boldsymbol{Q}) d\left(\frac{1}{2}\mu + \frac{1}{2}\nu\right)(\boldsymbol{Q}) \\ &= \frac{1}{2}\int_{\mathcal{O}(m)} f(\boldsymbol{Q}) g(\boldsymbol{Q}) d\mu(\boldsymbol{Q}) + \frac{1}{2}\int_{\mathcal{O}(m)} f(\boldsymbol{Q}) g(\boldsymbol{Q}) d\nu(\boldsymbol{Q}). \end{aligned} \tag{1.4.13}$$

You may check, and it should be intuitive, that f is then an orthogonally invariant function, i.e., having $f(\boldsymbol{QR}) = f(\boldsymbol{R})$ for any $\boldsymbol{Q}, \boldsymbol{R} \in \mathcal{O}(m)$. But then f must be a constant, and since it is a probability density we must have $f(\boldsymbol{Q}) = 1$. The above then implies that $\mu = \nu$. □

In summary, we have exactly characterized the law of the eigenvectors of $\boldsymbol{G}^\top \boldsymbol{G}$ having positive eigenvalue: there are exactly d of them, associated to distinct eigenvalues, and having Haar distribution in the Stiefel manifold $\mathsf{Stief}(m, d)$. In particular, their span is a "uniformly random" d-dimensional subspace of $\mathbb{R}^m$ (this is a perhaps more intuitive object, but its meaning is just the span of a Haar-distributed orthonormal basis from the Stiefel manifold).

1.4.2 Eigenvalues

To understand the eigenvalues of $\boldsymbol{G}^\top \boldsymbol{G}$, we will again instead work with those of $\boldsymbol{G}\boldsymbol{G}^\top$. Let the columns of $\boldsymbol{G}$ be $\boldsymbol{g}_1, \dots, \boldsymbol{g}_m \sim \mathcal{N}(\boldsymbol{0}, \frac{1}{d}\boldsymbol{I}_d)$. Intuitively, we should expect a law of large numbers to hold:

$$\boldsymbol{G}\boldsymbol{G}^\top = \sum_{i=1}^m \boldsymbol{g}_i \boldsymbol{g}_i^\top \overset{\text{(LLN)}}{\approx} m \cdot \mathbb{E}\boldsymbol{g}_1 \boldsymbol{g}_1^\top = \frac{m}{d}\boldsymbol{I}_d. \tag{1.4.14}$$

This heuristic statement is our first example of a *matrix concentration inequality*.

Consider the matrix $\boldsymbol{M} := \boldsymbol{G}\boldsymbol{G}^\top - \frac{m}{d}\boldsymbol{I}_d$. We will aim to show that $\|\boldsymbol{M}\|$ is relatively small. How do we move towards such a result? The trouble is that the operator norm $\|\boldsymbol{M}\|$ is a continuous quantity:

$$\|\boldsymbol{M}\| = \sup_{\boldsymbol{x} \in \mathbb{S}^{d-1}(1)} |\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}|. \tag{1.4.15}$$

If $\mathbb{S}^{d-1}(1)$ were a finite set, we could bound $\mathbb{P}[|\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}| > t]$ and use a union bound, much as for the Johnson-Lindenstrauss lemma, but this is not the case.

Fortunately, there is a powerful tool from real analysis and metric geometry that lets us import our discrete tools to such continuous settings.

Definition 1.4.8 (ϵ-net)**.** *A set* $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \mathcal{Y} \subset \mathbb{R}^d$ *is an* ϵ-net of $\mathcal{Y}$ *if, for all* $\boldsymbol{y} \in \mathcal{Y}$, *there is* $\boldsymbol{x}_i \in \mathcal{X}$ *such that* $\|\boldsymbol{y} - \boldsymbol{x}_i\| \le \epsilon$.

Another evocative term for an ϵ-net is an *ϵ-covering*: writing $B(\boldsymbol{x}, \epsilon)$ for a closed ball of radius ϵ around $\boldsymbol{x}$, the union of the balls $B(\boldsymbol{x}_i, \epsilon)$ contains all of $\mathcal{Y}$.

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

Lemma 1.4.9. *Suppose $\boldsymbol{M} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ and $\mathcal{X}$ is an ϵ-net of $\mathbb{S}^{d-1}(1)$ for $\epsilon < \frac{1}{2}$. Write $K := \max_{\boldsymbol{x}_i \in \mathcal{X}} |\boldsymbol{x}_i^\top \boldsymbol{M} \boldsymbol{x}_i|$. Then,*

$$K \le \|\boldsymbol{M}\| \le \frac{1}{1-2\epsilon} K. \tag{1.4.16}$$

Proof. The first inequality is immediate. For the second, let $\boldsymbol{x} \in \mathbb{S}^{d-1}(1)$ 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

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

and rearranging gives the result. □

To use this, we need two pieces of information: first, a small ϵ-net, and second, a bound on the probability of large values of $|\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}|$. There is a nice theory relating ϵ-nets and *packings*, which we describe here to address the first point.

Definition 1.4.10. *A set $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \mathcal{Y} \subseteq \mathbb{R}^d$ for $\mathcal{Y}$ a metric space is an* ϵ-packing *if the balls $B(\boldsymbol{x}_i, \epsilon)$ are pairwise disjoint.*

Lemma 1.4.11. *A maximal (under inclusion of sets) ϵ-packing is a 2ϵ-net.*

Proof. Suppose that $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \mathcal{Y}$ is a maximal ϵ-packing, and let $\boldsymbol{x} \in \mathcal{Y} \setminus \mathcal{X}$. Adding $\boldsymbol{x}$ to $\mathcal{X}$ must result in a set that is not an ϵ-packing. Thus, there exists $\boldsymbol{x}_i$ such that $B(\boldsymbol{x}, \epsilon) \cap B(\boldsymbol{x}_i, \epsilon) \neq \emptyset$. By the triangle inequality, $\|\boldsymbol{x} - \boldsymbol{x}_i\| \le 2\epsilon$. Since this holds for any $\boldsymbol{x} \notin \mathcal{X}$, the result follows. □

Actually, the relationship between sizes of packings and nets (or coverings) goes in both directions.

Lemma 1.4.12. *Let $\mathcal{X}^{(p)}$ be any ϵ-packing and $\mathcal{X}^{(n)}$ be any ϵ-net. Then, $|\mathcal{X}^{(p)}| \le |\mathcal{X}^{(n)}|$.*

Proof. Any $\boldsymbol{x}_i^{(p)} \in \mathcal{X}^{(p)}$ must belong to some $B(\boldsymbol{x}_j^{(n)}, \epsilon)$ for $\boldsymbol{x}_j^{(n)} \in \mathcal{X}^{(n)}$. On the other hand, any two $\boldsymbol{x}_i^{(p)}, \boldsymbol{x}_{i'}^{(p)}$ are at distance greater than 2ϵ, and thus cannot belong to the same $B(\boldsymbol{x}_j^{(n)}, \epsilon)$, whose diameter is 2ϵ. □

Corollary 1.4.13. *For a given $\mathcal{Y}$, let $N(\epsilon)$ be the minimum possible size of an ϵ-net and $P(\epsilon)$ be the maximum possible size of an ϵ-packing. Then,*

$$P(\epsilon) \le N(\epsilon) \le P(\epsilon/2). \tag{1.4.17}$$

What is especially convenient for our purposes is that *volumetric* arguments may be used to control maximal packings, allowing us to argue abstractly that small ϵ-nets exist. Below is a standard example.

Lemma 1.4.14. *For any $\epsilon > 0$, there is an ϵ-net $\mathcal{X}$ of $B(\mathbf{0}, 1) \subset \mathbb{R}^d$ with $|\mathcal{X}| \leq (1 + 2/\epsilon)^d$.*

Proof. Let μ denote the Lebesgue measure. We have for any $\boldsymbol{x}$ that $\mu(B(\boldsymbol{x}, r)) = r^d \mu(B(\mathbf{0}, 1))$. And, if $\boldsymbol{x} \in B(\mathbf{0}, 1)$, then $B(\boldsymbol{x}, \epsilon) \subseteq B(\mathbf{0}, 1 + \epsilon)$. Thus, ff $\mathcal{X}$ is a $\frac{\epsilon}{2}$-packing of $B(\mathbf{0}, 1)$, we must have

$$|\mathcal{X}| \leq 1 + \frac{\mu(B(\mathbf{0}, 1 + \frac{\epsilon}{2}))}{\mu(B(\mathbf{0}, \frac{\epsilon}{2}))} = \left(1 + \frac{2}{\epsilon}\right)^d. \tag{1.4.18}$$

Thus there exists a maximal $\frac{\epsilon}{2}$-packing of at most this size, which is also an ϵ-net. □

We can use this to obtain (rather sub-optimal) nets of the sphere as follows.

Proposition 1.4.15. *Suppose $\mathcal{X}$ is an ϵ-net of $B(\mathbf{0}, 1) \subset \mathbb{R}^d$ with $0 < \epsilon < 1/2$. Let $\widehat{\mathcal{X}} := \{\boldsymbol{x}/\|\boldsymbol{x}\| : \boldsymbol{x} \in \mathcal{X}\}$. Then, $\widehat{\mathcal{X}}$ is a $2\sqrt{\epsilon}$-net of $\mathbb{S}^{d-1}(1)$.*

Proof. Suppose $\boldsymbol{y} \in \mathbb{S}^{d-1}(1)$. There is $\boldsymbol{x} \in \mathcal{X}$ such that $\|\boldsymbol{y} - \boldsymbol{x}\| \leq \epsilon$. Let $\widehat{\boldsymbol{x}} := \boldsymbol{x}/\|\boldsymbol{x}\| \in \widehat{\mathcal{X}}$. We first make a few preliminary observations:

$$\begin{aligned}
\|\boldsymbol{x}\| &= \|\boldsymbol{y} + (\boldsymbol{x} - \boldsymbol{y})\| \\
&\geq \|\boldsymbol{y}\| - \|\boldsymbol{x} - \boldsymbol{y}\| \\
&\geq 1 - \epsilon, \\
\|\boldsymbol{x}\|^2 &\geq 1 - 2\epsilon + \epsilon^2 \\
&\geq 1 - 2\epsilon, \\
\frac{1}{4} &\geq \epsilon^2 \\
&\geq \|\boldsymbol{y} - \boldsymbol{x}\|^2 \\
&= 1 - 2\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \|\boldsymbol{x}\|^2 \\
&\geq 1 - 2\langle \boldsymbol{x}, \boldsymbol{y}\rangle + (1 - 2\epsilon),
\end{aligned}$$

and from this last observation we rearrange to find

$$\begin{aligned}
\langle \boldsymbol{x}, \boldsymbol{y}\rangle &\geq \frac{7}{8} - \epsilon \\
&> 0.
\end{aligned}$$

Now, we may bound:

$$\begin{aligned}
\|\boldsymbol{y} - \widehat{\boldsymbol{x}}\|^2 &= 2 - 2\langle \widehat{\boldsymbol{x}}, \boldsymbol{y}\rangle \\
&= 2 - \frac{2\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\|\boldsymbol{x}\|} \\
&\leq 2 - 2\langle \boldsymbol{x}, \boldsymbol{y}\rangle \\
&= 1 - 2\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \|\boldsymbol{x}\|^2 + (1 - \|\boldsymbol{x}\|^2) \\
&= \|\boldsymbol{y} - \boldsymbol{x}\|^2 + (1 - \|\boldsymbol{x}\|^2) \\
&\leq \epsilon^2 + 2\epsilon \\
&\leq 4\epsilon,
\end{aligned}$$

and the result follows. □

Corollary 1.4.16. *For any $\epsilon > 0$, there is an ϵ-net $\mathcal{X}$ of $\mathbb{S}^{d-1}(1) \subset \mathbb{R}^d$ with $|\mathcal{X}| \leq (1 + 8/\epsilon^2)^d$.*

This is actually very suboptimal; the true correct scaling of the smallest ϵ-net for small ϵ is like $d\epsilon^{-(d-1)}$, as you can show by a somewhat more complicated volumetric argument involving spherical surface areas rather than volumes. However, these details will not be important to us since we will only care about taking ϵ a small constant and it will suffice to have an ϵ-net of size at most $C(\epsilon)^d$ for some arbitrarily large $C(\epsilon) > 0$. Thus these observations conclude our construction of an adequately small ϵ-net.

For our second task, we can actually control the deviation probabilities of the individual quadratic forms with the same tools we have developed already.

Proposition 1.4.17. *For any $\boldsymbol{x} \in \mathbb{S}^{d-1}(1)$ and $t \leq \frac{m}{d}$,*

$$\mathbb{P}\left[|\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}| > t\right] \leq 2\exp\left(-\frac{d^2t^2}{8m}\right). \tag{1.4.19}$$

Proof. Note that $|\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}| = \sum_{i=1}^m \langle \boldsymbol{g}_i, \boldsymbol{x}\rangle^2 - \frac{m}{d} \overset{\text{(law)}}{=} \frac{\|\boldsymbol{x}\|^2}{d}(\sum_{i=1}^m h_i^2 - m)$ where $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$. This is precisely the setting treated by Lemma 1.3.3, which gives the result. □

We are now ready to prove our main statement on the concentration of eigenvalues of $\boldsymbol{G}\boldsymbol{G}^\top$ (equivalently, as we formulate below, of the singular values of $\boldsymbol{G}$).

Theorem 1.4.18. *Let $\boldsymbol{G} \sim \mathcal{N}(0, \frac{1}{d})^{\otimes d\times m}$. Then, there are absolute constants $c, C_1, C_2 > 0$ such that, if $m \geq C_1 d$, then*

$$\mathbb{P}\left[\left\|\boldsymbol{G}\boldsymbol{G}^\top - \frac{m}{d}\boldsymbol{I}_d\right\| \geq C_2\sqrt{\frac{m}{d}}\right] \leq \exp(-cd). \tag{1.4.20}$$

In other words, if instead $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d\times m}$, then we have

$$\mathbb{P}\left[||\boldsymbol{G}\boldsymbol{G}^\top - m\boldsymbol{I}_d|| \geq C_2\sqrt{md}\right] \leq \exp(-cd), \tag{1.4.21}$$

and, for possibly different constants $c, C_1, C_2 > 0$ and provided that $m \geq C_1 d$, we have

$$\mathbb{P}\left[\sqrt{m} - C_2\sqrt{d} \leq \sigma_d(\boldsymbol{G}) \leq \cdots \leq \sigma_1(\boldsymbol{G}) \leq \sqrt{m} + C_2\sqrt{d}\right] \geq 1 - \exp(-cd). \tag{1.4.22}$$

Proof. Write $\boldsymbol{M} := \boldsymbol{G}\boldsymbol{G}^\top - \frac{m}{d}\boldsymbol{I}_d$. Fix $\epsilon = 1/4$. Let $C > 0$ be such that there is an ϵ-net $\mathcal{X}$ of $\mathbb{S}^{d-1}(1)$ of size at most C^d for all d. Write $K := \max_{\boldsymbol{x}\in\mathcal{X}} |\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}|$. By Lemma 1.4.9, $\|\boldsymbol{M}\| \leq 2K$.

Then, by the union bound and Proposition 1.4.17,

$$\begin{aligned}
\mathbb{P}[\|\boldsymbol{M}\| \geq t] &\leq \mathbb{P}\left[K \geq \frac{t}{2}\right] \\
&\leq \sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}\left[|\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}| \geq \frac{t}{2}\right] \\
&\leq 2 \cdot |\mathcal{X}| \cdot \exp\left(-\frac{d^2t^2}{8m}\right) \\
&= 2\exp\left(\log C \cdot d - \frac{d^2t^2}{8m}\right) \\
&= 2\exp\left(-d\left(\frac{d}{8m}t^2 - \log C\right)\right),
\end{aligned}$$

and the result follows if we choose the constants in the statement correctly and take $t := C_2\sqrt{m/d}$. Note a technicality: in order to use Proposition 1.4.17, we must have $t/2 \le m/d$ or $t \le 2m/d$. Thus we need $C_2\sqrt{m/d} \le 2m/d$, or $m \ge \frac{C_2^2}{4}d$, which is where the constraint on m in the statement comes from. □

Remark 1.4.19. *In fact it can be shown that the typical size of the extreme singular values of $\boldsymbol{G}$ with $m \ge d$ is $\sqrt{m} \pm \sqrt{d}$, i.e., the "right" constant above is $C_2 = 1$. We will see some versions of this much more precise claim soon. For now, take this as a useful simple expression to remember the typical scaling of the extreme singular values of a matrix of standard Gaussians: the square root of the larger dimension plus/minus the square root of the smaller.*

1.4.3 Random Projection Analogy

When $m \gg d$ (say, along a sequence of d growing with $m = m(d)$), the above implies that $\boldsymbol{G}^\top\boldsymbol{G}$ is close to $\frac{m}{d}\boldsymbol{P}$, where $\boldsymbol{P}$ is a projection matrix to a uniformly random d-dimensional subspace of $\mathbb{R}^m$. Let us briefly see why this demystifies how $\boldsymbol{G}^\top\boldsymbol{G}$ preserves quadratic forms that are chosen obliviously to its randomness.

Consider $\boldsymbol{x} \in \mathbb{R}^m$, say with $\|\boldsymbol{x}\| = 1$. We expect from the above intuition that $\boldsymbol{x}^\top\boldsymbol{G}^\top\boldsymbol{G}\boldsymbol{x}$ behaves like $\boldsymbol{x}^\top(\frac{m}{d}\boldsymbol{P})\boldsymbol{x} = \frac{m}{d}\|\boldsymbol{P}\boldsymbol{x}\|^2$. Note that $\boldsymbol{P}$ has the same law as $\boldsymbol{Q}\boldsymbol{P}^{(0)}\boldsymbol{Q}^\top$, where $\boldsymbol{Q} \sim \mathsf{Haar}(\mathcal{O}(m))$ while $\boldsymbol{P}^{(0)}$ is the projection to the first d coordinates, i.e., to $\mathsf{span}(\boldsymbol{e}_1, \dots, \boldsymbol{e}_d)$. Then, we have

$$\frac{m}{d}\|\boldsymbol{P}\boldsymbol{x}\|^2 \overset{\text{(law)}}{=} \frac{m}{d}\|\boldsymbol{Q}\boldsymbol{P}^{(0)}\boldsymbol{Q}^\top\boldsymbol{x}\|^2 = \frac{m}{d}\|\boldsymbol{P}^{(0)}\boldsymbol{y}\|^2 = \frac{m}{d}\sum_{i=1}^{d} y_i^2, \tag{1.4.23}$$

where $\boldsymbol{y} = \boldsymbol{Q}^\top\boldsymbol{x}$ then has the law $\mathsf{Unif}(\mathbb{S}^{m-1}(1))$. Standard concentration arguments like we have done before (if you want to be very precise, you can use that $\boldsymbol{y}$ further has the law of $\boldsymbol{g}/\|\boldsymbol{g}\|$ for $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$) then imply that the above is close to 1 with high probability.

1.5 Application: Compressed Sensing

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

For general $\boldsymbol{x} \in \mathbb{R}^m$, we need $\boldsymbol{G}$ to be injective, which requires $d \ge m$. 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.

1.5.1 Null Space and Restricted Isometry Properties

Let us denote sparsity by

$$\|\boldsymbol{x}\|_0 := \#\{i \in [m] : x_i \ne 0\}. \tag{1.5.1}$$

Note that this "ℓ^0-norm" is not actually a norm.

Assuming that $\|\boldsymbol{x}\|_0 \le k$ 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 1.5.1. *We say that* $\boldsymbol{G}$ distinguishes k-sparse vectors *if* $\boldsymbol{G}\boldsymbol{x} \neq \boldsymbol{G}\boldsymbol{x}'$ *for any* $\boldsymbol{x} \neq \boldsymbol{x}'$ *with* $\|\boldsymbol{x}\|_0, \|\boldsymbol{x}'\|_0 \le k$.

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

Proposition 1.5.3. $\boldsymbol{G}$ *distinguishes* k*-sparse vectors if and only if* $\boldsymbol{G}$ *has the* $2k$*-NSP.*

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

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

Definition 1.5.4 (Restricted isometry property)**.** *We say that* $\boldsymbol{G}$ *has the* (k, δ)-restricted isometry property ((k, δ)-RIP) *if, for all* $\boldsymbol{x} \neq \boldsymbol{0}$ *with* $\|\boldsymbol{x}\|_0 \le k$,

$$(1 - \delta)\|\boldsymbol{x}\|^2 \le \|\boldsymbol{G}\boldsymbol{x}\|^2 \le (1 + \delta)\|\boldsymbol{x}\|^2. \tag{1.5.2}$$

The following implication is immediate.

Proposition 1.5.5. *If* $\boldsymbol{G}$ *has the* (k, δ)*-RIP for any* $\delta < 1$*, then* $\boldsymbol{G}$ *has the* k*-NSP.*

The RIP is more useful than the NSP both for analyzing compressed sensing when some noise is further added to $\boldsymbol{y} = \boldsymbol{G}\boldsymbol{x}$, and for analyzing algorithms for recovering $\boldsymbol{x}$.

1.5.2 Random Sensing Matrices

The following is the main result that we will show.

Theorem 1.5.6. *For any $k \geq 1$ and $\delta \in (0,1)$, there exists $C = C(\delta) > 0$ such that, if $d \geq Ck \log m$ and $\boldsymbol{G} \sim \mathcal{N}(0, \frac{1}{d})^{\otimes d \times m}$, then*

$$\mathbb{P}[\boldsymbol{G} \text{ has the } (k,\delta)\text{-RIP}] \geq 1 - \frac{2}{m^k}. \tag{1.5.3}$$

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

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 $\|\boldsymbol{x}\|_0 \leq k$ and the non-zero indices of $\boldsymbol{x}$ are contained in such S, then

$$\|\boldsymbol{x}\|^2 = \|\boldsymbol{x}^{(S)}\|^2, \tag{1.5.4}$$

$$\boldsymbol{G}\boldsymbol{x} = \boldsymbol{G}^{(S)}\boldsymbol{x}^{(S)}. \tag{1.5.5}$$

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

$$(1-\delta)\|\boldsymbol{x}^{(S)}\|^2 \leq \|\boldsymbol{G}^{(S)}\boldsymbol{x}^{(S)}\|^2 \leq (1+\delta)\|\boldsymbol{x}^{(S)}\|^2. \tag{1.5.6}$$

But this just amounts to asking that

$$1 - \delta \leq \lambda_k(\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)}) \leq \lambda_1(\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)}) \leq 1 + \delta, \tag{1.5.7}$$

or again equivalently that

$$\|\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)} - \boldsymbol{I}_k\| \leq \delta. \tag{1.5.8}$$

We will then proceed by union bounding,

$$\mathbb{P}[\boldsymbol{G} \text{ does not have the } (k,\delta)\text{-RIP}] \leq \binom{m}{k} \mathbb{P}[\|\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)} - \boldsymbol{I}_k\| > \delta], \tag{1.5.9}$$

where all probabilities coming from the union bound are the same since the $\boldsymbol{G}^{(S)}$ are identically distributed regardless of the choice of S.

This is almost exactly the kind of deviation that we bounded before in Theorem 1.4.18. First, let's decode our statement into that form. Introduce $\boldsymbol{H} \sim \mathcal{N}(0, \frac{1}{k})^{\otimes k \times d}$, the scaling that Theorem 1.4.18 addresses. We have $\boldsymbol{G}^{(S)} \overset{\text{(law)}}{=} \sqrt{\frac{k}{d}}\boldsymbol{H}^\top$. Thus,

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

The only difference between this and what Theorem 1.4.18 covered is that there we were concerned with deviations of the order $O(\sqrt{d/k})$, while here we are concerned with $O(d/k)$, which is much larger since $d \gg k$ under our assumption. But, we may still repeat the approach of that proof: revisiting that argument and setting $t = \delta\frac{d}{k}$, we notice that Lemma 1.3.3 applies the same way since $\delta < 1$, and we get that the above is bounded by, for an absolute constant $C' > 0$ involving the size of an ϵ-net,

$$\begin{aligned} &\leq 2\exp\left(-k\left(\frac{k}{8d}t^2 - \log C'\right)\right) \\ &= 2\exp\left(-k\left(\frac{k}{8d}t^2 - \log C'\right)\right) \\ &= 2\exp\left(-k\left(\delta^2\frac{d}{8k} - \log C'\right)\right) \end{aligned}$$

and choosing C in the statement sufficiently large, we may ensure, since $d \geq Ck\log m$, that

$$\leq 2\exp\left(-2k\log m\right). \tag{1.5.10}$$

Finally, we have

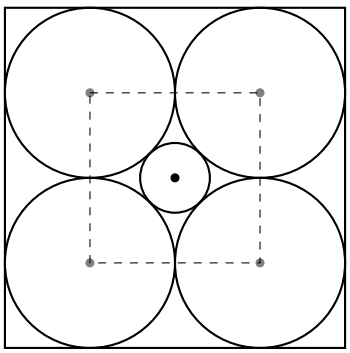
$$\begin{aligned} \mathbb{P}[\boldsymbol{G} \text{ does not have the } (k,\delta)\text{-RIP}] &\leq m^k \mathbb{P}[\|\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)} - \boldsymbol{I}_k\| > \delta] \\ &\leq 2\exp(k\log m - 2k\log m) \\ &\leq \frac{2}{m^k}, \end{aligned} \tag{1.5.11}$$

completing the proof. □

1.6 Exercises

Exercise 1.6.1. *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 $S_{\boldsymbol{s}}$. 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 1.6.2. *Show that there is a constant $c > 0$ such that, for all $\epsilon \in (0,1)$, for d sufficiently large, there are at least $N = \exp(c\epsilon^2 d)$ unit vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_N$ in $\mathbb{R}^d$ such that $|\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle| \le \epsilon$ for all $1 \le i < j \le N$ (i.e., such that the $\boldsymbol{v}_i$ are almost orthogonal). You may look up and use Hoeffding's inequality.*

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

Exercise 1.6.3. *Consider the shape $\Delta^d \subset \mathbb{R}^d$ that is the convex hull of the points $\mathbf{0}, \boldsymbol{e}_1, \boldsymbol{e}_1 + \boldsymbol{e}_2, \dots, \boldsymbol{e}_1 + \boldsymbol{e}_2 + \cdots + \boldsymbol{e}_d$ (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 $1 = \sqrt{1}, \sqrt{2}, \dots, \sqrt{d}$. 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 1.6.4. *In this exercise, you will prove a lower bound on the dimension required for embedding a particular point cloud that almost matches the Johnson-Lindenstrauss lemma. Along the way, you will see some linear algebra that you might not have been introduced to before.*

1. *Let $\lambda_1, \dots, \lambda_n \ge 0$. Show that*

$$\|\lambda\|_0 := \#\{i : \lambda_i \neq 0\} \ge \frac{(\sum_{i=1}^n \lambda_i)^2}{\sum_{i=1}^n \lambda_i^2} = \frac{\|\lambda\|_1^2}{\|\lambda\|_2^2}. \tag{1.6.1}$$

   Reinterpret this as a relationship between the rank, trace, and Frobenius norm of a positive semidefinite matrix.

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

3. *For $k \ge 1$ and $\boldsymbol{X} \in \mathbb{R}^{n \times n}_{\mathsf{sym}}$ with $\boldsymbol{X} \succeq \mathbf{0}$, write $\boldsymbol{X}^{\odot k}$ for the matrix that has entries $(\boldsymbol{X}^{\odot k})_{ij} = X_{ij}^k$, i.e., for the entrywise kth power of $\boldsymbol{X}$ (note that $\boldsymbol{X}^{\odot k} \neq \boldsymbol{X}^k$). Show that $\boldsymbol{X}^{\odot k} \succeq 0$, and that $\mathsf{rank}(\boldsymbol{X}^{\odot k}) \le \binom{\mathsf{rank}(\boldsymbol{X})+k}{k}$.*

   *(**Hint:** View $\boldsymbol{X}$ as a Gram matrix, $X_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$, and write $\boldsymbol{X}^{\odot k}$ in the same way. If you do this at all, the first part will follow (be sure to explain why). If you do it carefully, the second part will follow as well.)*

4. *Show that there are constants $c, \epsilon_0 > 0$ such that the following holds. For all $0 < \epsilon < \epsilon_0$ (i.e., ϵ sufficiently small), there exists $n_0 = n_0(\epsilon)$ such that, if $n \ge n_0(\epsilon)$ (i.e., n sufficiently large depending on ϵ), then there exists no pairwise ϵ-faithful embedding (that is, one preserving pairwise distances up to a factor of $1 \pm \epsilon$, as in the Johnson-Lindenstrauss lemma) of $\mathbf{0}, \boldsymbol{e}_1, \dots, \boldsymbol{e}_n \in \mathbb{R}^n$ into fewer than $\frac{c}{\log(1/\epsilon)} \cdot \frac{\log n}{\epsilon^2}$ dimensions. Thus, the dimension of the embedding provided by the Johnson-Lindenstrauss lemma for these points is tight up to a factor of $\log(1/\epsilon)$.*

*(**Hint:** Form the correlation matrix of the embeddings of the e_i. Raise it to a large enough entrywise power that Part 2 applies. Compare with Part 3.)*

Exercise 1.6.5. *The Gaussian measure is the most important one in probability theory, if not all of mathematics. Here you will derive some of its main algebraic properties.*

1. *Let $g \sim \mathcal{N}(0, \Sigma)$ for some $\Sigma \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ with $\Sigma \succeq 0$ (i.e., Σ is positive semidefinite). Prove that, for any smooth function $f : \mathbb{R}^d \to \mathbb{R}$ with $|f(x)| \leq C\|x\|^K$ for some $C, K > 0$ and all $x \in \mathbb{R}^d$, we have*

$$\mathbb{E}[g_i f(g)] = \sum_{j=1}^{d} \Sigma_{ij} \mathbb{E}[\partial_j f(g)] = (\Sigma\, \mathbb{E}[\nabla f(g)])_i \tag{1.6.2}$$

   where $\partial_i f$ is the partial derivative with respect to the ith argument and ∇ is the gradient (the second equality is just by the definition of gradient).

   *(**Hint:** Integrate by parts. You might also find it useful to first treat the case $\Sigma = I_d$, and then to observe that g and $\Sigma^{1/2}h$ for $h \sim \mathcal{N}(0, I_d)$ have the same law.)*

2. *Let $g \sim \mathcal{N}(0, 1)$ (a Gaussian scalar, not a vector). Prove that, for $k \geq 1$, $\mathbb{E}g^{2k-1} = 0$ and $\mathbb{E}g^{2k} = \prod_{i=1}^{k}(2i-1) =: (2k-1)!!$.*

3. *Let $g \sim \mathcal{N}(0, \Sigma)$ as in Part 1, and let $1 \leq i_1 < \cdots < i_k \leq d$. Let $\mathcal{M}$ be the set of all* matchings *of the set $I = \{i_1, \ldots, i_k\}$: a matching is a set of disjoint pairs $\{i_a, i_b\}$ whose union is I. For example, the three matchings of $\{1, 2, 3, 4\}$ are $\{\{1, 2\}, \{3, 4\}\}$, $\{\{1, 3\}, \{2, 4\}\}$, and $\{\{1, 4\}, \{2, 3\}\}$. Prove that*

$$\mathbb{E}\left[\prod_{a=1}^{k} g_{i_a}\right] = \sum_{M\in\mathcal{M}} \prod_{\{a,b\}\in M} \Sigma_{ab}. \tag{1.6.3}$$

   (For example, one case of the claim is that $\mathbb{E}g_1g_2g_3g_4 = \Sigma_{12}\Sigma_{34} + \Sigma_{13}\Sigma_{24} + \Sigma_{14}\Sigma_{23}$.) Generalize this to allow for repetitions among the $i_1, \ldots, i_k$. Try to be precise. Explain why Part 2 is a special case of this latter generalization.

   *(**Hint:** Induction.)*

Exercise 1.6.6. *Suppose that μ is a probability measure on $\mathbb{R}$ with a smooth density $\rho(x)$ that has $\rho(x) > 0$ for all $x \in \mathbb{R}$. Suppose that, for any smooth and compactly supported $f : \mathbb{R} \to \mathbb{R}$, $\mathbb{E}_{g\sim\mu}[gf(g)] = \mathbb{E}_{g\sim\mu}[f'(g)]$. Show that $\mu = \mathcal{N}(0, 1)$ (i.e., that the converse of the $d = 1$ case of Part 1 of the previous exercise holds).*

*(**Hint:** Write the expectations as integrals involving ρ. Integrate by parts.)*

2 | 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 dimension tend to infinity, without concrete bounds on probabilities of fluctuations like in, e.g., our previous centerpiece Theorem 1.4.18 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.

2.1 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

$$\boldsymbol{M} := \frac{1}{m}\boldsymbol{G}\boldsymbol{G}^\top = \frac{1}{m}\sum_{i=1}^m \boldsymbol{g}_i \boldsymbol{g}_i^\top. \tag{2.1.1}$$

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 Theorem 1.4.18 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, Theorem 1.4.18 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 $\approx \mathcal{N}(0, 1/m)$ 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 2.1 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*,

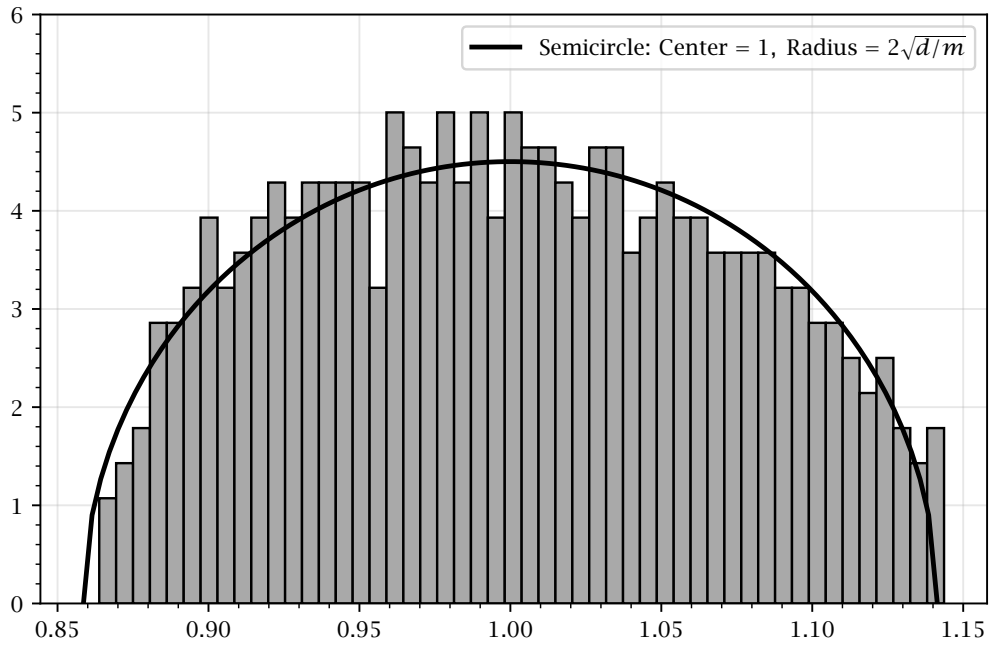


Figure 2.1: A histogram of the eigenvalues of $\boldsymbol{M}$ with $d = 500$ and $m = 100\,000$ along with the semicircle prediction proposed in the main text.

suggesting that the eigenvalues actually concentrate significantly more strongly than a hypothetical Gaussian limiting distribution would predict.

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, Theorem 1.4.18 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 $\boldsymbol{M} \succeq \boldsymbol{0}$). 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* $\mu_{\mathrm{MP}(c)}$. 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 2.2 illustrates, when $m = d$, then the density of eigenvalues converges as $d \to \infty$ to

$$\frac{d\mu_{\mathrm{MP}(1)}}{dx} = \frac{1}{2\pi}\sqrt{\frac{4-x}{x}}, \tag{2.1.2}$$

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 2.1.1 (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.*

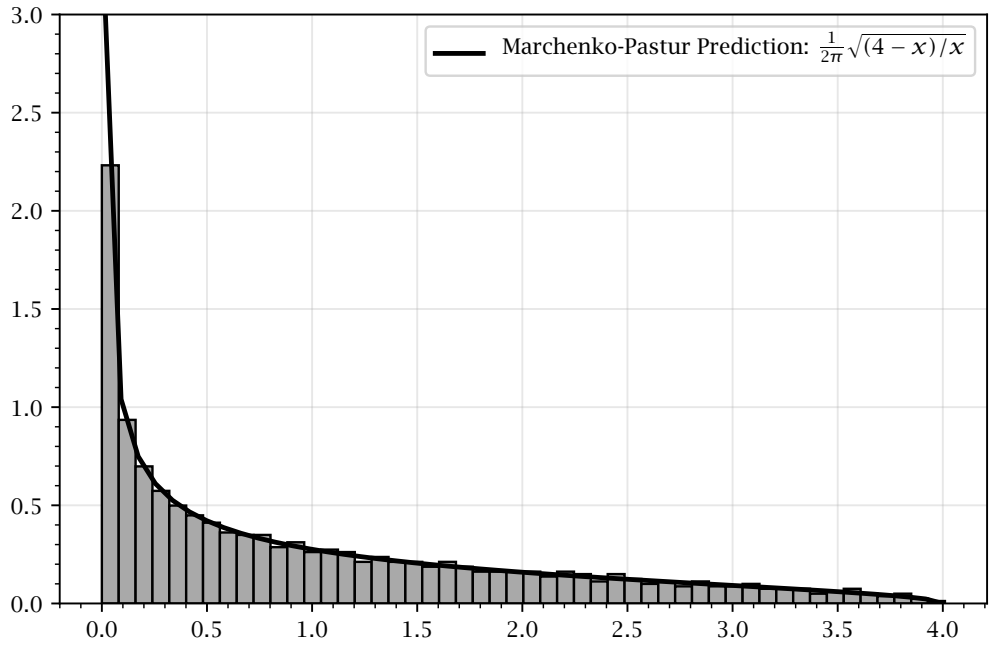


Figure 2.2: A histogram of the eigenvalues of $\boldsymbol{M}$ with $d = m = 1000$ along with the Marchenko-Pastur prediction proposed in the main text.

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 $\boldsymbol{g}_1, \dots, \boldsymbol{g}_d \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. 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 none other than $\boldsymbol{M}$ above.

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 $\boldsymbol{\Sigma}$ 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.)

2.2 Convergence of Random Measures

Let us first make more precise what it means for random "histograms" or, more precisely, 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.

2.2.1 Deterministic Weak Convergence

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

$$\mathsf{ed}(\boldsymbol{\lambda}) := \frac{1}{d} \sum_{i=1}^{d} \delta_{\lambda_i}. \tag{2.2.1}$$

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:

$$\mathsf{ed}(\boldsymbol{\lambda})(A) = \frac{\#\{i \in [d] : \lambda_i \in A\}}{d}. \tag{2.2.2}$$

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 2.2.1. *Let μ be a probability measure on $\mathbb{R}$. Its* cumulative distribution function (cdf) *is $\mathsf{cdf}(\mu) : \mathbb{R} \to \mathbb{R}$ with $\mathsf{cdf}(\mu)(t) := \mathbb{P}_{X\sim\mu}[X \le t] = \mu((-\infty, t])$.*

Definition 2.2.2 (Weak convergence)**.** *A sequence of probability measures $\mu^{(d)}$ on $\mathbb{R}$* converge weakly *to μ, written $\mu^{(d)} \to \mu$, if, whenever $\mathsf{cdf}(\mu)$ is continuous at $a, b \in \mathbb{R}$, then $\mu^{(d)}([a, b]) \to \mu([a, b])$ as $d \to \infty$.*

It is instructive to consider the special case that will usually arise for us: suppose $\mu^{(d)} = \mathsf{ed}(\boldsymbol{\lambda}^{(d)})$ and μ has a continuous density $\rho(t)$ on $\mathbb{R}$. Then, this claim is that, for all a, b (since $\mathsf{cdf}(\mu)$, the integral of ρ, is continuous everywhere, the caveat in the definition does not apply), we have

$$\frac{\#\{i \in [d] : \lambda_i^{(d)} \in [a, b]\}}{d} \to \int_a^b \rho(t)\,dt. \tag{2.2.3}$$

The condition that $\mathrm{cdf}(\mu)$ be continuous at a and b 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.

Let us check that weak convergence 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 $\mathsf{ed}(\boldsymbol{\lambda}^{(d)}) \to \mu$ which has a continuous density ρ, then for any choice of δ we will have, as $d \to \infty$,

$$\text{height of bin with left edge } a \to \int_a^{a+\delta} \rho(t)dt. \tag{2.2.4}$$

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

$$\lim_{\delta \to 0} \frac{1}{\delta} \lim_{d \to \infty} (\text{height of bin with left edge } a) = \lim_{\delta \to 0} \frac{1}{\delta} \int_a^{a+\delta} \rho(t)dt = \rho(a) \tag{2.2.5}$$

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.

2.2.2 Random Weak Convergence

The above discussion concerns deterministic $\lambda_i^{(d)}$. 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. Essentially the same statement as Definition 2.2.2 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 $C_d := \frac{1}{d}\#\{i : \lambda_i^{(d)} \in [a,b]\}$ are now *random*, and we are asking for them to converge to the deterministic number $C := \mu([a,b])$. 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. $C_d \to C$ *in expectation* if $\mathbb{E}C_d \to C$.
2. $C_d \to C$ *in probability* if, for all $\epsilon > 0$, $\mathbb{P}[|C_d - C| > \epsilon] \to 0$.
3. $C_d \to C$ *in* L^2 if $\mathbb{E}(C_d - C)^2 \to 0$, or equivalently (prove it if you do not see why it is true) if $C_d \to C$ in expectation and $\mathsf{Var}[C_d] \to 0$.

We therefore are justified in introducing the following definition.

Definition 2.2.3 (Random weak convergence)**.** *Let $\mu^{(d)}$ be random probability measures on $\mathbb{R}$: random variables taking their values in the space of probability measures on $\mathbb{R}$ (for example, $\mathsf{ed}(\boldsymbol{\lambda}^{(d)})$ for $\boldsymbol{\lambda}^{(d)}$ an arbitrary random vector). We say that $\mu^{(d)}$ converge* weakly in expectation *(respectively,* weakly in probability *and* weakly in L^2*) to μ, denoted $\mu^{(d)} \xrightarrow{\mathbb{E}} \mu$ (respectively, $\mu^{(d)} \xrightarrow{\mathbb{P}} \mu$ and $\mu^{(d)} \xrightarrow{L^2} \mu$), if, for all a, b where $\mathsf{cdf}(\mu)$ is continuous, $\mu^{(d)}([a,b])$ converges in expectation (respectively, in probability and in L^2) to $\mu([a,b])$.*

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.

2.3 Limit Theorems from Moments

Our goal will be to show random weak convergence for the empirical distributions of eigenvalues of random matrices, to which we give the following name.

Definition 2.3.1 (Empirical spectral distribution)**.** *The* empirical spectral distribution (e.s.d.) *of $\boldsymbol{M} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ is $\mathsf{esd}(\boldsymbol{M}) := \mathsf{ed}(\boldsymbol{\lambda}(\boldsymbol{M}))$.*

How are we going to do this? You might hope that we can write down the density of $\boldsymbol{\lambda}(\boldsymbol{M})$ and go from there. Unfortunately, this is only possible in some very special situations, and is not useful for proving *universality* of our results with respect to the details of the random matrix distributions involved.

Instead, let us draw inspiration from some results from scalar probability theory that are useful for proving results like the CLT. First, weak convergence is related to the convergence of expectations of test functions against the measures involved.

Lemma 2.3.2 (Portmanteau)**.** *$\mu^{(d)} \to \mu$ if and only if, for all $f : \mathbb{R} \to \mathbb{R}$ continuous and compactly supported, $\int f d\mu^{(d)} \to \int f d\mu$.*

Often this is given as the definition of weak convergence. Note that our original definition may be viewed as this kind of statement, but for discontinuous f that are the indicator functions of intervals $[a,b]$. Very roughly, the idea of the proof is to approximate such f by a linear combination of such step functions, and vice-versa.

This is still not very useful, because—even if we could formulate a version for random measures, which is non-trivial—while it reduces our task to saying something about the random empirical averages $\frac{1}{d}\sum_{i=1}^{d} f(\lambda_i(\boldsymbol{M}))$, we still do not have a convenient way to get a handle on these for general f. Fortunately, there is another family of results that shows the same for the very special class of polynomial functions f.

Theorem 2.3.3 (Carleman)**.** *Suppose that $\mu^{(d)}, \mu$ are probability measures that are all σ^2-subgaussian for some $\sigma^2 > 0$ (the same one for all of the measures). Then, $\mu^{(d)} \to \mu$ if and only if, for all $k \in \mathbb{N}$, $\int t^k d\mu^{(d)}(t) \to \int t^k d\mu(t)$.*

The proof idea here is similar, and is useful to note because it explains why a condition like subgaussianity is important. The idea, like the above, is to approximate any continuous and

compactly supported f by a polynomial. By the Weierstrass approximation theorem, this can be done to arbitrary pointwise accuracy, but *only on a compact interval*. E.g., f can be close to a polynomial on its support, but away from that support, the polynomial can (and nearly always will) fluctuate wildly while f will be zero. So, we must have some control of how quickly the $\mu^{(d)}$ decay to control this source of error, which is what the subgaussianity assumption achieves.

It is incredibly fortuitous to us that it is possible to establish limit theorems by working with moments. That is because moments are a *bridge between the entries and the eigenvalues of a matrix*. Consider: for $\boldsymbol{M} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$,

$$\int t^k d\,\mathsf{esd}(\boldsymbol{M})(t) = \frac{1}{d}\sum_{i=1}^{d} \lambda_i(\boldsymbol{M})^k = \frac{1}{d}\,\mathsf{Tr}(\boldsymbol{M}^k). \tag{2.3.1}$$

The final expression may be expanded in terms of the entries of $\boldsymbol{M}$ as a big polynomial. We will see that this lets us perform merely combinatorial computations on $\boldsymbol{M}$ in order to learn the distribution of its eigenvalues.

We therefore make the following definition.

Definition 2.3.4 (Convergence in moments)**.** *We say that random probability measures $\mu^{(d)}$* converge in (expected) moments *to μ, denoted $\mu^{(d)} \xrightarrow{\mathbb{E}\,mom.} \mu$, if, for all $k \geq 0$, $\mathbb{E}\int t^k d\mu^{(d)}(t) \to \int t^k d\mu(t)$.*

A typical argument by the moment-based method[1] will have two steps, which we leave informal for now and will illustrate in one special situation below:

1. Show that $\mathsf{esd}(\boldsymbol{M}^{(d)}) \xrightarrow{\mathbb{E}\,\mathrm{mom.}} \mu$.
2. Prove some sort of concentration result showing that, with high probability, the eigenvalues of $\boldsymbol{M}^{(d)}$ are not very large.

2.4 Wigner Semicircle Limit Theorem

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

Definition 2.4.1 (Wigner matrix)**.** *Let ν be a probability measure on $\mathbb{R}$. We write $\mathsf{Wig}(d,\nu)$ for the probability measure on $\mathbb{R}^{d\times d}_{\mathsf{sym}}$ where we sample $\boldsymbol{W} \sim \mathsf{Wig}(d,\nu)$ by drawing $W_{ij} = W_{ji} \sim \mu$ i.i.d. and setting $W_{ii} := 0$.*

Definition 2.4.2 (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 2.4.3 (Wigner's limit theorem)**.** *Suppose that ν is a probability measure all of whose moments are finite and with $\int x d\nu(x) = 0$ and $\int x^2 d\nu(x) = 1$. Let $\boldsymbol{W}^{(d)} \sim \mathsf{Wig}(\nu,d)$ for each $d \geq 1$. Then, $\mathsf{esd}(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}) \xrightarrow{\mathsf{mode}} \mu_{\mathsf{SC}}$ for any $\mathsf{mode} \in \{\mathbb{E},\mathbb{P},L^2\}$.*

[1] Not to be confused with the "method of moments" of statistical inference.

Remark 2.4.4 (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* $\sqrt{\frac{m}{d}}(\frac{1}{m}\sum_{i=1}^m \boldsymbol{g}_i\boldsymbol{g}_i^\top - \boldsymbol{I}_d) = \frac{1}{\sqrt{dm}}\sum_{i=1}^m(\boldsymbol{g}_i\boldsymbol{g}_i^\top - \boldsymbol{I}_d)$*. Indeed you may check numerically that this is the case. It is also reasonable that this random vector behaves somewhat like a Wigner matrix, at least if we ignore the diagonal. You may check that the off-diagonal entries of* $\boldsymbol{g}_i\boldsymbol{g}_i^\top$ *are pairwise uncorrelated, so if, hypothetically, we took* $m \to \infty$ *for* d *fixed, then the above matrix would converge to a symmetric Wigner matrix with Gaussian entries by the 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 2.4.5 (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). In fact, the moment assumptions in Theorem 2.4.3 can be relaxed even more: 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.*

We will proceed according to the strategy outlined before. Actually, before getting into moment calculations with random matrices, it will be helpful by way of analogy to recall the proof using the moment method of the CLT.

2.4.1 Warmup: Central Limit Theorem by Moments

We review a proof of the following version of the CLT.

Theorem 2.4.6. *Let* ν *be a subgaussian probability measure on* $\mathbb{R}$ *having* $\mathbb{E}_{X\sim\nu}X = 0$ *and* $\mathbb{E}_{X\sim\nu}X^2 = 1$*. Let* $X_1, X_2, X_3, \ldots \sim \nu$ *be an i.i.d. sequence. Then,* $\mathsf{Law}(\frac{1}{\sqrt{d}}\sum_{i=1}^d X_i) \to \mathcal{N}(0,1)$ *(in the sense of weak convergence).*

Proof. By Carleman's theorem (our Theorem 2.3.3), it suffices to show the convergence of moments

$$\mathbb{E}\left(\frac{1}{\sqrt{d}}\sum_{i=1}^d X_i\right)^k \overset{?}{\to} \mathop{\mathbb{E}}_{N\sim\mathcal{N}(0,1)} N^k =: m_k^{\mathcal{N}(0,1)} \text{ for all } k \geq 0. \tag{2.4.1}$$

(You can check that the subgaussianity assumption on ν implies that the corresponding condition of Carleman's theorem is satisfied.)

First, let us calculate the right-hand sides to know what we expect in seeking out a "main term" on the left. An induction together with a simple calculation with integration by parts shows that, for all $k \geq 0$,

$$m_k^{\mathcal{N}(0,1)} = \#\{\text{matchings of } [k]\} = \left\{ \begin{array}{ll} 0 & \text{if } k \text{ odd} \\ (k-1)!! & \text{if } k \text{ even} \end{array} \right\}. \tag{2.4.2}$$

The explicit formula is useful in other situations but will not be relevant here; we care more about the former enumerative interpretation.

Now consider the empirical moments. We write $\mathsf{Part}([k])$ for the set of *partitions* of k: a partition is a set of disjoint sets whose union is $[k]$. For instance, $\{\{1,2\},\{3,4,5\}\}$ is a partition of $[5]$. We expand directly, finding:

$$\mathbb{E}\left(\frac{1}{\sqrt{d}}\sum_{i=1}^{d} X_i\right)^k = \frac{1}{d^{k/2}} \sum_{i_1,\ldots,i_k \in [k]} \mathbb{E} X_{i_1} \cdots X_{i_k}$$

and now we note that to each term is associated a partition $\pi \in \mathsf{Part}([k])$ according to which of the i_a are equal to one another, and the value of the term only depends on π. We therefore have

$$= \frac{1}{d^{k/2}} \sum_{\pi \in \mathsf{Part}([k])} d(d-1)\cdots(d-|\pi|+1)\mathbb{E}\prod_{S\in\pi} X_S^{|S|}$$

where we introduce $X_S \sim \nu$ i.i.d. for each part $S \in \pi$. Since these random variables are independent, we can further simplify

$$= \frac{1}{d^{k/2}} \sum_{\pi \in \mathsf{Part}([k])} d(d-1)\cdots(d-|\pi|+1)\prod_{S\in\pi} \mathop{\mathbb{E}}_{X\sim\nu} X^{|S|}$$

and since $\mathbb{E}_{X\sim\nu} X = 0$, we may also restrict

$$= \frac{1}{d^{k/2}} \sum_{\substack{\pi \in \mathsf{Part}([k]) \\ |S|\geq 2 \text{ for all } S\in\pi}} d(d-1)\cdots(d-|\pi|+1)\prod_{S\in\pi} \mathop{\mathbb{E}}_{X\sim\nu} X^{|S|}$$

Now, recall that k is constant and the number of terms in this sum only depends on k, while we are interested in the limit as $d \to \infty$. The product of expectations inside also does not depend on d. Thus, any term where $|\pi| < k/2$ will tend to zero as $d \to \infty$. That is, in this limit we have the convergence:

$$\begin{aligned} &\to \sum_{\substack{\pi \in \mathsf{Part}([k]) \\ |S|=2 \text{ for all } S\in\pi}} \prod_{S\in\pi} \mathop{\mathbb{E}}_{X\sim\nu} X^{|S|} \\ &= \sum_{\substack{\pi \in \mathsf{Part}([k]) \\ |S|=2 \text{ for all } S\in\pi}} 1, \end{aligned} \tag{2.4.3}$$

which is merely the number of matchings of $[k]$ by definition, and the proof is complete. □

The proof is an elegant demonstration of how the moment method works and why it is so useful. The Gaussian distribution, as with many natural distributions of probability theory, has a natural combinatorial interpretation of its moments. The moment method lets us reduce proving a limit theorem to showing that a certain more complicated combinatorial quantity—the expanded empirical moment above—behaves like the Gaussian moment to leading order.

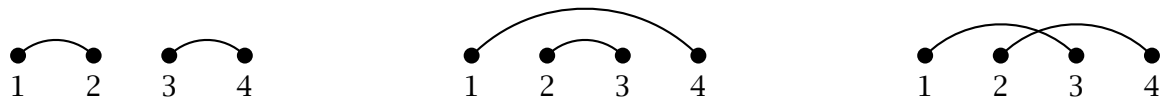


Figure 2.3: The three matchings, two non-crossing and one crossing, of four objects that arise in the computation of the fourth moments of $\mathcal{N}(0,1)$ and μ_{SC}.

k	0	2	4	6	8	10	12	14	16
$m_k^{\mathcal{N}(0,1)}$	1	1	3	15	105	735	6615	72 765	945 945
m_k^{SC}	1	1	2	5	14	42	132	429	1430

Table 2.1: A comparison of the first few moments of $\mathcal{N}(0,1)$ (the "matching numbers") and of μ_{SC} (the Catalan numbers).

2.4.2 Convergence of Expected Moments

We now execute the same strategy for Wigner's limit theorem.

Lemma 2.4.7. *In the setting of Theorem 2.4.3,* $\mathsf{esd}(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}) \xrightarrow{\mathbb{E}\ mom.} \mu_{\mathsf{SC}}$.

Proof. Again, we first compute the moments of the limiting distribution. This is another, slightly more tedious but still straightforward, exercise in integration by parts:

$$m_k^{\mathsf{SC}} := \mathop{\mathbb{E}}_{X\sim\mu_{\mathsf{SC}}} X^k = \#\{\text{non-crossing matchings of } [k]\} = \left\{ \begin{array}{ll} 0 & \text{if } k \text{ odd} \\ \mathsf{Cat}_\ell & \text{if } k = 2\ell \text{ even} \end{array} \right\}, \quad (2.4.4)$$

where $\mathsf{Cat}_\ell := \frac{1}{\ell+1}\binom{2\ell}{\ell}$ are the *Catalan numbers*, and a non-crossing matching is one that has no crossings when drawn as a matching of points arranged in a line. For example, the first two matchings of $[4]$ in Figure 2.3 are non-crossing, while the last has a crossing.

Generally, there are many fewer non-crossing matchings than general matchings, as we can see from the scaling behaviors that, roughly, $m_k^{\mathcal{N}(0,1)} \approx \sqrt{k}^k$ while $m_k^{\mathsf{SC}} \approx 2^k$ (both by Stirling's formula, omitting lower order terms). The first few of each sequence are given in Table 2.1.

With this in hand, we may proceed to the calculation, 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:

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

and by expanding the matrix multiplication in indices,

$$= \frac{1}{d^{k/2+1}} \sum_{i_1,\ldots,i_k\in[d]} \mathbb{E}W_{i_1i_2}W_{i_2i_3}\cdots W_{i_{k-1}i_k}W_{i_ki_1}$$

We associate to each such term a "closed walk graph" (my non-standard terminology): a connected graph with a distinguished root vertex whose edges are directed and numbered and that, when traversed in sequence, end at the same vertex where they started. 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 2.4 for an example. Writing $V(G)$ and $E(G)$ for the vertex and edge set of a graph, we then have

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

Here we may restrict to G simple since consecutive i_a are distinct in any non-zero term as $W_{ii} = 0$. Above, $UE(G)$ denotes 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 where we introduce independent $W_e \sim \nu$ for each undirected edge $e \in UE(G)$. 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 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 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 visit each vertex of the tree 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,

$$\mathbb{E} \frac{1}{d} \operatorname{Tr}\left(\frac{1}{\sqrt{d}} \boldsymbol{W}\right)^k \to \sum_{\substack{G \text{ a rooted tree traversal} \\ |E(G)|=k}} 1. \tag{2.4.5}$$

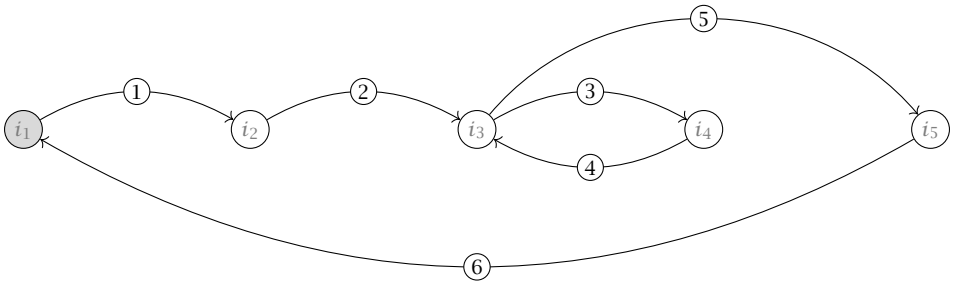


Figure 2.4: The closed walk graph associated to a term of the form $W_{i_1 i_2} W_{i_2 i_3} W_{i_3 i_4} W_{i_4 i_3} W_{i_3 i_5} W_{i_5 i_1}$ in $\mathsf{Tr}(\boldsymbol{W}^6)$, where $i_1, \dots, i_5$ are all distinct. We include these index labels for reference, but we think of a closed walk graph as *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.

It remains to show that the number of 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, the bijection between Dyck paths and tree traversals is just to "spread out" the tree traversal, viewing it over an x-axis of "time," or to "glue" levels of a Dyck path.

See Figure 2.5 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 visited fully determines the order in which the directed edges are traversed. □

2.4.3 Upgrading to Weak Convergence in Probability

We will now use the convergence of moments to show the version of Theorem 2.4.3 involving convergence in probability. In particular, we will show the following:

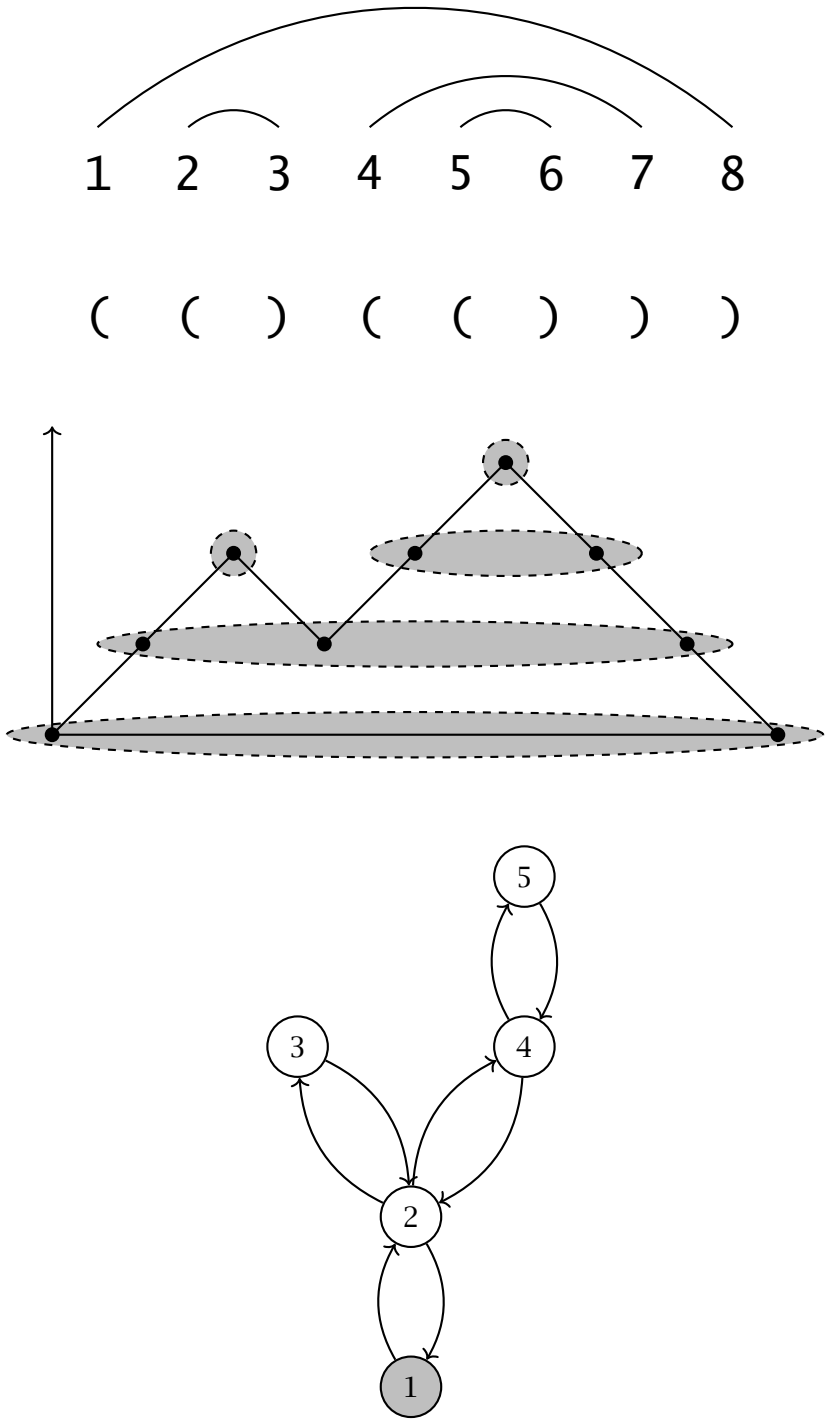


Figure 2.5: An example of the bijection between four kinds of objects that the Catalan numbers (Cat_4 in this case) count, as used in the proof of Lemma 2.4.7: non-crossing matchings, parenthesizations, Dyck paths, and rooted tree traversals. We also show the mapping from Dyck paths to tree traversals by showing which vertices along the Dyck path are identified to form the tree.

Theorem 2.4.8 (Wigner's limit theorem in probability)**.** *In the setting of Theorem 2.4.3, for any continuous bounded $f : \mathbb{R} \to \mathbb{R}$, $\frac{1}{d}\sum_{i=1}^{d} f(\lambda_i(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)})) \to \int f d\mu_{\mathsf{SC}}$ in probability.*

It is straightforward to deduce the version of this for convergence of counts of eigenvalues in an interval, as follows.

Corollary 2.4.9. *In the setting of Theorem 2.4.3, 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

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

for some absolute constant C, and likewise with f replaced by g. In particular, we find that, using Theorem 2.4.8 on f and g and combining with the above observation, for any ϵ,

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

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

Now let us proceed to the proof of our version of Wigner's theorem, where we will implement the full strategy of the moment method as outlined earlier. We note that you can use essentially the same strategy with small modifications to prove weak convergence either in expectation or in L^2 as well.

To wit, even below we will use the following "L^2 version" of Lemma 2.4.7. We do not give a careful proof, but sketch the main idea below.

Lemma 2.4.10. *Suppose that ν is a probability measure all of whose moments are finite and with $\int x d\nu(x) = 0$ and $\int x^2 d\nu(x) = 1$. Let $\boldsymbol{W}^{(d)} \sim \mathsf{Wig}(\nu, d)$ for each $d \geq 1$. For any k, there is a constant $C = C(\nu, k) > 0$ such that $\mathsf{Var}[\frac{1}{d}\sum_{i=1}^d \lambda_i (\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)})^k] \leq 1/d^2 + C/d^3$.*

Proof Sketch. Writing $\boldsymbol{W} := \boldsymbol{W}^{(d)}$, we may decompose by a similar expansion to Lemma 2.4.7,

$$\begin{aligned}\mathsf{Var}\left[\frac{1}{d}\sum_{i=1}^d \lambda_i \left(\frac{1}{\sqrt{d}}\boldsymbol{W}\right)^k\right] &= \mathsf{Var}\left[\frac{1}{d^{k/2+1}}\sum_{i_1,\dots,i_k\in[d]} W_{i_1 i_2}\cdots W_{i_k i_1}\right] \\ &= \frac{1}{d^{k+2}}\sum_{\substack{i_1,\dots,i_k\in[d]\\ j_1,\dots,j_k\in[d]}} \mathsf{Cov}[W_{i_1 i_2}\cdots W_{i_k i_1}, W_{j_1 j_2}\cdots W_{j_k j_1}].\end{aligned}$$

The terms here may be analyzed by a similar but slightly more complicated graphical approach as Lemma 2.4.7. The point becomes that the main contribution is from those terms where $i_1 = j_1, \dots, i_k = j_k$ and $i_1, \dots, i_k$ are all distinct. There are $d(d-1)\cdots(d-k+1)$ such terms, each of which makes a contribution of 1. Thus the resulting main term is $1/d^2$, as claimed. □

Remark 2.4.11. *Contrast the above with the case of independent random variables λ_i, say $\lambda_i \sim \mu_{\mathsf{SC}}$ drawn i.i.d. These will still have empirical distribution close to μ_{SC}, but will have $\mathsf{Var}[\frac{1}{d}\sum_{i=1}^d \lambda_i] = \frac{1}{d^2}\sum_{i=1}^d \mathsf{Var}[\lambda_i] = \Omega(1/d)$. Thus in a sense the λ_i have a much more "rigid" distribution than independent random variables. This is related to the "repulsion" between eigenvalues and the distribution of eigenvalue spacing, a set of phenomena we will not study here but that you are welcome to investigate on your own.*

We can now proceed to the main proof.

Proof of Theorem 2.4.8. Let us write $\lambda_i = \lambda_i^{(d)} := \frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}$ as before, to lighten the notation further. Our task is to show that, for any $\epsilon > 0$,

$$\lim_{d\to\infty} \mathbb{P}\left[\left|\frac{1}{d}\sum_{i=1}^d f(\lambda_i) - \int f d\mu_{\mathsf{SC}}\right| > \epsilon\right] = 0.$$

We expect the λ_i to fall close to the interval $[-2, 2]$. Let us therefore fix $B > 2$; certainly the λ_i should fall in $[-B, B]$.

Let $\delta > 0$ to be chosen later. By the Weierstrass approximation theorem, there is a polynomial p such that, for all $x \in [-B, B]$, we have $|f(x) - p(x)| \leq \delta$. We use this guarantee

together with our knowledge of convergence in moments to control the probability above. In particular, we decompose by a big triangle inequality

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

Here, the third and sixth terms are at most δ by the approximation guarantee between f and p, and, for sufficiently large d, the second term is also at most δ by Lemma 2.4.7 (expanding $p(t)$ in monomials and considering the limit of each one). Therefore,

$$\leq 3\delta + \underbrace{\left|\frac{1}{d}\sum_{i=1}^{d} p(\lambda_i) - \mathbb{E}\frac{1}{d}\sum_{i=1}^{d} p(\lambda_i)\right|}_{①} + \underbrace{\left|\frac{1}{d}\sum_{i:|\lambda_i|>B} f(\lambda_i)\right|}_{②} + \underbrace{\left|\frac{1}{d}\sum_{i:|\lambda_i|>B} p(\lambda_i)\right|}_{③}.$$

Suppose that we choose δ such that $6\delta < \epsilon$. Then, we may bound

$$\mathbb{P}\left[\left|\frac{1}{d}\sum_{i=1}^{d} f(\lambda_i) - \int f d\mu_{\mathsf{SC}}\right| > \epsilon\right] \leq \mathbb{P}\left[① > \delta\right] + \mathbb{P}\left[② > \delta\right] + \mathbb{P}\left[③ > \delta\right].$$

Since p is a polynomial, we may expand it as, for some D and $a_0, \ldots, a_D \in \mathbb{R}$,

$$p(t) = \sum_{k=0}^{D} a_k t^k.$$

Let us write $A := \max_{k=0}^{D} |a_k|$. To control ①, we use Chebyshev's inequality, finding

$$\begin{aligned}
\mathbb{P}\left[① > \delta\right] &\leq \frac{1}{\delta^2}\,\mathsf{Var}\left[\frac{1}{d}\sum_{i=1}^{d} p(\lambda_i)\right] \\
&= \frac{1}{\delta^2}\,\mathsf{Var}\left[\sum_{k=0}^{D}\left(\frac{a_k}{d}\sum_{i=1}^{d}\lambda_i^k\right)\right] \\
&\leq \frac{D+1}{\delta^2}\sum_{k=0}^{D}\mathsf{Var}\left[\frac{a_k}{d}\sum_{i=1}^{d}\lambda_i^k\right] \\
&\leq \frac{A^2(D+1)}{\delta^2}\sum_{k=0}^{D}\mathsf{Var}\left[\frac{1}{d}\sum_{i=1}^{d}\lambda_i^k\right],
\end{aligned}$$

which by Lemma 2.4.10 tends to zero as $d \to \infty$. (We have been moderately careful about keeping track of the constants in front of the variance, but they all do not matter: the point is just that D and the a_k are fixed while $d \to \infty$, and in this limit each variance above decays to zero.) We conclude that, for any fixed $\delta > 0$,

$$\lim_{d\to\infty} \mathbb{P}\left[① > \delta\right] = 0.$$

For ②, we recall that we have assumed f is bounded, so suppose that $|f(x)| \leq F$ for all x. We then bound by Markov's inequality

$$\begin{aligned}
\mathbb{P}\left[② > \delta\right] &= \mathbb{P}\left[\left|\frac{1}{d}\sum_{i:|\lambda_i|>B} f(\lambda_i)\right| > \delta\right] \\
&\leq \mathbb{P}\left[F \cdot \frac{1}{d}\#\{i : |\lambda_i| > B\} > \delta\right] \\
&\leq \frac{F}{\delta}\mathbb{E}\left[\frac{1}{d}\#\{i : |\lambda_i| > B\}\right]
\end{aligned}$$

and now we use a handy trick to relate this to the moments that Lemma 2.4.7 computes for us: for any $k \geq 1$,

$$\leq \frac{F}{\delta B^{2k}}\mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d}\lambda_i^{2k}\right]$$

which in the limit $d \to \infty$ goes to a Catalan number,

$$\begin{aligned}
&\to \frac{F}{\delta B^{2k}}\mathsf{Cat}_k \\
&\leq \frac{F}{\delta}\left(\frac{2}{B}\right)^{2k}.
\end{aligned}$$

Thus we learn that, for any $k \geq 1$,

$$\lim_{d\to\infty} \mathbb{P}\left[② > \delta\right] \leq \frac{F}{\delta}\left(\frac{2}{B}\right)^{2k}.$$

Taking $k \to \infty$ then gives

$$\lim_{d\to\infty} \mathbb{P}\left[\textcircled{2} > \delta\right] = 0.$$

Finally, for $\textcircled{3}$, we use another permutation of the ideas above. Using the expansion of p into monomials and the triangle inequality

$$\begin{aligned}\mathbb{P}\left[\textcircled{3} > \delta\right] &= \mathbb{P}\left[\left|\frac{1}{d}\sum_{i:|\lambda_i|>B} p(\lambda_i)\right| > \delta\right] \\ &\leq \sum_{k=0}^{D} \mathbb{P}\left[\frac{1}{d}\sum_{i:|\lambda_i|>B} |a_k||\lambda_i|^k > \frac{\delta}{D+1}\right]\end{aligned}$$

and now using the same Markov inequality argument from before

$$\leq \frac{A(D+1)}{\delta}\sum_{k=0}^{D}\mathbb{E}\left[\frac{1}{d}\sum_{i:|\lambda_i|>B}|\lambda_i|^k\right]$$

Here, let us choose some $\ell > D/2$, with which we may bound

$$\begin{aligned}&\leq \frac{A(D+1)}{\delta}\sum_{k=0}^{D}\mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d}\frac{\lambda_i^{2\ell}}{B^{2\ell-k}}\right] \\ &\leq \frac{A(D+1)^2B^D}{\delta B^{2\ell}}\mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d}\lambda_i^{2\ell}\right] \\ &\to \frac{A(D+1)^2B^D}{\delta B^{2\ell}}\mathsf{Cat}_\ell \\ &\leq \frac{A(D+1)^2B^D}{\delta}\left(\frac{2}{B}\right)^{2\ell}.\end{aligned}$$

Finishing the argument as before (again, note that we have been keeping track of the constants in front of the term exponentially decaying in ℓ, but these do not matter for the argument), we find

$$\lim_{d\to\infty} \mathbb{P}\left[\textcircled{3} > \delta\right] = 0.$$

Combining the analysis of $\textcircled{1}$, $\textcircled{2}$, and $\textcircled{3}$, we find

$$\begin{aligned}&\lim_{d\to\infty}\mathbb{P}\left[\left|\frac{1}{d}\sum_{i=1}^{d} f(\lambda_i) - \int f d\mu_{\mathsf{SC}}\right| > \epsilon\right] \\ &\quad\leq \lim_{d\to\infty}\mathbb{P}\left[\textcircled{1} > \delta\right] + \lim_{d\to\infty}\mathbb{P}\left[\textcircled{2} > \delta\right] + \lim_{d\to\infty}\mathbb{P}\left[\textcircled{3} > \delta\right] \\ &\quad= 0,\end{aligned}$$

and the proof is complete. □

2.5 Extreme Eigenvalues from Moments

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

We may write the finding of Lemma 2.4.7 in a way more relevant to our setting as follows. Given the entrywise distribution ν, for each k, there is a $\Delta_{2k} = \Delta_{2k}(\nu)$ such that

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

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

For extreme eigenvalues, we want to control the probability

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

and in particular 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-type inequality:

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

We see a few important features.

First, in order for the first term to decay once $t = 2 + \epsilon$, we will need to take $k = k(d) \gg \log(d)$, say $k(d) = (\log d)^{1+\delta}$ for some $\delta > 0$. (If, e.g., we try to run this argument with k constant, we will only get that the norm is $O(d^{1/2k})$, which is not useless but is far from tight.)

On the other hand, when we try to do this, we find that we need finer control over Δ_{2k} than we have established. Namely, in order for the second term to tend to zero with our choices above, we need that $\lim_{k\to\infty} \Delta_{2k}^{1/2k} \leq 2$. Our analysis thus far does not ensure this, and it is a non-trivial extension of the combinatorics we have encountered in Lemma 2.4.7 to do so, amounting to carefully examining the error terms that appear. The following is one version of this.

Lemma 2.5.1. *Suppose that ν is a probability measure satisfying the assumptions of Theorem 2.4.3, that in addition has $\mathbb{E}_{w\sim\nu}|w|^k \leq k^{Ck}$ for some $C > 0$ and for all $k \geq 1$ (for instance, ν can be any bounded or subgaussian distribution). Then, in the context above, there are $A, B > 0$ depending only on C such that we may take $\Delta_{2k} = A \cdot k^B \cdot 2^{2k}$.*

Using this together with the argument above gives the following limit theorem for the spectral norm.

Theorem 2.5.2. *In the context of Theorem 2.4.3, under the additional assumption on ν from Lemma 2.5.1, $\|\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}\| \to 2$ in probability.*

Remark 2.5.3. *In the same spirit as Remark 2.4.5, it is possible to loosen the moment conditions to only require that the fourth moment is finite. In an interesting nuance, this optimal and very general, but is slightly stronger than loosest possible condition for Theorem 2.4.3 (the weak convergence to the semicircle law) to hold. That is, 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, but has divergent spectral norm, $\|\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}\| = \omega(1)$ with high probability.*

2.6 Sketch of 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.

Definition 2.6.1. *For $c \in (0,\infty)$, define the measure $\mu_{\mathsf{MP}(c)}$ by*

$$d\mu_{\mathsf{MP}(c)}(x) = \max\left\{0, 1-\frac{1}{c}\right\}\delta_0 + \mathbb{1}\{x \in [\lambda_-,\lambda_+]\}\frac{1}{2\pi c}\sqrt{\left(\frac{\lambda_+}{x}-1\right)\left(1-\frac{\lambda_-}{x}\right)}\,dx,$$

where

$$\lambda_\pm = \lambda_\pm(c) = (1\pm\sqrt{c})^2 = 1+c\pm 2\sqrt{c}.$$

Theorem 2.6.2. *Suppose that $d = d_n$ and $m = m_n$ are increasing sequences with $d_n/m_n \to c \in (0,\infty)$. Suppose $\boldsymbol{G} = \boldsymbol{G}^{(d)} \sim \mathcal{N}(0,1)^{\otimes d\times m}$ and let $\boldsymbol{M} = \boldsymbol{M}^{(d)} := \frac{1}{m}\boldsymbol{G}\boldsymbol{G}^\top \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$. Then, $\mathsf{esd}(\boldsymbol{M}^{(d)}) \xrightarrow{\textit{mode}} \mu_{\mathsf{MP}(c)}$ for any mode $\in \{\mathbb{E},\mathbb{P},L^2\}$.*

To interpret the result, consider two cases. First, if $c \le 1$, then $d \le m$, so $\boldsymbol{M}$ is full rank (generically) and we expect none of its eigenvalues to 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 $2\sqrt{c} = 2\sqrt{d/m}$. 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 2.1.

Second, if $c > 1$, then $d > m$, so $\boldsymbol{M}$ is rank deficient (generically), having rank only m, and should have a fraction of $\frac{d-m}{d} = 1-\frac{m}{d} = 1-\frac{1}{c}$ 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 is again centered at $c+1$ (where now c is the "dominant" part) and has 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 $\mu_{\mathrm{MP}(c)}$ and $\mu_{\mathrm{MP}(1/c)}$ for all c (which is slightly obscured since we view $\boldsymbol{G}\boldsymbol{G}^\top$ as normalized by $m = d/c$, changing the scaling of the continuous part).

The proof by moments is a more complicated version of our argument for Theorem 2.4.3. We see the similarities and differences upon taking the first few steps:

$$\begin{aligned}\frac{1}{d}\operatorname{Tr}(\boldsymbol{M}^k) &= \frac{1}{dm^k}\operatorname{Tr}((\boldsymbol{G}\boldsymbol{G}^\top)^k)\\ &= \frac{c^k}{d^{k+1}}\sum_{i_1,\dots,i_k\in[d]}(\boldsymbol{G}\boldsymbol{G}^\top)_{i_1i_2}\cdots(\boldsymbol{G}\boldsymbol{G}^\top)_{i_ki_1}\\ &= \frac{c^k}{d^{k+1}}\sum_{\substack{i_1,\dots,i_k\in[d]\\ j_1,\dots,j_k\in[m]}}G_{i_1j_1}G_{i_2j_1}G_{i_2j_2}G_{i_3j_2}\cdots G_{i_kj_k}G_{i_1j_k}.\end{aligned}$$

Thus we will end up counting a kind of bipartite version of the closed walk structures that appeared for Theorem 2.4.3. Moreover, the parameter c will appear as a weight given to these objects, since the number of distinct $j_\bullet$ indices that a given bipartite graph can be labelled by will be of the form $m^b = d^b/c^b$, while the number of distinct $i_\bullet$ indices will just be of the form d^a. Still, it is possible to carry out these calculations and prove Theorem 2.6.2 with essentially the same tools we have seen already.

2.7 Stieltjes Transform and Resolvent Arguments

Finally, we sketch another approach to the semicircle and Marchenko-Pastur limit theorems, that is somewhat "slicker" and easier to generalize in certain directions.

Definition 2.7.1. *The* Stieltjes transform *of a probability measure μ is*

$$G_\mu(z) := \mathop{\mathbb{E}}_{X\sim\mu}\left[\frac{1}{z-X}\right],$$

defined on $z \in \mathbb{C}\setminus\operatorname{supp}(\mu)$. (This definition differs by a sign from the usual definition but is used in some references like [PB20]; the function G_μ is sometimes called the Green's function *also.)*

Note that this is a kind of moment generating function: for $|z|$ large enough and μ compactly supported, we may expand

$$G_\mu(z) = \frac{1}{z}\mathbb{E}\frac{1}{1-\frac{X}{z}} = \sum_{k\geq 0}\frac{1}{z^{k+1}}\mathbb{E}X^k, \tag{2.7.1}$$

so in particular the *ordinary moment generating function* (ordinary in the technical sense of lacking the $1/k!$ factors appearing in the expansion of $\mathbb{E}\exp(zX)$) is

$$\frac{1}{z}G_\mu\left(\frac{1}{z}\right) = \sum_{k\geq 0}z^k\mathbb{E}X^k.$$

It is then not surprising that it is possible to invert the Stieltjes transform to determine μ, as with the moment generating function $\phi_\mu(z) := \mathbb{E}_{X\sim\mu}\exp(zX)$, which for purely imaginary

$z = it$ gives the *characteristic function*, which in turn for μ with a density are just the Fourier transform of that density.

Indeed, we may give more intuition about the Stieltjes transform by considering its real and imaginary parts: if $z = s + it$, then

$$\begin{aligned}
\mathsf{Im}(G_\mu(z)) &= \mathop{\mathbb{E}}_{X\sim\mu} \mathsf{Im}\left(\frac{1}{z-X}\right) \\
&= \mathop{\mathbb{E}}_{X\sim\mu} \frac{\mathsf{Im}(\overline{z-X})}{|z-X|^2} \\
&= -\mathop{\mathbb{E}}_{X\sim\mu} \frac{t}{(X-s)^2+t^2} \\
\mathsf{Re}(G_\mu(z)) &= \mathop{\mathbb{E}}_{X\sim\mu} \frac{s-X}{(s-X)^2+t^2}
\end{aligned}$$

The imaginary part in particular has a nice interpretation: the *Cauchy distribution* with shape parameter t, denoted $\mathsf{Cauchy}(t)$, has density

$$\rho_t(x) = \frac{1}{\pi}\frac{t}{x^2+t^2} \text{ for } x \in \mathbb{R}.$$

Thus, we have

$$\mathsf{Im}(G_\mu(s+it)) = -\pi \mathop{\mathbb{E}}_{X\sim\mu} \rho_t(s-X) = -\pi \int \rho_t(s-x)d\mu(x).$$

Up to the constant in front, if μ has a density, then the latter is the density of $X + Z$ for $X \sim \mu$ and $Z \sim \mathsf{Cauchy}(t)$ at s (in other words, the convolution of the density of μ with the density of $\mathsf{Cauchy}(t)$). While a Cauchy distribution is very heavy-tailed (neither its first or second moment are finite!), we still expect this convolution to approach the identity as $t \to 0$, so the following should not be surprising.

Theorem 2.7.2 (Stieltjes inversion formula)**.** *If* $\mathsf{cdf}(\mu)$ *is continuous at* $a, b \in \mathbb{R}$ *with* $a < b$*, then*

$$\mu([a,b]) = \lim_{t\to 0} \int_a^b -\frac{1}{\pi}\mathsf{Im}(G_\mu(s+it))ds.$$

Moreover, if μ *has a continuous density* $\rho(x)$*, then*

$$\rho(x) = \lim_{t\to 0} -\frac{1}{\pi}\mathsf{Im}(G_\mu(x+it)).$$

This implies, after some bookkeeping, the following results.

Theorem 2.7.3. *The following hold:*

1. *If* μ *and* ν *are probability measures for which* $G_\mu(z) = G_\nu(z)$ *for all* $z \in \mathbb{C} \setminus \mathbb{R}$*, then* $\mu = \nu$*.*

2. *If* μ_n, μ *are probability measures such that* $G_{\mu_n}(z) \to G_\mu(z)$ *for all* $z \in \mathbb{C} \setminus \mathbb{R}$*, then* $\mu_n \to \mu$ *(in the sense of weak convergence).*

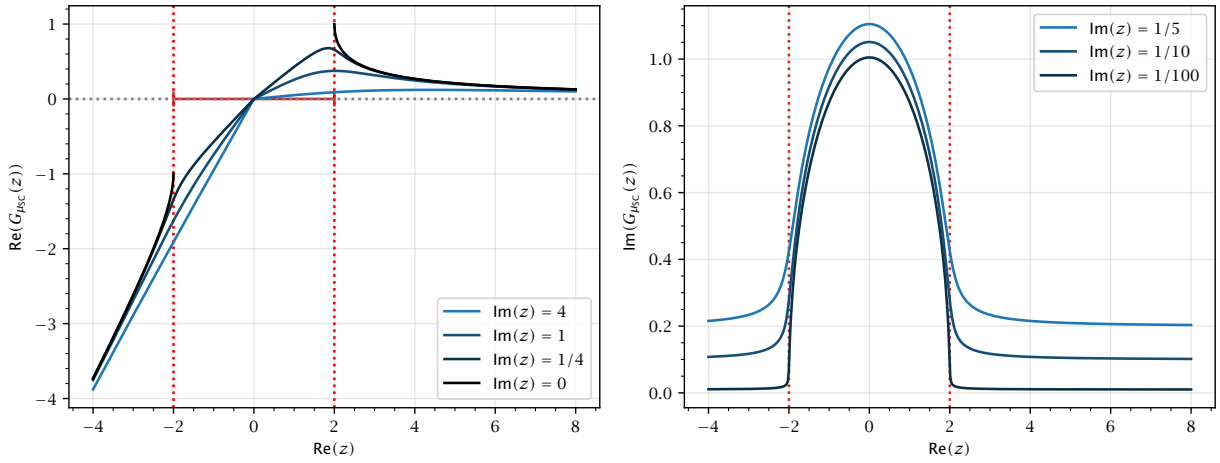


Figure 2.6: Plots of the real and imaginary parts of $G_{\mu_{\mathrm{SC}}}(z)$ for various choices of $\mathrm{Im}(z) > 0$ as $\mathrm{Re}(z)$ varies. Observe that the function is only defined away from $[-2,2]$, and that $\mathrm{Im}(G_{\mu_{\mathrm{SC}}}(z))$ indeed approximates the semicircle density as $\mathrm{Im}(z) \to 0$. The branch cut of the function $z \mapsto \sqrt{z^2-4}$ is the interval $[-2,2]$; while $\mathrm{Re}(G_{\mu_{\mathrm{SC}}}(z))$ has a well-defined limit on this interval, $\mathrm{Im}(G_{\mu_{\mathrm{SC}}}(z))$ changes sign and thus the function cannot be extended to the interval.

3. *If μ_n are random probability measures and μ is a deterministic one and $G_{\mu_n}(z) \xrightarrow{\mathbb{P}} G_\mu(z)$ for all $z \in \mathbb{C} \setminus \mathbb{R}$, then $\mu_n \xrightarrow{\mathbb{P}} \mu$ (that is, with weak convergence in probability).*

The upshot is that we may prove limit theorems by proving convergence of Stieltjes transforms. The final insight making this useful is that the Stieltjes transform of an empirical distribution of eigenvalues also has a natural interpretation in matrix algebra:

$$G_{\mathrm{esd}(\boldsymbol{W})}(z) = \frac{1}{d}\sum_{i=1}^{d} \frac{1}{z - \lambda_i(\boldsymbol{W})} = \frac{1}{d}\,\mathrm{Tr}(z\boldsymbol{I}_d - \boldsymbol{W})^{-1}.$$

The matrix $(z\boldsymbol{I}_d - \boldsymbol{W})^{-1}$ appearing here is called the *resolvent* and has many uses in random matrix theory.

2.7.1 Sketch of Second Proof of Semicircle Limit Theorem

Let us sketch how working with the Stieltjes transform would give an alternative proof of the semicircle limit theorem. First, the following is a straightforward calculation of an integral.

Proposition 2.7.4. *$G_{\mu_{\mathrm{SC}}}(z) = \frac{1}{2}(z - \sqrt{z^2-4})$ for all $z \in \mathbb{C} \setminus [-2,2]$.*

We will also use the following basic linear algebra fact.

Proposition 2.7.5 (Schur complement for matrix inversion)**.** *Suppose that $\boldsymbol{A} \in \mathbb{R}^{m\times m}$, $\boldsymbol{B} \in \mathbb{R}^{m\times n}$, $\boldsymbol{C} \in \mathbb{R}^{n\times m}$, and $\boldsymbol{D} \in \mathbb{R}^{n\times n}$. If $\boldsymbol{D}$ and $\boldsymbol{S} := \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}$ are invertible, then*

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{S}^{-1} & -\boldsymbol{S}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{S}^{-1} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{S}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \end{bmatrix}.$$

The same also holds for complex-valued matrices.

This scary-looking result is just keeping track of the Gaussian elimination derivation of the matrix inverse; the matrix $\boldsymbol{S}$ is called the *Schur complement* of $\boldsymbol{D}$.

Now, suppose that $\boldsymbol{W}^{(d)} \sim \mathrm{Wig}(\mu, d)$ for some μ of mean zero and variance 1. Write $\widehat{\boldsymbol{W}}^{(d)} := \frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}$. We have

$$G_{\mathsf{esd}(\widehat{\boldsymbol{W}}^{(d)})}(z) = \frac{1}{d}\sum_{i=1}^{d}((z\boldsymbol{I}_d - \widehat{\boldsymbol{W}}^{(d)})^{-1})_{ii}$$

and, using Proposition 2.7.5 to derive the value of the single entry (i,i) of this matrix inverse,

$$= \frac{1}{d}\sum_{i=1}^{d}\frac{1}{z - \widehat{\boldsymbol{w}}_i^\top(z\boldsymbol{I}_{d-1} - \widehat{\boldsymbol{W}}_{\sim i}^{(d)})^{-1}\widehat{\boldsymbol{w}}_i},$$

where $\widehat{\boldsymbol{W}}_{\sim i}^{(d)}$ denotes the matrix with the ith row and column both removed, and $\widehat{\boldsymbol{w}}_i$ is the ith row of $\widehat{\boldsymbol{W}}^{(d)}$ with the ith entry removed (leaving a $(d-1)$-dimensional vector). Note also that we are using $\widehat{W}_{ii}^{(d)} = 0$, though this is not an essential assumption here. The next step is to note that, by symmetry, the terms of this sum are identically distributed. While they are not independent, we might hope they are weakly dependent, justifying the law of large numbers approximation (also sometimes called a "self-averaging" property in such contexts)

$$\approx \mathbb{E}\frac{1}{z - \widehat{\boldsymbol{w}}_1^\top(z\boldsymbol{I}_{d-1} - \widehat{\boldsymbol{W}}_{\sim 1}^{(d)})^{-1}\widehat{\boldsymbol{w}}_1}$$

Now, if we believe the random variable in the denominator is concentrated around its mean, we should also be able to approximate

$$\approx \frac{1}{z - \mathbb{E}\widehat{\boldsymbol{w}}_1^\top(z\boldsymbol{I}_{d-1} - \widehat{\boldsymbol{W}}_{\sim 1}^{(d)})^{-1}\widehat{\boldsymbol{w}}_1}$$

Finally, note that $\widehat{\boldsymbol{w}}_1$ and $(z\boldsymbol{I}_{d-1} - \widehat{\boldsymbol{W}}_{\sim 1}^{(d)})^{-1}$ are independent. Thus, since $\widehat{\boldsymbol{w}}_1$ has independent entries that are centered and have variance $1/d$, expanding this quadratic form lets us evaluate

$$\begin{aligned}
&= \frac{1}{z - \frac{1}{d}\sum_{i=1}^{d-1}\mathbb{E}((z\boldsymbol{I}_{d-1} - \widehat{\boldsymbol{W}}_{\sim 1}^{(d)})^{-1})_{ii}} \\
&= \frac{1}{z - \frac{d-1}{d}\mathbb{E}G_{\mathsf{esd}(\widehat{\boldsymbol{W}}^{(d-1)})}(z)}.
\end{aligned}$$

We have argued both that these Stieltjes transforms, evaluated at a given z, converge to deterministic numbers, and that those numbers should satisfy a "self-consistency" equation as $d \to \infty$. Namely, we expect $G_{\mathsf{esd}(\widehat{\boldsymbol{W}}^{(d)})}(z) \xrightarrow{\mathbb{P}} G(z)$, for $G(z)$ a solution of

$$G(z) = \frac{1}{z - G(z)}.$$

Solving this quadratic equation gives precisely $G(z) = G_{\mu_{\mathsf{SC}}}(z)$.

Remark 2.7.6. *Actually, solving the quadratic equation would reach a juncture where we must choose one of two signs in $G(z) = \frac{1}{2}(z \pm \sqrt{z^2-4})$. To choose the right sign, we may refer back to the moment generating function expression (2.7.1). This leads us to expect that, as $z \to 0$, $G(z)$ should scale like $1/z$ (this is true for* any *Stieltjes transform, not just that of the semicircle law). Above, the two choices of signs give the scalings $G(z) = \frac{1}{2}(z \pm \sqrt{z^2-4}) = \frac{z}{2}(1 \pm \sqrt{1-\frac{4}{z^2}}) \approx \frac{z}{2}(1 \pm (1-\frac{2}{z^2}))$, and only choosing the minus sign gives the correct behavior.*

The key to making this precise is to argue, slightly differently, that with high probability we have $\widehat{\boldsymbol{w}}_1^\top (z\boldsymbol{I}_{d-1} - \widehat{\boldsymbol{W}}_{\sim 1}^{(d)})^{-1} \widehat{\boldsymbol{w}}_1 \approx \frac{1}{d} \mathsf{Tr}(z\boldsymbol{I}_{d-1} - \widehat{\boldsymbol{W}}_{\sim 1}^{(d)})^{-1}$, which makes rigorous the approximate recursion alluded to above. In fact, it is possible to show that with high probability $\widehat{\boldsymbol{w}}_1^\top \boldsymbol{X} \widehat{\boldsymbol{w}}_1 \approx \frac{1}{d}\mathsf{Tr}(\boldsymbol{X})$; this is a general property of quadratic forms with well-behaved random vectors. One strong version of such a result is the *Hanson-Wright inequality*; see [RV13].

2.8 Exercises

Exercise 2.8.1. *Let $\boldsymbol{G} \in \mathbb{R}^{d\times d}$ have i.i.d. entries distributed as $\mathcal{N}(0,1)$ (with no symmetry constraint). You will study the random variable $|\det(\boldsymbol{G})|$, one interpretation of which is the volume of the random parallelopiped generated by d independent standard Gaussian vectors.*

1. *Show that $\mathbb{E}[|\det(\boldsymbol{G})|] \le \sqrt{d!}$. (Do something much simpler than using Part 2 below.)*
2. *Let $\boldsymbol{g}^{(k)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ for $k = 1, \dots, d$, drawn independently (that is, $\boldsymbol{g}^{(k)}$ is a standard Gaussian vector in $\mathbb{R}^k$). Show that $|\det(\boldsymbol{G})|$ has the same law as $\prod_{k=1}^d \|\boldsymbol{g}^{(k)}\|$.*

   *(**Hint:** Consider the QR decomposition of $\boldsymbol{G}$.)*
3. *Show that, for a constant $c > 0$, $\mathbb{E}[|\det(\boldsymbol{G})|] \ge \frac{c}{d}\sqrt{d!}$, almost matching the upper bound of Part 1.*

   *(**Hint:** Prove that $\sqrt{x} \ge \frac{1}{2}(1 + x - (x-1)^2)$ for all $x \ge 0$. Apply this with $x := \|\boldsymbol{g}^{(k)}\|^2/k$.)*
4. *Show that, for all square $\boldsymbol{G}$ (not random), $|\det(\boldsymbol{G})| = \prod_{i=1}^d \sigma_i(\boldsymbol{G})$, where σ_i are the singular values. By computing $\mathbb{E}|\det(\boldsymbol{G})|^2$ (for random Gaussian $\boldsymbol{G}$ again now), make a heuristic but intuitively justified prediction for the value of*

$$\lim_{d\to\infty} \mathbb{E}\left[\frac{1}{d}\sum_{i=1}^d \log\left(\lambda_i\left(\frac{1}{d}\boldsymbol{G}^\top \boldsymbol{G}\right)\right)\right]. \tag{2.8.1}$$

   Confirm that your prediction is compatible with the Marchenko-Pastur limit theorem.

   *(**Hint:** Make a heuristic leap of the form $\mathbb{E}[\log(\cdots)] \approx \log(\mathbb{E}[\cdots])$. Don't be afraid. If you like, speculate about when you expect this to be accurate.)*

Exercise 2.8.2. *This exercise will concern the* Gaussian orthogonal ensemble (GOE) *from Definition 3.2.1. Note that $\boldsymbol{W} \sim \mathsf{GOE}(n)$ is almost surely a symmetric matrix, and thus has real eigenvalues $\lambda_1 \ge \cdots \ge \lambda_n$, whose distribution we will study.*

Below, we write $\chi^2(d,\sigma^2)$ for the law of $\|\boldsymbol{g}\|^2 = g_1^2 + \cdots + g_d^2$ for $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_d)$ (the usual χ^2 distribution, but allowing for rescaling). Similarly, we write $\chi(d,\sigma^2)$ for the law of $\|\boldsymbol{g}\|$.

1. *One small theoretical task: show that the* eigenvalue spacing $\lambda_1 - \lambda_2 \geq 0$ *when $\boldsymbol{W} \sim \mathsf{GOE}(2)$ (a 2×2 matrix) has the law $\chi(2,\sigma^2)$ for some σ^2 (calculate and give this value). Look up and write down the density of this distribution—you will need it later.*

2. *On the computer, sample $\boldsymbol{W} \sim \mathsf{GOE}(n)$ for a sequence of growing n. Go at least up to $n = 1000$. Plot histograms of the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ for a few growing n and make sure you observe convergence to a semicircle shape. Also plot λ_1 and λ_n versus n, taking the mean over several random trials for each value of n and including error bars. Make a prediction about the typical scaling of λ_1 and λ_n (each of the form $\mathbb{E}\lambda_i \sim a_i n^{b_i}$ for $a, b \in \mathbb{R}$), each supported by a convincing plot.*

3. *Now fix a large n, at least $n = 1000$ (the larger the better), and plot a histogram of the distribution of the* bulk spacing $\lambda_{n/2} - \lambda_{n/2+1}$ *over many independent draws of $\boldsymbol{W}$ (at least 2000). Come up with a procedure to try to find a good σ^2 to approximate this distribution by $\chi(2,\sigma^2)$ (i.e., to approximate the distribution of spacings for large n by a rescaling of the closed form distribution of spacings you found for $n = 2$ in Part 1). You can define a reasonable "loss function" of σ^2 and use any optimization library your language has to minimize it, for instance. Draw a plot to illustrate the quality of the fit.*

4. *Consider another distribution of $(\lambda_1, \ldots, \lambda_n)$ where λ_i are chosen uniformly at random independently in the interval that you conjectured $[\lambda_n, \lambda_1]$ to scale like in Part 1 (that is, n independent draws from a distribution of the form $\mathsf{Unif}([a_n n^{b_n}, a_1 n^{b_1}])$). Repeat the spacing experiment: fix the same n as in Part 3, sample n independent numbers $\lambda_1, \ldots, \lambda_n$ uniformly at random in the predicted interval, sort them to form $\tilde{\lambda}_1 \leq \cdots \leq \tilde{\lambda}_n$, and plot a histogram of the spacing $\tilde{\lambda}_{n/2} - \tilde{\lambda}_{n/2+1}$ over many independent trials of this procedure. Comment on the differences between the distribution of actual eigenvalue spacings and the distribution of spacings under this alternative model. What does this say about the structure of the eigenvalues? (Focus on the behavior of these distributions near zero.)*

5. *Download a list of the imaginary parts of the first 100 000 non-trivial zeros of the Riemann zeta function (the famous Riemann Hypothesis conjectures that the real parts of all such zeros are equal to $\frac{1}{2}$) from this website:*

   https://www-users.cse.umn.edu/~odlyzko/zeta_tables/zeros1

   Calculate the differences between consecutive values (the distances between consecutive zeros along the imaginary axis.). Plot a histogram of the spacings. You should observe similar qualitative phenomena to before. Repeat the procedure you chose before to find σ^2 to fit the density of $\chi(2,\sigma^2)$ to this distribution. Find another d such that $\chi(d,\sigma^2)$ for some σ^2 achieves an exceptionally good fit. Illustrate the best choice of d (and σ^2) by plotting this density over the histogram of spacings of zeros.

Exercise 2.8.3. *We have seen 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.

1. *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$. Show the perturbation inequality*

$$\min_{\pi \text{ permutation of } [d]} \sum_{i=1}^{d} (\lambda_i(\boldsymbol{A}) - \lambda_{\pi(i)}(\boldsymbol{B}))^2 \le \|\boldsymbol{A} - \boldsymbol{B}\|_F^2.$$

   You may use the Birkhoff–von Neumann theorem*, which states that the set of doubly stochastic $d \times d$ matrices (i.e., $\boldsymbol{P} \in \mathbb{R}^{d\times d}$ such that $P_{ij} \ge 0$ for all $i, j \in [d]$, $\sum_j P_{ij} = 1$ for all $i \in [d]$, and $\sum_i P_{ij} = 1$ for all $j \in [d]$) is the convex hull of the set of the $d \times d$ permutation matrices (those $\boldsymbol{P}$ with exactly one 1 in each row and each column and all other entries 0, of which there are $d!$). You may also use that a linear function over a convex polytope is maximized at one of the vertices.*

   *(**Hint:** Write the spectral decomposition of $\boldsymbol{A}$ and $\boldsymbol{B}$. Consider the matrix of $\langle \boldsymbol{u}_i, \boldsymbol{v}_j\rangle^2$ for $\boldsymbol{u}_i$ eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{v}_j$ eigenvectors of $\boldsymbol{B}$.)*

2. *Prove that, for arbitrary numbers $\lambda_1 \ge \cdots \lambda_d$ and $\rho_1 \ge \cdots \ge \rho_d$,*

$$\min_{\pi \text{ permutation of } [d]} \sum_{i=1}^{d} (\lambda_i - \rho_{\pi(i)})^2 = \sum_{i=1}^{d} (\lambda_i - \rho_i)^2.$$

3. *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 $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$,*

$$\left| \frac{1}{d}\sum_{i=1}^{d} f(\lambda_i(\boldsymbol{A})) - \frac{1}{d}\sum_{i=1}^{d} f(\lambda_i(\boldsymbol{B})) \right| \le \frac{K}{\sqrt{d}}\|\boldsymbol{A} - \boldsymbol{B}\|_F.$$

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

   *(**Hint:** Define a version of $\boldsymbol{W} = \boldsymbol{W}^{(d)}$ where entries W_{ij} are replaced with the centered* truncations *$W_{ij}\mathbb{1}\{|W_{ij}| \le C\} - \mathbb{E}[W_{ij}\mathbb{1}\{|W_{ij}| \le C\}]$ for a large C and use the limit theorem we have shown already, as cited above, on this matrix.)*

5. *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 (the statement that $\|\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}\| \to 2$ in probability) does require further moment assumptions on ν.*

   *(**Hint:** Prove and use that $\|\boldsymbol{W}\| \ge \max_{i,j\in[d]} |W_{ij}|$. 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.)*

Exercise 2.8.4. *This problem is a continuation of Exercise 1.6.5. Further problems in the sequence will have you derive powerful consequences of these ideas for random matrices. For now, you will derive some more general tools.*

1. *Suppose $F : \mathbb{R}^d \to \mathbb{R}$ is a smooth function with $\max\{|F(\boldsymbol{x})|, \|\nabla F(\boldsymbol{x})\|_2^2, \|\nabla^2 F(\boldsymbol{x})\|_F^2\} \le C(1+\|\boldsymbol{x}\|)^K$ for some $C, K > 0$ and all $\boldsymbol{x} \in \mathbb{R}^d$, where $\nabla^2 F$ is the $d \times d$ Hessian matrix of second derivatives. Let $\boldsymbol{\Sigma}, \boldsymbol{\Lambda} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ be positive semidefinite. Define $\boldsymbol{\Sigma}(t) := (1-t)\boldsymbol{\Sigma} + t\boldsymbol{\Lambda}$ for $t \in [0,1]$, and write*

$$f(t) := \mathop{\mathbb{E}}_{\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}(t))} F(\boldsymbol{g}).$$

That is, we are considering the value of an expectation of a general function of a Gaussian vector as the covariance matrix moves along a line in matrix space. Show that the derivative of this value is

$$f'(t) = \frac{1}{2}\left\langle \boldsymbol{\Lambda} - \boldsymbol{\Sigma}, \mathop{\mathbb{E}}_{\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}(t))} \nabla^2 F(\boldsymbol{g}) \right\rangle.$$

Here, $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathsf{Tr}(\boldsymbol{A}\boldsymbol{B}) = \sum_{i,j} A_{ij}B_{ij}$ is the Frobenius inner product.

You may differentiate under the expectation (i.e., bring a derivative inside an expectation) without justification, but you should consider on your own time what the justification would be.

*(**Hint:** If $\boldsymbol{g} \sim \mathcal{N}(0,\Sigma)$ and $\boldsymbol{h} \sim \mathcal{N}(0,\Lambda)$ independently, construct a Gaussian vector with covariance $\Sigma(t)$ to make differentiating under the expectation easier. Then, use Exercise 1.6.5.)*

2. *Show that, if F as above is also convex, and $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})$ and $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{\Gamma})$ are independent Gaussian vectors (that is, the entries of $\boldsymbol{g}$ may be correlated with one another, and likewise for $\boldsymbol{h}$, but entries of $\boldsymbol{g}$ are independent of entries of $\boldsymbol{h}$) for any $\boldsymbol{\Sigma}, \boldsymbol{\Gamma} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ positive semidefinite, then*

$$\mathbb{E}F(\boldsymbol{g}) \le \mathbb{E}F(\boldsymbol{g}+\boldsymbol{h}).$$

Informally, expectations of convex functions of Gaussians are only increased by adding noise. Show that the same also holds for $F(\boldsymbol{x}) = \max_{i\in[d]} x_i$, though it is not smooth.

*(**Hint:** $\mathsf{Law}(g+h) = \mathcal{N}(0,\Lambda)$ for some Λ—write this out and use Part 1. For the last part, consider the "soft-max" function $F(x) = \beta^{-1}\log(\sum_{i=1}^d \exp(\beta x_i))$ and take $\beta \to \infty$.)*

3. *Suppose that $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})$ and $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{\Lambda})$ are arbitrary centered Gaussian vectors as in Part 1. Suppose that, for all $i, j \in [d]$, we have $\mathbb{E}(g_i - g_j)^2 \le \mathbb{E}(h_i - h_j)^2$. Show that*

$$\mathbb{E}\max_{i\in[d]} g_i \le \mathbb{E}\max_{i\in[d]} h_i.$$

*(**Hint:** Expand the condition on $\boldsymbol{g}$ and $\boldsymbol{h}$ into a condition on Σ and Λ. Again consider the soft-max function and use Part 1, explicitly calculating the Hessian.)*

3 | Free Probability

3.1 Warmup: Central Limit Theorem by Renormalization

We suggested above a curious relationship between the CLT and the semicircle limit theorem to do with the moments of the limiting distributions: in the former case they count all matchings, and in the latter case only non-crossing matchings. As an entry point into free probability, let us make our way towards a more precise analogy between the two theorems.

To do this, we sketch another approach to the CLT. Consider the operation $\mathcal{S}$ mapping probability measures on $\mathbb{R}$ to probability measures on $\mathbb{R}$ where

$$\mathcal{S}(\mu) := \mathsf{Law}\left(\frac{X_1 + X_2}{\sqrt{2}}\right) \text{ for } X_1, X_2 \overset{\text{iid}}{\sim} \mu.$$

We then have that iterating this map gives

$$\mathcal{S}^{(k)}(\mu) := \mathsf{Law}\left(\frac{X_1 + X_2 + \cdots + X_{2^k}}{\sqrt{2^k}}\right) \text{ for } X_i \overset{\text{iid}}{\sim} \mu,$$

and thus the CLT (modulo the issue of the number of summands being restricted to powers of 2, which we will not deal with here) amounts to the elegant statement

$$\mathcal{S}^{(k)}(\mu) \xrightarrow[k\to\infty]{} \mathcal{N}(0,1)$$

for any μ with mean zero and variance 1. Indeed, $\mathcal{S}(\mathcal{N}(0,1)) = \mathcal{N}(0,1)$, so this perspective lets us view the CLT as a *fixed point theorem* about the operation $\mathcal{S}$.

Moreover, we may use this fixed point property to derive the moments of $\mathcal{N}(0,1)$, if we did not know them already. Suppose μ is some probability measure with mean zero, variance 1, and $\mathcal{S}(\mu) = \mu$, and write m_k for the kth moment of μ. We may then derive

$$\begin{aligned}
m_k &= \mathop{\mathbb{E}}_{X\sim\mu} X^k \\
&= \mathop{\mathbb{E}}_{X_1,X_2\sim\mu} \left(\frac{X_1 + X_2}{\sqrt{2}}\right)^k \\
&= \frac{1}{2^{k/2}} \sum_{a=0}^{k} \binom{k}{a} \mathbb{E} X_1^a X_2^{k-a} \\
&= \frac{1}{2^{k/2}} \sum_{a=0}^{k} \binom{k}{a} m_a m_{k-a} \\
&= \frac{2}{2^{k/2}} m_k + \frac{1}{2^{k/2}} \sum_{a=1}^{k-1} \binom{k}{a} m_a m_{k-a},
\end{aligned}$$

whereby we may solve to find

$$m_k = \frac{1}{2^{k/2} - 2} \sum_{a=1}^{k-1} \binom{k}{a} m_a m_{k-a} \tag{3.1.1}$$

for all $k \geq 3$, with the "initial conditions" $m_0 = 1$, $m_1 = 0$, $m_2 = 1$. You may check that, indeed, the solution to this recursion gives the moments of $\mathcal{N}(0,1)$ discussed before.

To summarize, the limiting distribution $\mu = \mathcal{N}(0,1)$ appearing in the CLT is uniquely determined by the following simple properties:

$$\begin{aligned} \mathop{\mathbb{E}}_{X\sim\mu} X &= 0, \\ \mathop{\mathbb{E}}_{X\sim\mu} X^2 &= 1, \\ \mathcal{S}(\mu) &= \mu. \end{aligned}$$

For more on this approach, including precise proofs of the CLT by this kind of argument (using, curiously, the entropy of a probability measure as a "potential" or "Lyapunov function" that is monotonically increasing with respect to applications of the map $\mathcal{S}$), see the original reference [Lin59], the more recent [Bar86, Ans99, ABBN04, Ott23] or the textbook treatment in [KS07, Section 10.3].

3.2 Tangled Joint Moments

We might wonder if we can take the same approach as above to the semicircle limit theorem (our Theorem 2.4.3), and learn a fixed point interpretation of the limit theorem and a similar characterization of the semicircle law. Indeed, the natural Gaussian random symmetric matrix is also a fixed point of $\mathcal{S}$, if we view $\mathcal{S}$ now as acting on probability measures over random symmetric matrices.

Definition 3.2.1 (Gaussian orthogonal ensemble)**.** *The* Gaussian orthogonal ensemble (GOE)*, denoted* $\mathsf{GOE}(d)$*, is the law of a* $d \times d$ *symmetric matrix* $\boldsymbol{W}$ *where* $W_{ij} = W_{ji} \sim \mathcal{N}(0,1)$ *and* $W_{ii} \sim \mathcal{N}(0,2)$*, with all entries with* $i \leq j$ *distributed independently.*

The reason for the specific scaling of the diagonal of the GOE is as follows.

Proposition 3.2.2. *For any* $\boldsymbol{Q} \in \mathcal{O}(n)$*, if* $\boldsymbol{W} \sim \mathsf{GOE}(d)$*, then* $\mathsf{Law}(\boldsymbol{W}) = \mathsf{Law}(\boldsymbol{Q}\boldsymbol{W}\boldsymbol{Q}^\top)$*.*

On the other hand, the diagonal does not contribute materially to the limiting spectral moments, and $\boldsymbol{W}^{(d)} \sim \mathsf{GOE}(d)$ still satisfy the conclusion of Theorem 2.4.3, i.e., $\mathsf{esd}(\frac{1}{\sqrt{d}}\boldsymbol{W}^{(d)}) \to \mu_{\mathsf{SC}}$. And, we indeed have $\mathcal{S}(\mathsf{GOE}(d)) = \mathsf{GOE}(d)$.

Suppose now that we have some sequence of probability measures $\mu^{(d)}$ on $\mathbb{R}^{d\times d}_{\mathsf{sym}}$ which satisfy $\mathcal{S}(\mu^{(d)}) = \mu^{(d)}$ and which converge in expected moments after rescaling as we do for Wigner matrices to some μ,

$$\lim_{d\to\infty} \frac{1}{d} \mathop{\mathbb{E}}_{\boldsymbol{W}\sim\mu^{(d)}} \mathsf{Tr}\left(\frac{1}{\sqrt{d}}\boldsymbol{W}\right)^k = m_k = \mathbb{E}_{X\sim\mu}[X^k].$$

Can we show, without using the structure of the GOE specifically, that, if $m_2 = 1$, then the m_k can only be the moments of $\mu = \mu_{\mathsf{SC}}$, coinciding with the limit achieved by the sequence $\mu^{(d)} = \mathsf{GOE}(d)$ (and by more general Wigner matrices)?

You might immediately notice an issue: the conditions

$$\begin{aligned} \mathop{\mathbb{E}}_{X\sim\mu} X &= 0, \\ \mathop{\mathbb{E}}_{X\sim\mu} X^2 &= 1, \\ \mathcal{S}(\mu^{(d)}) &= \mu^{(d)}. \end{aligned}$$

in fact do not uniquely characterize the GOE or its semicircle limit. These conditions are also satisfied by a sequence of laws $\mu^{(d)}$ of growing diagonal matrices of i.i.d. Gaussians, replicating in the matrix setting the scalar CLT setup. Namely, consider $\boldsymbol{D}^{(d)} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ diagonal with $D_{ii} \sim \mathcal{N}(0,d)$ independently. Let us write $\mu^{(d)} = \mathsf{Diag}(\mathcal{N}(0,d))$ for this model. Then, these laws satisfy $\mathcal{S}(\mu^{(d)}) = \mu^{(d)}$, and the limiting distribution (in expected moments) is $\mu = \mathcal{N}(0,1)$. On the other hand, the $\mu^{(d)} = \mathsf{GOE}(d)$ *also* satisfy $\mathcal{S}(\mu^{(d)}) = \mu^{(d)}$, while the limiting distribution is the semicircle $\mu = \mu_{\mathsf{SC}}$. (You may check that both limiting distributions also satisfy the mean and variance conditions above.)

Thus understanding the semicircle limit theorem as a fixed point theorem will require a different approach. It is instructive, though, to see how exactly this ambiguity arises when we try to imitate the above derivation of a combinatorial recursion. Let $\boldsymbol{W} = \boldsymbol{W}^{(d)}, \boldsymbol{W}_1 = \boldsymbol{W}_1^{(d)}, \boldsymbol{W}_2 = \boldsymbol{W}_2^{(d)} \sim \mu^{(d)}$ independently; we leave the superscripts implicit below to lighten the notation. As before, write

$$m_k := \lim_{d\to\infty} \mathbb{E}\frac{1}{d}\,\mathsf{Tr}\left(\frac{1}{\sqrt{d}}\boldsymbol{W}\right)^k.$$

Following the idea from the treatment of the CLT, we expand:

$$\begin{aligned} m_k &= \lim_{d\to\infty} \mathbb{E}\frac{1}{d}\,\mathsf{Tr}\left(\frac{1}{\sqrt{d}}\boldsymbol{W}\right)^k \\ &= \lim_{d\to\infty} \frac{1}{d^{k/2+1}}\mathbb{E}\,\mathsf{Tr}\,\boldsymbol{W}^k \\ &= \lim_{d\to\infty} \frac{1}{d^{k/2+1}}\mathbb{E}\,\mathsf{Tr}\left(\frac{\boldsymbol{W}_1+\boldsymbol{W}_2}{\sqrt{2}}\right)^k \\ &= \frac{1}{2^{k/2}} \sum_{a_1,\dots,a_k\in\{1,2\}} \lim_{d\to\infty} \frac{1}{d^{k/2+1}}\mathbb{E}\,\mathsf{Tr}\,\boldsymbol{W}_{a_1}\cdots\boldsymbol{W}_{a_k}. \end{aligned} \tag{3.2.1}$$

Unlike in the scalar case, the $\boldsymbol{W}_i$ do not necessarily commute. In fact, that is the key difference between the two cases above: the diagonal choice of $\boldsymbol{W}_i$ do commute, while the GOE-distributed choice do not. But let us continue and see how this affects the calculation very concretely. There are a few terms above that we can compute directly.

- Most simply, we have by definition

$$\lim_{d\to\infty} \frac{1}{d^{k/2+1}}\mathbb{E}\,\mathsf{Tr}(\boldsymbol{W}_1^k) = \lim_{d\to\infty} \frac{1}{d^{k/2+1}}\,\mathsf{Tr}(\boldsymbol{W}_2^k) = m_k.$$

- More generally, you may also show that, for both of the above two choices of $\mu^{(d)}$ (GOE and diagonal Gaussian) as well as for quite general classes of "sufficiently symmetric" models, $\mathbb{E}\boldsymbol{W}^k = \frac{1}{d}(\mathbb{E}\,\mathrm{Tr}\,\boldsymbol{W}^k)\boldsymbol{I}_d$. Thus we can handle terms of the form

$$\begin{aligned}\lim_{d\to\infty}\frac{1}{d^{k/2+1}}\mathbb{E}\,\mathrm{Tr}(\boldsymbol{W}_1^a\boldsymbol{W}_2^{k-a}) &= \lim_{d\to\infty}\frac{1}{d^{k/2+1}}\cdot\frac{1}{d^2}\cdot\mathbb{E}\,\mathrm{Tr}(\boldsymbol{W}_1^a)\cdot\mathbb{E}\,\mathrm{Tr}(\boldsymbol{W}_2^{k-a})\cdot\mathrm{Tr}(\boldsymbol{I}_d)\\ &= \lim_{d\to\infty}\frac{1}{d^{a/2+1}}\cdot\mathbb{E}\,\mathrm{Tr}(\boldsymbol{W}_1^a)\cdot\frac{1}{d^{(k-a)/2+1}}\mathbb{E}\,\mathrm{Tr}(\boldsymbol{W}_2^{k-a})\\ &= m_a m_{k-a}. \end{aligned} \tag{3.2.2}$$

For $\mu^{(d)}$ the diagonal Gaussian law, all terms in (3.2.1) may be treated by the above rule, since the $\boldsymbol{W}_i$ commute and may be gathered into like terms. By applying this rule, we find a recursion for the m_k, which is just the same one as in (3.1.1), describing the moments of $\mathcal{N}(0,1)$.

But when $\mu^{(d)}$ is the GOE, this rule is not enough to recursively compute all the terms appearing in (3.2.1). We can even throw in one more rule:

- By the cyclic property of trace, terms of the form $\mathrm{Tr}(\boldsymbol{W}_1^a\boldsymbol{W}_2^{k-a-b}\boldsymbol{W}_1^b)$ may be brought into the above form as

$$\mathrm{Tr}(\boldsymbol{W}_1^a\boldsymbol{W}_2^{k-a-b}\boldsymbol{W}_1^b) = \mathrm{Tr}(\boldsymbol{W}_1^{a+b}\boldsymbol{W}_2^{k-a-b}),$$

  and similarly with the roles of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ reversed.

This lets us treat a few more terms, but still not all of them. Let us see that some other reasoning specific to the GOE is necessary.

Example 3.2.3. *Consider $k = 4$. We know already that, for the GOE, $m_4 = \mathrm{Cat}_2 = 2$. But let us see if we can derive this recursively from (3.2.1), supposing we know already that $m_0 = 1$, $m_1 = m_3 = 0$, and $m_2 = \mathrm{Cat}_1 = 1$ (we are substituting in moments of the semicircle distribution here). We have $k/2+1 = 3$ and $2^{k/2} = 4$, so that equation reads*

$$m_4 = \frac{1}{4}\sum_{a_1,a_2,a_3,a_4\in\{1,2\}}\lim_{d\to\infty}\frac{1}{d^3}\mathbb{E}\,\mathrm{Tr}(\boldsymbol{W}_{a_1}\boldsymbol{W}_{a_2}\boldsymbol{W}_{a_3}\boldsymbol{W}_{a_4}).$$

Note that, by symmetry, all terms where each $\boldsymbol{W}_i$ occurs an odd number of times (1 or 3, in this case) equal zero. Our above observations give values for the limits of most of the remaining terms, as shown in Table 3.1. The remaining two terms are

$$\lim_{d\to\infty}\frac{1}{d^3}\mathbb{E}\,\mathrm{Tr}(\boldsymbol{W}_1\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{W}_2),$$

and the similar term with $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ having reversed roles. In fact both are equal by the cyclic property of the trace. Since the previous terms already contribute 2 to m_4, we expect to have

$$\lim_{d\to\infty}\frac{1}{d^3}\mathbb{E}\,\mathrm{Tr}(\boldsymbol{W}_1\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{W}_2) = 0.$$

This is true: you may check it by expanding this "mixed moment" in a similar fashion to our treatment of the moments of a single Wigner matrix in Lemma 2.4.7. But it cannot

a_1	a_2	a_3	a_4	$\lim_{d\to\infty} \mathbb{E}\,\mathsf{Tr}(\boldsymbol{W}_{a_1}\boldsymbol{W}_{a_2}\boldsymbol{W}_{a_3}\boldsymbol{W}_{a_4})/d^3$
1	1	1	1	$2 = m_4$
2	2	2	2	$2 = m_4$
1	1	2	2	$1 = m_2 \cdot m_2$
2	2	1	1	$1 = m_2 \cdot m_2$
1	2	2	1	$1 = m_2 \cdot m_2$
2	1	1	2	$1 = m_2 \cdot m_2$
1	2	1	2	0?
2	1	2	1	0?

Table 3.1: The limiting expectations of traces of various words of length 4 in two independent GOE matrices $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, as discussed in Example 3.2.3.

be derived without some reasoning specific to the GOE; indeed, it is not true of the diagonal Gaussian matrix discussed above. Indeed, you may check that, if instead $\boldsymbol{W}_i \sim \mathsf{Diag}(\mathcal{N}(0,d))$ independently, then

$$\lim_{d\to\infty} \frac{1}{d^3} \mathop{\mathbb{E}}_{\boldsymbol{W}_i\sim\mathsf{Diag}(\mathcal{N}(0,d))} \mathsf{Tr}(\boldsymbol{W}_1\boldsymbol{W}_2\boldsymbol{W}_1\boldsymbol{W}_2) = \lim_{d\to\infty} \frac{1}{d^3} \mathop{\mathbb{E}}_{\boldsymbol{W}_i\sim\mathsf{Diag}(\mathcal{N}(0,d))} \mathsf{Tr}(\boldsymbol{W}_1^2\boldsymbol{W}_2^2) = 1.$$

That this term is negligible in the GOE case does not follow merely from the independence of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ and the behaviors of their individual moments—we must be using some other relationship between these two matrices. While we only have the blunt tool of expanding traces of products available to derive this behavior for now, you might think that at least the overall behavior of the recursion is simple in that all "tangled" joint moments as above—those that cannot be reduced to $\mathsf{Tr}(\boldsymbol{W}_1^a\boldsymbol{W}_2^{k-a})$ by the cyclic property of the trace—are negligible. But that is also wrong, as the following example shows.

Example 3.2.4. *Consider $k = 8$. The expansion of m_8 as above will contain the following term:*

$$\lim_{d\to\infty} \frac{1}{d^5} \mathbb{E}\,\mathsf{Tr}(\boldsymbol{W}_1^2\boldsymbol{W}_2^2\boldsymbol{W}_1^2\boldsymbol{W}_2^2).$$

This limit is non-zero for $\boldsymbol{W}_1, \boldsymbol{W}_2 \sim \mathsf{GOE}(d)$, as you may verify by considering terms in the expanded trace of the form

$$\mathbb{E}(\boldsymbol{W}_1)_{ij}(\boldsymbol{W}_1)_{ji}(\boldsymbol{W}_2)_{ik}(\boldsymbol{W}_2)_{ki}(\boldsymbol{W}_1)_{i\ell}(\boldsymbol{W}_1)_{\ell i}(\boldsymbol{W}_2)_{im}(\boldsymbol{W}_2)_{mi} = 1$$

with i, j, k, ℓ, m distinct, of which there are $\Omega(d^5)$ in the expanded trace.

Our plan to understand the semicircle limit theorem as a fixed point theorem seems to have failed: the above examples are forcing us to expand traces in sums of products of entries to understand some terms in the fixed point expansion, and of course once we are willing to do that we might as well reproduce our original proof of convergence in expected moments to the semicircle law. Yet, without such direct moment calculations we have no way (yet) to determine which terms in the fixed point expansion contribute in the $d \to \infty$ limit or how much they contribute. It seems that we need to somehow change the formulation

of our putative fixed point theorem in a way that "picks out" the behavior of GOE matrices rather than diagonal Gaussian matrices.

We will see below that *free probability* is precisely the theory of how traces of all "words" of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ behave in this kind of situation of independent GOE matrices, and more generally for pairs of "very non-commuting matrices" in a suitable sense. Note that this property (or at least just commutativity vs. non-commutativity) distinguishes a pair of independent GOE matrices from a pair of independent diagonal Gaussian matrices. Moving in this direction, below we will introduce the main definition behind free probability and see how it recursively determines the values of expected traces of words that we ran into above.

3.3 Asymptotic Freeness

The main idea is as follows. It is easier to cross-reference to the previous discussion if we think not of $\boldsymbol{W} \sim \mathsf{GOE}$ but of $\widehat{\boldsymbol{W}} := \frac{1}{\sqrt{d}}\boldsymbol{W}$ (and similarly defining $\widehat{\boldsymbol{W}}_1, \widehat{\boldsymbol{W}}_2$), whose bulk eigenvalues scale as $O(1)$. The below definition, which we will see momentarily is satisfied by the $\widehat{\boldsymbol{W}}_i$, may then be viewed as a wide-ranging generalization of our observation that $\lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathsf{Tr}(\widehat{\boldsymbol{W}}_1\widehat{\boldsymbol{W}}_2\widehat{\boldsymbol{W}}_1\widehat{\boldsymbol{W}}_2) = 0$.

Definition 3.3.1. *Let $\boldsymbol{X} = \boldsymbol{X}^{(d)} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ and $\boldsymbol{Y} = \boldsymbol{Y}^{(d)} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ be random matrices satisfying the convergences of moments*

$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathsf{Tr}\,\boldsymbol{X}^k = m_k,$$
$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathsf{Tr}\,\boldsymbol{Y}^k = n_k.$$

(Note that there are no other assumptions on their joint distribution: they can be deterministic, random but correlated, etc.) We say that the pair of sequences $(\boldsymbol{X}, \boldsymbol{Y})$ are asymptotically free *if, for all $k_1, \dots, k_t \geq 1$ and $\ell_1, \dots, \ell_t \geq 1$, we have*

$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathsf{Tr}\left((\boldsymbol{X}^{k_1} - m_{k_1}\boldsymbol{I}_d)(\boldsymbol{Y}^{\ell_1} - n_{\ell_1}\boldsymbol{I}_d)\cdots(\boldsymbol{X}^{k_t} - m_{k_t}\boldsymbol{I}_d)(\boldsymbol{Y}^{\ell_t} - n_{\ell_t}\boldsymbol{I}_d)\right) = 0. \tag{3.3.1}$$

The point here is that each term has limiting trace zero:

$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathsf{Tr}(\boldsymbol{X}^{k_a} - m_{k_a}\boldsymbol{I}_d) = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathsf{Tr}(\boldsymbol{Y}^{\ell_b} - n_{\ell_b}\boldsymbol{I}_d) = 0.$$

Asymptotic freeness states 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.,

$$\begin{aligned}
&\lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathsf{Tr}\left((\boldsymbol{X}^{k_a} - m_{k_a}\boldsymbol{I}_d)(\boldsymbol{Y}^{\ell_b} - n_{\ell_b}\boldsymbol{I}_d)(\boldsymbol{X}^{k_a} - m_{k_a}\boldsymbol{I}_d)(\boldsymbol{Y}^{\ell_b} - n_{\ell_b}\boldsymbol{I}_d)\right) \\
&\quad = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathsf{Tr}\left((\boldsymbol{X}^{k_a} - m_{k_a}\boldsymbol{I}_d)^2(\boldsymbol{Y}^{\ell_b} - n_{\ell_b}\boldsymbol{I}_d)^2\right) \\
&\quad = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\langle(\boldsymbol{X}^{k_a} - m_{k_a}\boldsymbol{I}_d)^2, (\boldsymbol{Y}^{\ell_b} - n_{\ell_b}\boldsymbol{I}_d)^2\rangle,
\end{aligned}$$

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 $(\boldsymbol{X}^{k_a} - m_{k_a}\boldsymbol{I}_d)^2$ and $(\boldsymbol{Y}^{\ell_b} - m_{\ell_b}\boldsymbol{I}_d)^2$ have all eigenvalues at least $\epsilon > 0$). Generally, you should hold the intuition that

"freeness is maximal non-commutativity."

The following (completely non-obvious and quite difficult) result captures the other side of this intuition, guaranteeing the asymptotic freeness of very many sequences of "nice" random matrices whose eigenvectors are randomized, including the ones we were struggling with before.

Theorem 3.3.2 (Voiculescu)**.** *Let $\boldsymbol{X} = \boldsymbol{X}^{(d)}, \boldsymbol{Y} = \boldsymbol{Y}^{(d)}$ be sequences of random matrices such that $\mathsf{esd}(\boldsymbol{X}^{(d)}) \xrightarrow{\mathbb{E}\ mom.} \mu$ and $\mathsf{esd}(\boldsymbol{Y}^{(d)}) \xrightarrow{\mathbb{E}\ mom.} \nu$. Let $\boldsymbol{Q} = \boldsymbol{Q}^{(d)} \sim \mathsf{Haar}(\mathcal{O}(d))$, drawn independently of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Then, the sequences $(\boldsymbol{X}, \boldsymbol{Q}\boldsymbol{Y}\boldsymbol{Q}^\top)$ are asymptotically free.*

Note that indeed $\boldsymbol{Q}\boldsymbol{Y}\boldsymbol{Q}^\top$ has the same eigenvalues as $\boldsymbol{Y}$, but uniformly random eigenvectors that are independent of those eigenvalues.

Corollary 3.3.3. *Let $\boldsymbol{X}_0^{(d)}, \boldsymbol{Y}_0^{(d)} \sim \mathsf{GOE}(d)$ be independent, and let $\boldsymbol{X}^{(d)} := \frac{1}{\sqrt{d}}\boldsymbol{X}_0^{(d)}$, $\boldsymbol{Y}^{(d)} := \frac{1}{\sqrt{d}}\boldsymbol{Y}_0^{(d)}$. Then, the sequences $(\boldsymbol{X}, \boldsymbol{Y})$ are asymptotically free, with the $m_k = n_k$ the moments of the semicircle law.*

These results are *extremely* powerful for the kinds of calculations we were struggling with above, allowing us to circumvent any expansions of traces in matrix entries.

Example 3.3.4. *It immediately follows that $\lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathsf{Tr}(\widehat{\boldsymbol{W}}_1\widehat{\boldsymbol{W}}_2\widehat{\boldsymbol{W}}_1\widehat{\boldsymbol{W}}_2) = 0$, by taking $t = 2$ and $k_1 = k_2 = \ell_1 = \ell_2 = 1$, noting that $m_1 = 0$.*

Example 3.3.5. *We may also rederive the factorization of traces of "separable" words*

$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathsf{Tr}(\widehat{\boldsymbol{W}}_1^k\widehat{\boldsymbol{W}}_2^\ell).$$

Indeed, expanding the associated "centered" trace that asymptotic freeness controls, we have

$$\begin{aligned} 0 &= \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathsf{Tr}\left((\widehat{\boldsymbol{W}}_1^k - m_k)(\widehat{\boldsymbol{W}}_2^\ell - m_\ell)\right) \\ &= \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathsf{Tr}(\widehat{\boldsymbol{W}}_1^k\widehat{\boldsymbol{W}}_2^\ell) - m_k m_\ell - m_k m_\ell + m_k m_\ell, \end{aligned}$$

which immediately implies

$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathsf{Tr}(\widehat{\boldsymbol{W}}_1^k\widehat{\boldsymbol{W}}_2^\ell) = m_k m_\ell,$$

as we derived earlier.

Example 3.3.6. *Finally, we may also use asymptotic freeness to treat in a simple way the example from earlier of a non-zero tangled joint moment,* $C := \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathrm{Tr}(\widehat{\boldsymbol{W}}_1^2\widehat{\boldsymbol{W}}_2^2\widehat{\boldsymbol{W}}_1^2\widehat{\boldsymbol{W}}_2^2) \neq 0$*. Indeed, by fully expanding the product of sums, asymptotic freeness implies*

$$\begin{aligned}0 &= \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathrm{Tr}((\widehat{\boldsymbol{W}}_1^2 - m_2\boldsymbol{I}_d)(\widehat{\boldsymbol{W}}_2^2 - m_2\boldsymbol{I}_d)(\widehat{\boldsymbol{W}}_1^2 - m_2\boldsymbol{I}_d)(\widehat{\boldsymbol{W}}_2^2 - m_2\boldsymbol{I}_d)) \\ &= C - m_2(4m_2m_4) + m_2^2(4m_2^2 + 2m_4) - m_2^3(4m_2) + m_2^4 \\ &= C - 2m_2^2m_4 + m_2^4,\end{aligned}$$

where we have used Example 3.3.6 on the terms other than C*. Thus,*

$$C = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathrm{Tr}(\widehat{\boldsymbol{W}}_1^2\widehat{\boldsymbol{W}}_2^2\widehat{\boldsymbol{W}}_1^2\widehat{\boldsymbol{W}}_2^2) = 2m_2^2m_4 - m_2^4,$$

and if we know that $m_2 = 1$ *and* $m_4 = 2$*, we would find that* $C = 3$*.*

Much more generally, the following holds, though explicitly writing down the polynomials involved is difficult without further machinery.

Proposition 3.3.7. *Suppose that* $\boldsymbol{X}, \boldsymbol{Y}$ *are an asymptotically free sequence, as in Definition 3.3.1. Then, the value of any limiting mixed moment*

$$\lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\left(\boldsymbol{X}^{k_1}\boldsymbol{Y}^{\ell_1}\cdots\boldsymbol{X}^{k_t}\boldsymbol{Y}^{\ell_t}\right) \tag{3.3.2}$$

is a polynomial of the limiting individual moments m_a, n_b*.*

Proof. By induction on t and expanding the asymptotic freeness condition as in the examples above. □

Thus we have, to some extent, redeemed our approach to the semicircle theorem: applying Proposition 3.3.7 together with Corollary 3.3.3, we get a recursion for the semicircle moments roughly in the style of the one we derived for the Gaussian moments. Sadly, the recursion is not in closed form, and seems hard to write down directly in closed form. We will briefly see later other tools for actually computing with these kinds of identities.

But Proposition 3.3.7 is much deeper than just that. Think of the traces of all words as the "spectral mixed moments" of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Then, the Proposition says that, under asymptotic freeness, the "spectral joint distribution" of $\boldsymbol{X}$ and $\boldsymbol{Y}$ (the spectral mixed moments) is fully determined by the individual "spectral distributions" (the individual spectral moments) of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

You should view this as parallel to the scalar case. If X and Y are scalar random variables, then their individual distributions determine their joint distribution if and only if they are *independent*. Thus asymptotic freeness is one version of scalar independence for the spectra random matrices (for this reason it is also sometimes called *free independence*). In fact, what we have seen is that there are two incompatible kinds of independence that random matrices can have: the individual spectral moments will determine the joint ones if $\boldsymbol{X}$ and $\boldsymbol{Y}$ commute and are independent as random vectors, or if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are asymptotically free, but the resulting joint distributions will be different! We delve further into a special case of this below.

3.4 Free Convolutions

3.4.1 Additive Free Convolution

Let us continue to think about the consequences of asymptotic freeness at an abstract level. The following direct consequence of Proposition 3.3.7 will be critical.

Corollary 3.4.1. *Suppose that $\boldsymbol{X}, \boldsymbol{Y}$ are an asymptotically free sequence, as in Definition 3.3.1. Then, there is a polynomial p_k (whose coefficients are universal constants not depending on $\boldsymbol{X}, \boldsymbol{Y}$, or their moments m_a and n_b) such that the value of any limiting moment of $\boldsymbol{X} + \boldsymbol{Y}$ is*

$$\lim_{d\to\infty} \frac{1}{d} \mathbb{E}\,\mathrm{Tr}(\boldsymbol{X} + \boldsymbol{Y})^k = p_k(m_1, \ldots, m_k, n_1, \ldots, n_k). \tag{3.4.1}$$

The proof is just to use the same procedure from Example 3.3.6 and underlying Proposition 3.3.7 inductively on every term formed by expanding $\mathrm{Tr}(\boldsymbol{X} + \boldsymbol{Y})^k$.

The following is not hard to show, but requires going into some technical hinterlands that we will not have time for.

Proposition 3.4.2. *In the setting of Corollary 3.4.1, if m_k are the moments of a compactly supported measure μ and n_k are the moments of a compactly supported measure ν, then $p_k(m_1, \ldots, m_k, n_1, \ldots, n_k)$ are the moments of another compactly supported measure, which is also uniquely determined by its moments.*

Definition 3.4.3 (Additive free convolution)**.** *We write $\mu \boxplus \nu$ for the measure discussed above, called the* additive free convolution *of μ and ν.*

Thus what we have shown is the following remarkable fact.

Theorem 3.4.4. *Suppose that $\boldsymbol{X}, \boldsymbol{Y}$ are an asymptotically free sequence, as in Definition 3.3.1, where $\mathrm{esd}(\boldsymbol{X}) \xrightarrow{\mathbb{E}\ mom.} \mu$ and $\mathrm{esd}(\boldsymbol{Y}) \xrightarrow{\mathbb{E}\ mom.} \nu$. Then, $\mathrm{esd}(\boldsymbol{X} + \boldsymbol{Y}) \xrightarrow{\mathbb{E}\ mom.} \mu \boxplus \nu$.*

This may seem a bit abstract, but you can check that it is really possible to compute directly any order moment of $\mu \boxplus \nu$ that you want, though the formulas become long. Here are the first few:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\lambda\sim\mu\boxplus\nu} \lambda^0 &= p_0(\boldsymbol{m}, \boldsymbol{n}) = 1, \\
\mathop{\mathbb{E}}_{\lambda\sim\mu\boxplus\nu} \lambda^1 &= p_1(\boldsymbol{m}, \boldsymbol{n}) = m_1 + n_1, \\
\mathop{\mathbb{E}}_{\lambda\sim\mu\boxplus\nu} \lambda^2 &= p_2(\boldsymbol{m}, \boldsymbol{n}) = m_2 + 2m_1n_1 + n_2, \\
\mathop{\mathbb{E}}_{\lambda\sim\mu\boxplus\nu} \lambda^3 &= p_3(\boldsymbol{m}, \boldsymbol{n}) = m_3 + 3m_2n_1 + 3m_1n_2 + n_3, \\
\mathop{\mathbb{E}}_{\lambda\sim\mu\boxplus\nu} \lambda^4 &= p_4(\boldsymbol{m}, \boldsymbol{n}) = m_4 + 4m_3n_1 + 4m_2n_2 + 4m_1n_3 + n_4 + 2m_2n_1^2 + 2m_1^2n_2 - 2m_1^2n_1^2.
\end{aligned}$$

It is helpful and explains the name of free convolution to contrast it with ordinary convolution, which should be familiar from scalar probability. We use the notation "$\oplus$" instead of the more conventional "$*$" to emphasize the analogy.

Theorem 3.4.5. *Suppose that $X = X^{(d)}, Y = Y^{(d)}$ are sequences of independent scalar random variables, where $\mathsf{Law}(X) \xrightarrow{\mathbb{E}\ mom.} \mu$ and $\mathsf{Law}(Y) \xrightarrow{\mathbb{E}\ mom.} \nu$. Then, there is a probability measure $\mu \oplus \nu$ such that $\mathsf{Law}(X+Y) \xrightarrow{\mathbb{E}\ mom.} \mu \oplus \nu$, where the moments of $\mu \oplus \nu$ are (different) polynomials of the moments of μ and ν:*

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

These polynomials are just given by the binomial theorem. Listing the first few, we see that $\mu \boxplus \nu$ and $\mu \oplus \nu$ agree up to the third moment, but then start to disagree:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\lambda\sim\mu\oplus\nu} \lambda^0 &= q_0(\boldsymbol{m}, \boldsymbol{n}) = 1, \\
\mathop{\mathbb{E}}_{\lambda\sim\mu\oplus\nu} \lambda^1 &= q_1(\boldsymbol{m}, \boldsymbol{n}) = m_1 + n_1, \\
\mathop{\mathbb{E}}_{\lambda\sim\mu\oplus\nu} \lambda^2 &= q_2(\boldsymbol{m}, \boldsymbol{n}) = m_2 + 2m_1n_1 + n_2, \\
\mathop{\mathbb{E}}_{\lambda\sim\mu\oplus\nu} \lambda^3 &= q_3(\boldsymbol{m}, \boldsymbol{n}) = m_3 + 3m_2n_1 + 3m_1n_2 + n_3, \\
\mathop{\mathbb{E}}_{\lambda\sim\mu\oplus\nu} \lambda^4 &= q_4(\boldsymbol{m}, \boldsymbol{n}) = m_4 + 4m_3n_1 + 6m_2n_2 + 4m_1n_3 + n_4.
\end{aligned}$$

Again, this is an indication that you should think of asymptotic freeness as a kind of independence for the spectra of random matrices. In the same way that for random scalars X and Y we can calculate $\mathsf{Law}(X+Y)$ from $\mathsf{Law}(X)$ and $\mathsf{Law}(Y)$ when X and Y are independent, for sequences of random matrices $\boldsymbol{X}^{(d)}$ and $\boldsymbol{Y}^{(d)}$ we can calculate the limiting law of $\mathsf{esd}(\boldsymbol{X}^{(d)} + \boldsymbol{Y}^{(d)})$ from the individual limits of $\mathsf{esd}(\boldsymbol{X}^{(d)})$ and $\mathsf{esd}(\boldsymbol{Y}^{(d)})$ when $\boldsymbol{X}^{(d)}$ and $\boldsymbol{Y}^{(d)}$ are asymptotically free.

3.4.2 Additive Free Limit Theorems

Definition 3.4.6. *For μ a measure, we write $c \cdot \mu := \mathsf{Law}(cX)$ where $X \sim \mu$.*

With the above notation, we may again revisit our renormalization formulation of the CLT. Indeed, with this notation the renormalization map may be written in terms of operations on measures only as

$$\mathcal{S}(\mu) = \frac{1}{\sqrt{2}} \cdot (\mu \oplus \mu),$$

and the CLT in its original form is just that, for μ a probability measure with mean zero and variance one,

$$\frac{1}{\sqrt{k}} \cdot (\underbrace{\mu \oplus \cdots \oplus \mu}_{k \text{ times}}) \xrightarrow[k\to\infty]{} \mathcal{N}(0,1), \tag{3.4.2}$$

which also implies the renormalization form

$$\mathcal{S}^{(k)}(\mu) \xrightarrow[k\to\infty]{} \mathcal{N}(0,1),$$

and the fixed point condition reading

$$\mathcal{S}(\mathcal{N}(0,1)) = \frac{1}{\sqrt{2}} \cdot (\mathcal{N}(0,1) \oplus \mathcal{N}(0,1)) = \mathcal{N}(0,1).$$

To understand the semicircle limit theorem in parallel to this, the natural choice is to consider a different renormalization map,

$$\mathcal{S}_{\mathsf{free}}(\mu) := \frac{1}{\sqrt{2}} \cdot (\mu \boxplus \mu).$$

This is a bit of a subtle operation: $\mathcal{S}_{\mathsf{free}}$ should be seen as operating on the *limiting distributions* of random matrices. If $\boldsymbol{X} = \boldsymbol{X}^{(d)}$ has $\mathsf{esd}(\boldsymbol{X}) \to \mu$ and $\boldsymbol{Y}$ is an independent copy of $\boldsymbol{X}$, then $\mathcal{S}_{\mathsf{free}}(\mu)$ is the limit of $\mathsf{esd}(\frac{1}{\sqrt{2}}(\boldsymbol{X} + \boldsymbol{Q}\boldsymbol{Y}\boldsymbol{Q}^\top))$ for $\boldsymbol{Q} \sim \mathsf{Haar}(\mathcal{O}(d))$. Since if $\boldsymbol{X} \sim \frac{1}{\sqrt{d}} \cdot \mathsf{GOE}(d)$ then $\mathsf{Law}(\boldsymbol{X} + \boldsymbol{Q}\boldsymbol{Y}\boldsymbol{Q}^\top) = \frac{1}{\sqrt{d}} \cdot \mathsf{GOE}(d)$ again, and $\mathsf{esd}(\boldsymbol{X}) \to \mu_{\mathsf{SC}}$, we find that

$$\mathcal{S}_{\mathsf{free}}(\mu_{\mathsf{SC}}) = \frac{1}{\sqrt{2}} \cdot (\mu_{\mathsf{SC}} \boxplus \mu_{\mathsf{SC}}) = \mu_{\mathsf{SC}},$$

so that indeed μ_{SC} is a fixed point of this "free renormalization map."

Remark 3.4.7. *In fact, note that it is possible just as well to take $\boldsymbol{Y} = \boldsymbol{X}$ above: $\boldsymbol{X}$ and $\boldsymbol{Q}\boldsymbol{X}\boldsymbol{Q}^\top$ are asymptotically free by Theorem 3.3.2, even though they are* not *independent as random vectors.*

Moreover, via our results above, though we have not worked out the combinatorics needed to actually write out this recursion in closed form, the fixed point condition $\mathcal{S}_{\mathsf{free}}(\mu) = \mu$ determines the moments of μ recursively. In particular, you can show that the following conditions uniquely describe μ_{SC}:

$$\mathop{\mathbb{E}}_{X \sim \mu} X = 0,$$
$$\mathop{\mathbb{E}}_{X \sim \mu} X^2 = 1,$$
$$\mu \text{ compactly supported},$$
$$\mathcal{S}_{\mathsf{free}}(\mu) = \mu.$$

Finally, it is possible to show a free version of the CLT in the stronger form (3.4.2) above:

Theorem 3.4.8 (Free central limit theorem)**.** *For any compactly supported μ of mean zero and variance 1,*

$$\frac{1}{\sqrt{k}} \cdot (\underbrace{\mu \boxplus \cdots \boxplus \mu}_{k \text{ times}}) \xrightarrow[k \to \infty]{} \mu_{\mathsf{SC}}.$$

Remark 3.4.9 (Free Poisson limit theorem)**.** *You might be familiar with the following limit theorem from classical probability, concerning the limiting distribution of a count of rare events: as $k \to \infty$, the sum of k i.i.d. random variables distributed as $\mathsf{Ber}(\lambda/k)$ for a fixed $\lambda > 0$ converges weakly to $\mathsf{Pois}(\lambda)$. In our above convolution notation, this says*

$$\underbrace{\mathsf{Ber}\left(\frac{\lambda}{k}\right) \oplus \cdots \oplus \mathsf{Ber}\left(\frac{\lambda}{k}\right)}_{k \text{ times}} \xrightarrow[k \to \infty]{} \mathsf{Pois}(\lambda).$$

We might ask, what happens in the case of additive free convolution rather than ordinary convolution? It turns out that the answer is one of the distributions we have seen already, the Marchenko-Pastur law:

$$\underbrace{\mathsf{Ber}\left(\frac{\lambda}{k}\right) \boxplus \cdots \boxplus \mathsf{Ber}\left(\frac{\lambda}{k}\right)}_{k \text{ times}} \xrightarrow[k\to\infty]{} \mu_{\mathsf{MP}(\lambda)}.$$

Indeed, $\mathsf{Ber}(\lambda/k)$ *is the limiting e.s.d. of any projection matrix whose rank is a fraction* λ/k *of the ambient dimension. Thus the left-hand side above is a sum of many free low-rank projections, which is indeed reminiscent of (though not identical to) the original matrices whose limiting e.s.d. gave the Marchenko-Pastur distribution, a sum of i.i.d. rank one matrices* $\boldsymbol{g}_i \boldsymbol{g}_i^\top$ *for* $\boldsymbol{g}_i$ *standard Gaussian.*

3.4.3 Multiplicative Free Convolution

There is a parallel theory for studying *products* of random matrices as well. A bit of care is required in the setup, since in general the product of symmetric matrices is not symmetric. Still, we have:

Theorem 3.4.10. *Suppose that* $\boldsymbol{X}, \boldsymbol{Y}$ *are an asymptotically free sequence of random matrices, as in Definition 3.3.1, such that moreover* $\boldsymbol{X}, \boldsymbol{Y} \succeq \boldsymbol{0}$ *almost surely. Then, there is a polynomial* r_k *(whose coefficients are universal constants not depending on* $\boldsymbol{X}$, $\boldsymbol{Y}$*, or their moments* m_a *and* n_b*) such that the value of any limiting moment of* $\boldsymbol{X}^{1/2}\boldsymbol{Y}\boldsymbol{X}^{1/2}$ *is*

$$\lim_{d\to\infty} \frac{1}{d} \mathbb{E}\,\mathsf{Tr}(\boldsymbol{X}^{1/2}\boldsymbol{Y}\boldsymbol{X}^{1/2})^k = r_k(m_1, \ldots, m_k, n_1, \ldots, n_k). \tag{3.4.3}$$

Further, if m_k *are the moments of a compactly supported measure* μ *and* n_k *are the moments of a compactly supported measure* ν*, then* $p_k(m_1, \ldots, m_k, n_1, \ldots, n_k)$ *are the moments of another compactly supported measure, which is also uniquely determined by its moments.*

Definition 3.4.11. *We write* $\mu \boxtimes \nu$ *for the measure discussed above, called the* multiplicative free convolution *of* μ *and* ν*.*

The argument is actually very simple with the tools we have developed already: we just observe that these moments, after an application of the cyclic property of trace, are a special case of the tangled joint moments that asymptotic freeness lets us derive the limiting values of via Proposition 3.3.7:

$$\begin{aligned}\mathsf{Tr}(\boldsymbol{X}^{1/2}\boldsymbol{Y}\boldsymbol{X}^{1/2})^k &= \mathsf{Tr}(\boldsymbol{X}^{1/2}\boldsymbol{Y}(\boldsymbol{X}\boldsymbol{Y})^{k-1}\boldsymbol{X}^{1/2}) \\ &= \mathsf{Tr}(\boldsymbol{X}\boldsymbol{Y})^k.\end{aligned}$$

Let us give a few examples of how this can be useful to justify the perhaps unusual-looking form of the matrices it handles.

Example 3.4.12. *Consider the case of* $\boldsymbol{X}$ *a projection matrix. Then,* $\boldsymbol{X}^{1/2} = \boldsymbol{X}$*, so the above reduces to describing the spectrum of* $\boldsymbol{X}\boldsymbol{Y}\boldsymbol{X}$*. Moreover, if* $\boldsymbol{X}$ *is a* coordinate *projection to, say, a choice of* αd *random coordinates for some* $\alpha \in (0,1)$*, then its limiting empirical*

spectral distribution is $\mathsf{Ber}(\alpha)$*. So, for any rotationally invariant* $\boldsymbol{Y}$ *with limiting empirical spectral distribution* ν*, we have* $\mathsf{esd}(\boldsymbol{XYX}) \to \mathsf{Ber}(\alpha) \boxtimes \nu$*. But, up to padding by zeros, the matrix* $\boldsymbol{XYX}$ *is just a random* $\alpha d \times \alpha d$ *submatrix of* $\boldsymbol{Y}$*. Thus multiplicative free convolution allows us to describe the spectra of* random submatrices *of many matrices (including, for instance, random subgraphs of given graphs in some cases).*

Example 3.4.13. *If, say,* $\boldsymbol{G} \sim \mathcal{N}(0, \frac{1}{m})^{\otimes d \times m}$ *with* $d/m \to c$*, then the Marchenko-Pastur limit theorem gives that* $\mathsf{esd}(\boldsymbol{GG}^\top) \to \mu_{\mathsf{MP}(c)}$*. Write* $\sqrt{\mu} := \mathsf{Law}(\sqrt{X})$ *for* $X \sim \mu$*, when* $X \geq 0$ *almost surely, we also find that* $\mathsf{ed}(\boldsymbol{\sigma}(\boldsymbol{G})) \to \sqrt{\mu_{\mathsf{MP}(c)}}$ *for the singular values. With multiplicative free convolution, we can treat products of rectangular matrices in a similar spirit. Suppose that* $\boldsymbol{G} \sim \mathcal{N}(0, \frac{1}{m})^{\otimes d \times m}$ *and* $\boldsymbol{H} \sim \mathcal{N}(0, \frac{1}{n})^{\otimes f \times m}$*, where* $\frac{d}{m} \to b$ *and* $\frac{f}{m} \to c$*. Suppose that* $0 < b < c < 1$ *(other cases may be handled similarly), so that* $\boldsymbol{GH}^\top$*, a* $d \times f$ *matrix, generically has* d *positive singular values since* $d < f$*. Then, we have*

$$\begin{aligned}\mathsf{ed}(\boldsymbol{\sigma}(\boldsymbol{GH}^\top)) &= \sqrt{\mathsf{esd}(\boldsymbol{GH}^\top \boldsymbol{HG}^\top)} \\ &= \sqrt{\mathsf{esd}(\boldsymbol{G}(\boldsymbol{H}^\top \boldsymbol{H})^{1/2}(\boldsymbol{H}^\top \boldsymbol{H})^{1/2}\boldsymbol{G}^\top)}.\end{aligned}$$

This is a $d \times d$ *matrix. We will convert it below to an* $m \times m$ *matrix, which will introduce* $m - d = (1-b)m$ *zero eigenvalues. Thus:*

$$\begin{aligned}(1-b)\delta_0 + b\,\mathsf{ed}(\boldsymbol{\sigma}(\boldsymbol{GH}^\top)) &\approx \sqrt{\mathsf{esd}((\boldsymbol{H}^\top \boldsymbol{H})^{1/2}\boldsymbol{G}^\top \boldsymbol{G}(\boldsymbol{H}^\top \boldsymbol{H})^{1/2})} \\ &\to \sqrt{((1-b)\delta_0 + b\mu_{\mathsf{MP}(b)}) \boxtimes ((1-f)\delta_0 + f\mu_{\mathsf{MP}(f)})},\end{aligned}$$

and a limit for $\mathsf{ed}(\boldsymbol{\sigma}(\boldsymbol{GH}^\top))$ *may be calculated by first taking this multiplicative free convolution and then removing the atom at zero that will result.*

3.4.4 R- AND S-TRANSFORMS

We have thus far only given abstract means of computing additive and multiplicative free convolutions via moments using the implicit recursion in Proposition 3.3.7. In fact, there are powerful tools for performing explicit computations that rely on generating functions constructed from the Stieltjes transform from Section 2.7.

For additive free convolution, we have the following tool.

Definition 3.4.14. *The* R-transform *of a probability measure* μ *is* $R_\mu(z) = G_\mu^{-1}(z) - \frac{1}{z}$*.*

Here, G_μ^{-1} is the functional inverse of the Stieltjes transform, viewed as an inverse of formal power series: using the moment generating function expression from (2.7.1), we find a power series such that $G_\mu^{-1}(G_\mu(z)) = z$ when the composition of power series is computed term by term. The useful property of the R-transform is as follows:

Theorem 3.4.15. $R_{\mu \boxplus \nu}(z) = R_\mu(z) + R_\nu(z)$*.*

Moreover, the above power series expansion of $R_\mu(z)$ leads to a series of the form

$$R_\mu(z) = \sum_{k \geq 1} \kappa_k^{\mathsf{free}}(\mu) z^{k-1}.$$

Here the $\kappa_k^{\text{free}}(\mu)$ are each polynomials of the moments of μ called the *free cumulants*, which parallel the classical cumulants. If $m_k := \mathbb{E}_{X\sim\mu}X^k$, then the first few free cumulants are

$$\begin{aligned}
\kappa_1^{\text{free}} &= m_1,\\
\kappa_2^{\text{free}} &= m_2 - m_1^2,\\
\kappa_3^{\text{free}} &= m_3 - 3m_2m_1 + 2m_1^3,\\
\kappa_4^{\text{free}} &= m_4 - 4m_3m_1 - 2m_2^2 + 10m_2m_1^2 - 5m_1^4.
\end{aligned}$$

In contrast, the classical cumulants are

$$\begin{aligned}
\kappa_1 &= m_1,\\
\kappa_2 &= m_2 - m_1^2,\\
\kappa_3 &= m_3 - 3m_2m_1 + 2m_1^3,\\
\kappa_4 &= m_4 - 4m_3m_1 - 3m_2^2 + 12m_2m_1^2 - 6m_1^4.
\end{aligned}$$

As in the case of moment formulas for additive free convolution, the first three free and classical free cumulants are the same, while for the fourth and beyond they are different.

Example 3.4.16. *One may compute $R_{\mu_{\mathsf{SC}}}(z) = z$. This means that the second free cumulant of μ_{SC} is 1, and all others are zero. This characterizes the semicircle law, and is parallel to how the second classical cumulant of the Gaussian measure is 1 and all others are zero.*

The R transform gives a fairly practical (though often algebraically involved) recipe for calculating the free convolution, as illustrated below:

$$\begin{array}{ccccccc}
\mu & \to & G_\mu & \to & R_\mu & \searrow & \\
 & & & & & & R_\mu + R_\nu = R_{\mu\boxplus\nu} \to G_{\mu\boxplus\nu} \to \mu\boxplus\nu.\\
\nu & \to & G_\nu & \to & R_\nu & \nearrow &
\end{array}$$

There is also a similar transform and corresponding recipe for multiplicative free convolution.

Definition 3.4.17. *The S-*transform *of a probability measure μ is $S_\mu(z) = \frac{1}{z}R_\mu^{-1}(z)$.*

Theorem 3.4.18. *$S_{\mu\boxtimes\nu}(z) = S_\mu(z)S_\nu(z)$.*

Again, using this result and our tools for Stieltjes transforms, you can compute multiplicative free convolutions as follows:

$$\begin{array}{ccccccccc}
\mu & \to & G_\mu & \to & R_\mu & \to & S_\mu & \searrow & \\
 & & & & & & & & S_\mu S_\nu = S_{\mu\boxtimes\nu} \to R_{\mu\boxtimes\nu} \to G_{\mu\boxtimes\nu} \to \mu\boxtimes\nu.\\
\nu & \to & G_\nu & \to & R_\nu & \to & S_\nu & \nearrow &
\end{array}$$

3.5 Application: Spectra of Expanders [LM23]

Let us work through an example that shows both how additive free convolution and the free central limit theorem can be informative about concrete combinatorial objects and hints at the underlying algebraic structure of asymptotic freeness.

3.5.1 Cycles

We begin with a mystery. Consider on the one hand $\boldsymbol{D} = \boldsymbol{D}^{(d)} \in \mathbb{R}^{d\times}_{\mathsf{sym}}$ a random diagonal matrix with $D_{ii} \sim \mathsf{Unif}(\{\pm 1\})$ i.i.d., and write $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top$ and $\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^\top$ for $\boldsymbol{Q}, \boldsymbol{R} \sim \mathsf{Haar}(\mathcal{O}(d))$ (omitting the (d) superscripts from here on out). We may view the individual laws of these matrices as being differences of projections: $\boldsymbol{X} = \boldsymbol{P}_1 - \boldsymbol{P}_2$ for a pair of projections $\boldsymbol{P}_1, \boldsymbol{P}_2$ having $\boldsymbol{P}_1 + \boldsymbol{P}_2 = \boldsymbol{I}$ and $\mathsf{rank}(\boldsymbol{P}_i) \approx d/2$ (though with some fluctuations, as really $\mathsf{rank}(\boldsymbol{P}_i)$ is a sum of d i.i.d. $\mathsf{Ber}(1/2)$ random variables). In particular, $\boldsymbol{X} = \boldsymbol{P}_1 - (\boldsymbol{I} - \boldsymbol{P}_1) = 2\boldsymbol{P}_1 - \boldsymbol{I}$, so it may be interpreted as a *reflection* across the (random) row space of $\boldsymbol{P}_1$ (and thus is an orthogonal matrix, as we can check by noting that $\boldsymbol{X}^2 = \boldsymbol{I}$).

$\boldsymbol{X}$ and $\boldsymbol{Y}$ are asymptotically free by Theorem 3.3.2. They have the respective convergences

$$\mathsf{esd}(\boldsymbol{X}) \xrightarrow{\mathbb{E}\text{ mom.}} \mathsf{Unif}(\{\pm 1\}),$$
$$\mathsf{esd}(\boldsymbol{Y}) \xrightarrow{\mathbb{E}\text{ mom.}} \mathsf{Unif}(\{\pm 1\}).$$

Thus we also have

$$\mathsf{esd}(\boldsymbol{X} + \boldsymbol{Y}) \xrightarrow{\mathbb{E}\text{ mom.}} \mathsf{Unif}(\{\pm 1\})^{\boxplus 2}.$$

Using the R-transform, the right-hand side may be computed to be the *arcsine distribution*, having density

$$d\mu_{\mathsf{AS}}(x) = \mathbb{1}\{x \in (-2,2)\}\frac{1}{\pi\sqrt{4-x^2}}dx.$$

Remark 3.5.1. *By the remarks above, this also shows that a shifted and rescaled arcsine distribution is the limiting empirical spectral distribution of a sum of two independent projections to two uniformly random subspaces of dimension $d/2$, i.e.,* $\mathsf{Ber}(1/2)^{\boxplus 2}$.

Remark 3.5.2. *In either of the cases above of* $\mathsf{Unif}(\{\pm 1\})^{\boxplus 2}$ *or* $\mathsf{Ber}(1/2)^{\boxplus 2}$*, we see the striking phenomenon that the additive free convolution of two discrete measures can be, and in fact almost always is, a continuously supported measure. This is in constrast to classical convolution, where $\mu \oplus \mu$ for any μ supported on two atoms $\{a, b\}$ can of course be supported only on at most three atoms $\{2a, 2b, a+b\}$.*

The arcsine distribution also appears as a limiting e.s.d. in the following, apparently completely different, situation. Consider the cycle graph C_d on d vertices. Its adjacency matrix is

$$\boldsymbol{A}_{C_d} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

This matrix is *circulant*, and therefore is diagonalized by the discrete Fourier transform. Performing this calculation, you may show that the eigenvalues are, up to reordering, of the

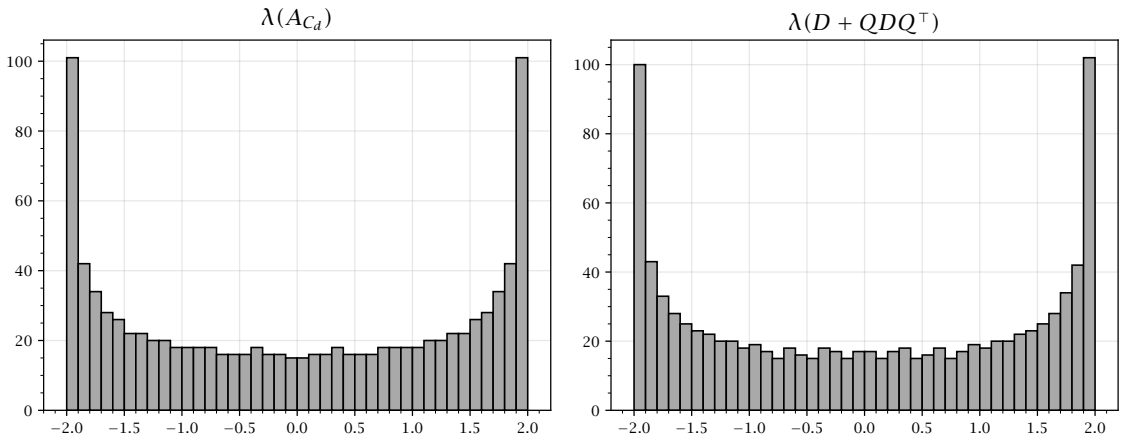


Figure 3.1: Histograms of the empirical spectral distributions of the adjacency matrix of C_{1000} and of $\boldsymbol{D} + \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top \in \mathbb{R}^{1000\times 1000}$ for $D_{ii} \sim \mathsf{Unif}(\{\pm 1\})$ and $\boldsymbol{Q} \sim \mathsf{Haar}(1000)$.

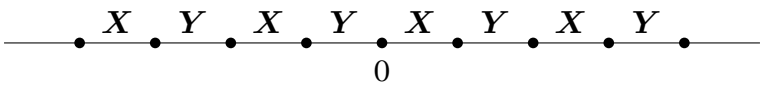


Figure 3.2: An illustration of the labelling of the graph on $\mathbb{Z}$ used in the argument of Section 3.5.1.

form $2\cos(2\pi j/n)$ for $j = 0, \ldots, d-1$. A bit more computation shows that these (deterministic!) collections of numbers again have the arcsine distribution as their empirical limiting distribution:

$$\mathsf{esd}(\boldsymbol{A}_{C_d}) \to \mu_{\mathsf{AS}} = \mathsf{Unif}(\{\pm 1\})^{\boxplus 2}. \tag{3.5.1}$$

See Figure 3.1 for an illustration.

Let us find an explanation of this unusual coincidence: we will see that, in fact, the combinatorics of moments of the cycle graph have a natural interpretation in terms of freeness. We have the interpretation

$$\begin{aligned} m_k &:= \lim_{d\to\infty} \frac{1}{d} \mathsf{Tr}(\boldsymbol{A}_{C_d}^k) \\ &= \lim_{d\to\infty} \frac{1}{d} \#\{\text{closed walks of length } k \text{ in } C_d\}, \end{aligned}$$

where a *closed walk* is any sequence $i_1, i_2, \ldots, i_{k-1}, i_k = i_1$ such that i_a and i_{a+1} are adjacent in C_d (or whatever graph is under discussion). Write $\mathbb{Z}$ for the infinite graph on the vertex set of the integers, where each pair $\{i, i+1\}$ are connected. Note that, in C_d, once $d > k/2$, the number of closed walks starting from any given i_1 is same as the number of closed walks in $\mathbb{Z}$ starting from 0. That is, the neighborhood of radius $k/2$ around any vertex of C_d looks just like the same neighborhood around 0 in $\mathbb{Z}$. Thus,

$$m_k = \#\{\text{closed walks of length } k \text{ in } \mathbb{Z} \text{ starting from } 0\} = \left\{ \begin{array}{ll} 0 & \text{if } k \text{ is odd,} \\ \binom{k}{k/2} & \text{if } k \text{ is even} \end{array} \right\}.$$

As an aside, comparing these moments to those of μ_{AS} is one way to prove (3.5.1) above.

Now, recall our two matrices $\boldsymbol{X}, \boldsymbol{Y}$ above. They satisfy $\boldsymbol{X}^2 = \boldsymbol{Y}^2 = \boldsymbol{I}_d$, and, by asymptotic freeness,

$$\lim_{d\to\infty} \frac{1}{d} \mathbb{E}\,\mathsf{Tr}(\boldsymbol{X}\boldsymbol{Y}\cdots\boldsymbol{X}\boldsymbol{Y}) = 0,$$

for any number of interleaved repetitions of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Consider labelling each edge of $\mathbb{Z}$ with $\boldsymbol{X}$ or $\boldsymbol{Y}$, so that the two labels alternate (see Figure 3.2). View a walk on $\mathbb{Z}$ as a sequence of edge traversals, which corresponds to a sequence such as $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{X}, \boldsymbol{Y}, \dots)$. Convince yourself of the following: a walk is closed if and only if this sequence, when written as a *product* $\boldsymbol{X}\boldsymbol{Y}\boldsymbol{X}\boldsymbol{X}\boldsymbol{Y}\cdots$, is *cancellable*, meaning that it can be rewritten down to $\boldsymbol{I}_d$ by only using the identities $\boldsymbol{X}^2 = \boldsymbol{Y}^2 = \boldsymbol{I}_d$. For example, $\boldsymbol{X}\boldsymbol{Y}\boldsymbol{Y}\boldsymbol{X}\boldsymbol{Y}\boldsymbol{Y}\boldsymbol{X}\boldsymbol{X} = \boldsymbol{X}\boldsymbol{Y}^2\boldsymbol{X}\boldsymbol{Y}^2\boldsymbol{X}^2 = \boldsymbol{X}^2 = \boldsymbol{I}_d$ corresponds to a closed walk, while $\boldsymbol{Y}\boldsymbol{Y}\boldsymbol{X}\boldsymbol{Y}\boldsymbol{X}\boldsymbol{X} = \boldsymbol{Y}^2\boldsymbol{X}\boldsymbol{Y}\boldsymbol{X}^2 = \boldsymbol{X}\boldsymbol{Y}$ does not. Moreover, any non-cancellable sequence when viewed as a product, can be rewritten down to a product of the form $\boldsymbol{X}\boldsymbol{Y}\boldsymbol{X}\boldsymbol{Y}\cdots\boldsymbol{X}\boldsymbol{Y}$.

Thus by the above observations, we have

$$\lim_{d\to\infty} \frac{1}{d} \mathbb{E}\,\mathsf{Tr}(\text{ word in } \boldsymbol{X}, \boldsymbol{Y}\ ) = \mathbb{1}\{\text{word is cancellable}\}.$$

Therefore,

$$\begin{aligned} m_k &= \#\{\text{closed walks of length } k \text{ in } \mathbb{Z} \text{ starting from } 0\} \\ &= \#\{\text{cancellable words of length } k\} \\ &= \sum_{\text{words of length } k} \lim_{d\to\infty} \frac{1}{d} \mathbb{E}\,\mathsf{Tr}(\text{ word in } \boldsymbol{X}, \boldsymbol{Y}\ ) \\ &= \lim_{d\to\infty} \frac{1}{d} \mathbb{E}\,\mathsf{Tr}(\boldsymbol{X}+\boldsymbol{Y})^k. \end{aligned}$$

Thus, even if we did not know that the limit was μ_{AS}, this argument would show us that

$$\mathsf{esd}(\boldsymbol{A}_{C_d}) \to \mathsf{Unif}(\{\pm 1\})^{\boxplus 2}.$$

3.5.2 Generalization to Higher Degree

The above argument, while nice, is perhaps not very impressive, because after all it is easy to also do direct calculations with μ_{AS}. Let us now show how, by pushing it just a little further, we can quickly obtain a large amount of rather more difficult information.

Definition 3.5.3 (Girth)**.** *The* girth *of a graph, written* $\mathsf{girth}(G)$*, is the length of the shortest cycle in* G *(or* $+\infty$ *if there are no cycles).*

Consider a sequence of graphs $G = G^{(d)}$, each on d vertices, such that G is p-regular, i.e., such that every vertex has degree p. (There are no p-regular graphs on d vertices unless pd is even, so we should really take only even d when p is odd, but we will not introduce special notation to handle this caveat.) Suppose also that $\mathsf{girth}(G^{(d)}) \to \infty$ as $d \to \infty$.

For $p = 2$, such a sequence of graphs is given by the cycles C_d, since of course $\mathsf{girth}(C_d) = d$. For $p \geq 3$, it is not at all obvious that such graphs exist. Such graphs would be a very

strong form of so-called *expander* graphs. The reason for this terminology is that, for G p-regular, every neighborhood around every vertex of G of radius smaller than $\mathsf{girth}(G)$ is a p-regular *tree*. In particular, the sizes of these neighborhoods grow very fast with the radius r, roughly as p^r. Such graphs therefore have many favorable properties such as fast mixing of random walks, making them useful in computational applications; see, e.g., [HLW06] for an introduction to this area. You should thus also think of graphs of high girth as *locally tree-like*.

Theorem 3.5.4 (Erdős, Sachs [ES63])**.** *There exists a sequence $G^{(d)}$ as above with* $\mathsf{girth}(G^{(d)}) \geq c_p \log_{p-1}(d)$ *for a constant $c_p > 0$.*

The above result is based on a non-constructive argument with the probabilistic method, but explicit constructions for various classes of special p have been found as well, like the famous Lubotzky-Phillips-Sarnak graphs of [LPS88]. See also [LS21] for some discussion of the history of this problem.

Remark 3.5.5. *Actually, even if we choose $G^{(d)}$ uniformly at random from the set of all p-regular graphs on d vertices, it will typically have a relatively small number of short cycles, and the results we prove below also apply to these* random regular graphs*. This is again a field unto itself, but see [Wor99] for an introduction.*

Here we will ask: what does the empirical spectral distribution of a p-regular graph of high girth look like? (This question is intimately related to quantifying the mixing time and related expansion properties of such graphs as mentioned above.) One can answer this question with combinatorics calculations of moments, of course. But, following our argument for cycles and using our general understanding of free probability, watch how quickly we can reason about this problem.

Write $\mathbb{T}_p$ for the infinite p-regular tree, and choose some vertex v_0 in this tree to serve as a root (like the role of 0 in $\mathbb{Z}$ before). By the same argument as for C_d and $\mathbb{Z}$, by the diverging girth of $G^{(d)}$, we have

$$\begin{aligned} m_k &:= \lim_{d\to\infty} \frac{1}{d} \mathsf{Tr}(\boldsymbol{A}_G^k) \\ &= \#\{\text{closed walks of length } k \text{ in } \mathbb{T}_p \text{ starting from } v_0\}. \end{aligned}$$

We may again label the edges of $\mathbb{T}_p$ with matrices $\boldsymbol{X}_1, \dots, \boldsymbol{X}_p$, such that each vertex is adjacent to exactly one edge having each label. Suppose that $\boldsymbol{X}_i = \boldsymbol{Q}_i \boldsymbol{D} \boldsymbol{Q}_i^\top$ for $\boldsymbol{D}$ random diagonal as before, and $\boldsymbol{Q}_i \sim \mathsf{Haar}(\mathcal{O}(d))$ i.i.d.; in particular, these $\boldsymbol{X}_i$ are asymptotically free.[1] The same exact argument as before then gives

$$\begin{aligned} &= \#\{\text{cancellable words of length } k \text{ in } \boldsymbol{X}_1, \dots, \boldsymbol{X}_p\} \\ &= \sum_{\text{words of length } k \text{ in } \boldsymbol{X}_1,\dots,\boldsymbol{X}_p} \lim_{d\to\infty} \frac{1}{d} \mathbb{E}\, \mathsf{Tr}(\text{ word in } \boldsymbol{X}_1, \dots \boldsymbol{X}_p\,) \\ &= \lim_{d\to\infty} \frac{1}{d} \mathbb{E}\, \mathsf{Tr}(\boldsymbol{X}_1 + \cdots + \boldsymbol{X}_p)^k. \end{aligned}$$

Thus we find a natural generalization of the arcsine limit theorem for the e.s.d. of large cycles, namely:

$$\mathsf{esd}(G^{(d)}) \to \mathsf{Unif}(\{\pm 1\})^{\boxplus p}.$$

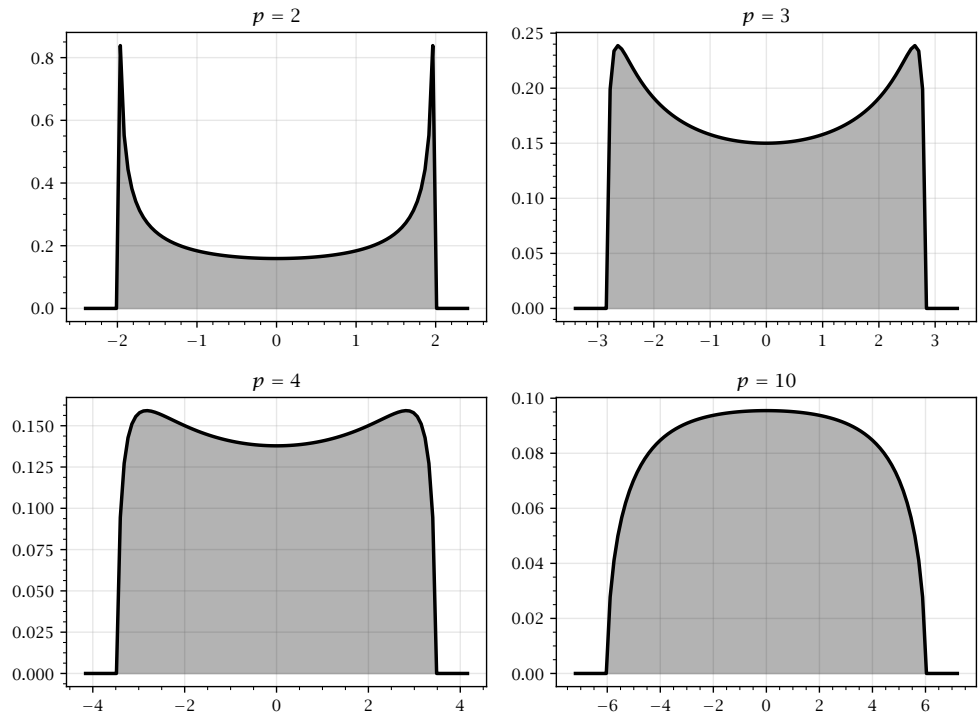


Figure 3.3: Examples of densities of the Kesten-McKay laws $\mu_{\mathsf{KM}(p)}$ for different values of the parameter p.

By moment calculations or using the R-transform, you may compute this free convolution to be the *Kesten-McKay law*, $\mathsf{Unif}(\{\pm 1\})^{\boxplus p} = \mu_{\mathsf{KM}(p)}$, having density

$$d\mu_{\mathsf{KM}(p)}(x) = \mathbb{1}\left\{x \in \left[-2\sqrt{p-1}, 2\sqrt{p-1}\right]\right\} \cdot \frac{p\sqrt{4(p-1)-x^2}}{2\pi(p^2-x^2)}dx.$$

If you are familiar with *Ramanujan graphs*, the interval of support of the eigenvalues above, $[-2\sqrt{p-1}, 2\sqrt{p-1}]$, is the same as appears in their definition. Note also that $\mu_{\mathsf{KM}(2)} = \mu_{\mathsf{AS}}$, matching our previous observation.

Finally, the free central limit theorem immediately implies what these limiting distributions look like in the further limit $p \to \infty$ (note that here we are discussing a sequence of two limits: first $d \to \infty$, and then the behavior of a limiting object arising there as $p \to \infty$). The above expression shows that $\frac{1}{\sqrt{p}} \cdot \mu_{\mathsf{KM}(p)}$ is supported on $[-2 + o(1), 2 + o(1)]$, and the free CLT implies more specifically

$$\frac{1}{\sqrt{p}} \cdot \mu_{\mathsf{KM}(p)} = \frac{1}{\sqrt{p}} \cdot \mathsf{Unif}(\{\pm 1\})^{\boxplus p} \xrightarrow[p\to\infty]{} \mu_{\mathsf{SC}}.$$

See Figure 3.3 for examples of the Kesten-McKay densities illustrating this convergence even for $p \approx 10$.

3.6 Application: Neural Network Loss Landscapes [PB17]

We outline an argument making certain predictions about the structure of the landscape of the optimization arising in training a neural network.

3.6.1 Setup and Definitions

At a high level, like our earlier example of kernel regression, this is a matter of learning a general non-linear mapping $F : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$. We want to approximate this map by one of the form

$$\hat{\boldsymbol{y}}(\boldsymbol{x}; \boldsymbol{V}, \boldsymbol{W}) = \boldsymbol{W}(\boldsymbol{V}\boldsymbol{x})_+$$

for $\boldsymbol{V} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{inter}}}$ and $\boldsymbol{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{inter}}}$. where $(x)_+ = \max\{0, x\}$ and this operation applies entrywise to vectors (this is the so-called *rectified linear unit* or *ReLU* non-linearity commonly used in deep learning). We also abbreviate $\boldsymbol{\theta} = (\boldsymbol{V}, \boldsymbol{W})$, and view this as a vector $\boldsymbol{\theta} \in \mathbb{R}^{d_{\text{in}} d_{\text{inter}} + d_{\text{inter}} d_{\text{out}}}$, the vector of all parameters available for tuning in $\hat{\boldsymbol{y}}$. We then also write $\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}(\boldsymbol{x}; \boldsymbol{\theta})$.

We are also given a *training set* of examples $(\boldsymbol{x}_1, \boldsymbol{y}_1), \dots, (\boldsymbol{x}_m, \boldsymbol{y}_m)$ for $\boldsymbol{x}_k \in \mathbb{R}^{d_\in}$ and $\boldsymbol{y}_k \in \mathbb{R}^{d_{\text{out}}}$, and want to learn $\boldsymbol{\theta}$ such that $\hat{\boldsymbol{y}}(\boldsymbol{x}_k; \boldsymbol{\theta}) \approx \boldsymbol{y}_k$ for each $k \in [m]$.

We make the following assumptions about the dimensions $d_\in, d_{\text{inter}}, d_{\text{out}}, m$ involved. The first is for simplicity:

$$d := d_\in = d_{\text{inter}} = d_{\text{out}},$$

saying that our network is "square" in shape (in particular, $\boldsymbol{V}$ and $\boldsymbol{W}$ are square). Note then that $\boldsymbol{\theta} \in \mathbb{R}^{2d^2}$. The second expresses that we are in the "big data" regime of having a comparable amount of data to the dimensionality of the data:

$$\frac{d}{m} \to c \in (0, \infty),$$

as in the setting of the Marchenko-Pastur limit theorem.

In training our network, we seek to minimize a *training loss*,

$$L(\boldsymbol{\theta}) := \frac{1}{2} \sum_{k=1}^{m} \|\hat{\boldsymbol{y}}(\boldsymbol{x}_k; \boldsymbol{\theta}) - \boldsymbol{y}_k\|^2 = \frac{1}{2} \sum_{k=1}^{m} \sum_{i=1}^{d} \left(\hat{y}_{ki}(\boldsymbol{\theta}) - y_{ki}\right)^2,$$

where we introduce new notation $\hat{y}_{ki}(\boldsymbol{\theta}) = (\hat{\boldsymbol{y}}(\boldsymbol{x}_k; \boldsymbol{\theta}))_i$, suppressing for now the dependence on $\boldsymbol{x}_k$. We then run some procedure like gradient descent, stochastic gradient descent, adaptive gradient descent (Adam, AdaGrad, etc.) to attempt to minimize $L(\boldsymbol{\theta})$.

In general, in doing so we should expect to approach a *critical point* of L, one where $\nabla L(\boldsymbol{\theta}) = \boldsymbol{0}$. Near a critical point, the landscape of L is determined by the Hessian matrix $\nabla^2 L(\boldsymbol{\theta})$; in particular, by multivariate Taylor expansion, near a critical point $\boldsymbol{\theta}_0$, $L(\boldsymbol{\theta}) \approx \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \nabla^2 L(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$. This local landscape can have the structure of a *saddle point*: along certain one-dimensional directions it can look like an upward-opening parabola, and along others like a downward-opening parabola. The eigenvalues of $\nabla^2 L(\boldsymbol{\theta}_0)$ describe how many "principal" directions of each kind there are (for a total of the ambient dimension $2d^2$. We call the fraction of negative eigenvalues the *(normalized) index* of a critical point.

We then want to answer the following kind of question: as training proceeds, how does the index of the critical points we encounter change? This (somewhat indirectly) says something about the convergence of training, since saddle points of higher index have a larger number of favorable directions and thus are "easier to escape" and continue descending the loss landscape, at least for noisy gradient methods like stochastic gradient descent.

3.6.2 Free Probability Heuristics for Hessian Spectrum

To understand these questions, we need to calculate the gradient and Hessian, a simple exercise in the chain and product rules. We write α, β below for general indices of $\boldsymbol{\theta}$, corresponding to an index in either $\boldsymbol{V}$ or $\boldsymbol{W}$.

$$\begin{aligned}
(\nabla L)(\boldsymbol{\theta})_\alpha &= \partial_\alpha L(\boldsymbol{\theta}) \\
&= \sum_{k=1}^{m}\sum_{i=1}^{d}(\hat{y}_{ki}(\boldsymbol{\theta}) - y_{ki})\partial_\alpha \hat{y}_{ki}(\boldsymbol{\theta}), \\
(\nabla^2 L)(\boldsymbol{\theta})_{\alpha\beta} &= \partial_\alpha \partial_\beta L(\boldsymbol{\theta}) \\
&= \underbrace{\sum_{k=1}^{m}\sum_{i=1}^{d}\partial_\alpha \hat{y}_{ki}(\boldsymbol{\theta})\partial_\beta \hat{y}_{ki}(\boldsymbol{\theta})}_{①_{\alpha\beta}} + \underbrace{\sum_{k=1}^{m}\sum_{k=1}^{d} r_{ki}(\boldsymbol{\theta})\partial_\alpha\partial_\beta \hat{y}_{ki}(\boldsymbol{\theta})}_{②_{\alpha\beta}},
\end{aligned}$$

where we define the *residuals*

$$r_{ki}(\boldsymbol{\theta}) := \hat{y}_{ki}(\boldsymbol{\theta}) - y_{ki},$$

the errors in approximating coordinate i of training data point k. We would like to make a prediction about $\mathrm{esd}(\nabla^2 L(\boldsymbol{\theta}))$ in terms of the value of $L(\boldsymbol{\theta})$. To do this, we start with the following assumption:

Heuristic 0: ① and ② are freely independent.

We will not be precise with terms like asymptotic freeness here but will just proceed at an intuitive level. If we believe Heuristic 1, it should be enough to understand the spectra of ① and ② individually.

For ①, note that if we define the *Jacobian* of the function $\hat{\boldsymbol{y}} : \mathbb{R}^{2d^2} \to \mathbb{R}^{dm}$ as $\boldsymbol{J}(\boldsymbol{\theta}) \in \mathbb{R}^{2d^2 \times dm}$ with entries

$$J_{\alpha,(k,i)}(\boldsymbol{\theta}) = \partial_\alpha \widehat{by}_{ki}(\boldsymbol{\theta}),$$

then we have

$$① = \boldsymbol{J}(\boldsymbol{\theta})\boldsymbol{J}(\boldsymbol{\theta})^\top.$$

We make two remarks. First, $① \succeq \boldsymbol{0}$, so this term is "fighting against" the loss landscape having more descent directions or critical points having high index. Second, ① does not depend on the training outputs $\boldsymbol{y}_k$, and thus is just a function of the neural network architecture as well as the relative over- or under-parametrization, the relationship of d to m. We have seen that the Marchenko-Pastur law is the "canonical" limiting distribution for such matrices, which suggests that we might guess:

Heuristic 1: (1) obeys the Marchenko-Pastur limit theorem: for some suitable scaling $S^{(J)} = S^{(J)}(d,m)$,

$$\mathsf{esd}\left(\frac{1}{S^{(J)}} \cdot (1)\right) \approx \mu_{\mathsf{MP}(2c)},$$

where we calculate $2d^2/dm = 2d/m \to c$ by our assumption.

For (2), we calculate in some more detail. Note that $\hat{y}_{ki}(\boldsymbol{\theta}) = \hat{y}_{ki}(\boldsymbol{V}, \boldsymbol{W})$, for any $k \in [m]$ and $i \in [d]$, is piecewise bilinear in $\boldsymbol{V}$ and $\boldsymbol{W}$. Thus, second derivatives with respect to two entries both in $\boldsymbol{V}$ or both in $\boldsymbol{W}$ will always be zero, so (2) will have the block matrix structure

$$(2) = \begin{bmatrix} \mathbf{0} & \boldsymbol{R} \\ \boldsymbol{R}^\top & \mathbf{0} \end{bmatrix}$$

for an (asymmetric) $\boldsymbol{B} \in \mathbb{R}^{d^2 \times d^2}$, where the entries of $\boldsymbol{B}$ are

$$B_{(p,q),(s,t)} = \frac{\partial^2 L}{\partial V_{pq} \partial W_{st}} = \sum_{k=1}^{m} \sum_{i=1}^{d} r_{ki}(\boldsymbol{V}, \boldsymbol{W}) \frac{\partial^2 \hat{y}_{ki}}{\partial V_{pq} \partial W_{st}}(\boldsymbol{V}, \boldsymbol{W}).$$

The eigenvalues of (2) are then $\pm\sigma_i(\boldsymbol{B})$ for $i = 1, \ldots, d$, so it suffices to compute the latter.

Expanding the expression for $\hat{y}_{ki}$, we have

$$\hat{y}_{ki} = (\boldsymbol{W}(\boldsymbol{V}\boldsymbol{x}_k)_+)_i = \sum_{t=1}^{d} W_{it} \left(\sum_{q=1}^{d} V_{tq} x_{kq} \right)_+ .$$

Thus many of the derivatives involved above are zero; we may compute

$$\frac{\partial^2 \hat{y}_{ki}}{\partial V_{pq} \partial W_{st}}(\boldsymbol{V}, \boldsymbol{W}) = \mathbb{1}\{i = s\} \cdot \mathbb{1}\{p = t\} \cdot \mathbb{1}\{(\boldsymbol{V}\boldsymbol{x}_k)_p \geq 0\} \cdot x_{kq}.$$

The sum for an entry of $\boldsymbol{B}$ then reduces to

$$B_{(p,q),(s,t)} = \mathbb{1}\{p = t\} \cdot \sum_{k=1}^{m} r_{ks}(\boldsymbol{V}, \boldsymbol{W}) \cdot \mathbb{1}\{(\boldsymbol{V}\boldsymbol{x}_k)_p \geq 0\} \cdot x_{kq}.$$

In particular, there is a $\widetilde{\boldsymbol{B}}$ that differs from $\boldsymbol{B}$ only by a permutation of the columns (and therefore has the same singular values) that has a block-diagonal structure because of the $\mathbb{1}\{p = t\}$ term above:

$$\widetilde{\boldsymbol{B}} = \begin{bmatrix} \widetilde{\boldsymbol{B}}^{(1)} & & \\ & \ddots & \\ & & \widetilde{\boldsymbol{B}}^{(d)} \end{bmatrix}.$$

Here we have $\widetilde{\boldsymbol{B}}^{(p)} \in \mathbb{R}^{d \times d}$, having entries

$$\widetilde{B}^{(p)}_{qs} = \sum_{k=1}^{m} r_{ks}(\boldsymbol{V}, \boldsymbol{W}) \cdot \mathbb{1}\{(\boldsymbol{V}\boldsymbol{x}_k)_p \geq 0\} \cdot x_{kq}.$$

Another way to write this is

$$\widetilde{\boldsymbol{B}}^{(p)} = \boldsymbol{R}\boldsymbol{D}^{(p)}\boldsymbol{X},$$

where $\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{V}, \boldsymbol{W}) \in \mathbb{R}^{d\times m}$ has entries $R_{sk} = r_{ks}$, $\boldsymbol{D}^{(p)} \in \mathbb{R}^{m\times m}$ is a diagonal matrix with $D^{(p)}_{kk} = \mathbb{1}\{(\boldsymbol{V}\boldsymbol{x}_k)_p \geq 0\}$, and $\boldsymbol{X} \in \mathbb{R}^{m\times d}$ has entries $X_{kq} = x_{kq}$ (i.e., the matrix of the training data). The singular values of $\boldsymbol{B}$ are the disjoint union of the singular values of the $\widetilde{\boldsymbol{B}}^{(p)}$ for $p = 1, \ldots, d$. We now make the following (dramatic!) heuristic assumptions about this setting:

Heuristic 2: $\boldsymbol{X}$ and $\boldsymbol{R}$ have i.i.d. Gaussian entries and $\boldsymbol{D}^{(p)}$ has independent $\mathsf{Unif}(\{0,1\})$ diagonal entries. Specifically, $R_{ks} \sim \mathcal{N}(0, \epsilon S^{(R)})$ i.i.d. for a parameter ϵ to be fixed later and some scaling $S^{(R)} = S^{(R)}(d, m)$. Note first that this assumption fixes the scale of the loss of $\boldsymbol{\theta}$, which will depend on ϵ:

$$L(\boldsymbol{\theta}) = \sum_{k=1}^{m}\sum_{i=1}^{d} R_{ki}^2 \approx \epsilon \cdot S^{(R)} md =: \epsilon S^{(L)}$$

for some further scaling $S^{(L)} = S^{(L)}(d, m)$. Also, we then find that the singular values of $\boldsymbol{B}$, which are the same as those of $\widetilde{\boldsymbol{B}}$, which are the disjoint unions of those of $\widetilde{\boldsymbol{B}}^{(p)}$, have, upon rescaling, the same empirical distribution as those of $\sqrt{\epsilon}\boldsymbol{G}\boldsymbol{H}^\top$ for $\boldsymbol{G}, \boldsymbol{H} \sim \mathcal{N}(0,1)^{d\otimes m/2}$, where we reduce m to $m/2$ since $\boldsymbol{D}^{(p)}$ zeroes out about half of the initial m dimensions.

Thus, concretely we learn the following. First, for a probability measure ν, write $\mathsf{sym}(\nu) := \mathsf{Law}(sX)$ where $s \sim \mathsf{Unif}(\{\pm 1\})$ and $X \sim \nu$ independently. By the case of multiplicative free convolution discussed in Example 3.4.13, we have

$$\mathsf{ed}\left(\boldsymbol{\sigma}\left(\frac{2}{m}\boldsymbol{G}\boldsymbol{H}^\top\right)\right) \to \sqrt{\mu_{\mathsf{MP}(2c)}^{\boxtimes 2}}$$

when $d, m \to \infty$ with $\frac{d}{m} \to c$ (recall that $\sqrt{\mu}$ is the law of $\sqrt{X}$ when $X \sim \mu$). Note that we normalize by $2/m$ above because that is the dimension of $\boldsymbol{G}$ and $\boldsymbol{H}$. Then, we have

$$\mathsf{esd}\left(\frac{1}{m\sqrt{S^{(R)}}} \cdot \textcircled{2}\right) \to \frac{\sqrt{\epsilon}}{2} \cdot \mathsf{sym}\left(\sqrt{\mu_{\mathsf{MP}(2c)}^{\boxtimes 2}}\right).$$

Now, suppose that $S := m\sqrt{S^{(R)}} = S^{(J)}$ so that $\textcircled{1}$ and $\textcircled{2}$ are on the same scale. Then, combining our heuristic arguments, we find that we expect the following relationship between $L(\boldsymbol{\theta})$ and the e.s.d. of the Hessian at $\boldsymbol{\theta}$:

$$\mathsf{esd}\left(\frac{1}{S}\cdot\nabla^2 L(\boldsymbol{\theta})\right) \approx \mu_{\mathsf{MP}(2c)} \boxplus \left(\frac{\sqrt{\epsilon}}{2}\cdot\mathsf{sym}\left(\sqrt{\mu_{\mathsf{MP}(2c)}^{\boxtimes 2}}\right)\right) =: \rho_{c,\epsilon} \quad \text{when} \quad L(\boldsymbol{\theta}) \approx \epsilon \cdot mdS^{(R)}.$$

This is a rather involved combination of additive and multiplicative free convolution, but in principle we have the tools to compute it using R- and S-transforms.

3.6.3 Implications

In particular, from $\rho_{c,\epsilon}$ we can calculate the normalized index for a given c (overparametrization) and ϵ (loss) as

$$\mathsf{ind}(c, \epsilon) := \rho_{c,\epsilon}((-\infty, 0]) \in [0, 1].$$

In [PB17], this is computed and analyzed asymptotically (the details are not included) and the following two qualitative phenomena are extracted:

1. For $c \in (0, \frac{1}{2})$, there is an $\epsilon_* = \epsilon_*(c) > 0$ such that, if $\epsilon < \epsilon_*$, then $\mathsf{ind}(c, \epsilon) = 0$. That is, the theory predicts that, at a "macroscopically" large scaling of the loss value, all critical points achieving that loss will be local minima, and thus one expects noisy gradient descent to become stuck at roughly this level of loss.

   On the other hand, $\epsilon_* \to 0$ as $c \to \frac{1}{2}$. Note that, when $c = \frac{1}{2}$, then $m = 2d$ so $2d^2 = md$ and the total number of dimensions of the training data ($= md$) equals the total number of parameters ($= 2d^2$). Thus we expect to be able to interpolate the training data exactly once $c > \frac{1}{2}$, and the theory suggests that in fact gradient descent will be able to reach such an interpolation.

2. Moreover, in the underparametrized regime $c \in (0, \frac{1}{2})$, for ϵ slightly greater than ϵ_*, we have the scaling

$$\mathsf{ind}(c, \epsilon) \sim f(c) \left(\frac{\epsilon - \epsilon_*}{\epsilon_*} \right)^{3/2}$$

   for some function $f(c)$. We may interpret this as saying that, towards the end of training, when we are approaching the best loss level that gradient descent will reach ($= \epsilon_*$), then the normalized index of the nearby critical points will decrease superlinearly with the loss.

These findings are compatible with previous models for neural network loss landscapes and even can be corroborated to some extent (most convincingly for small c) in numerical experiments, including the specific exponent $3/2$. See [PB17] for full details.

Remark 3.6.1. *The paper also offers the simplifying heuristic of replacing* $\mathsf{sym}(\sqrt{\mu_{\mathsf{MP}(2c)}^{\boxtimes 2}})$ *with* μ_{SC}*, independent of* c*, which is a coarser model but one that might be sensible when we do not assume the special structure coming from the ReLU non-linearity.*

3.7 Application: Covariance Estimation [EK08]

As we have proposed before, one of the main statistical uses of free probability is to help us understand the estimation of covariance matrices when we have inadequate data for the sample covariance to be a consistent estimator. Consider the following setting: we have a sequence $\mathbf{\Sigma} = \mathbf{\Sigma}^{(d)} \in \mathbb{R}_{\mathsf{sym}}^{d \times d}$ positive semidefinite matrices, with $\mathsf{esd}(\mathbf{\Sigma}^{(d)}) \to \mu$. The $\mathbf{\Sigma}^{(d)}$ are covariances of an increasing number of features we are interested in of a population, and μ describes the limiting correlation structure of this growing collection of features.

Suppose, though, that we only have access to a relatively small number of samples of this population, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{(d)})$ independently, with $\frac{d}{m} \to c$ as in our previous discussions. We want to estimate the limiting spectrum μ from these observations. A natural choice of estimator of $\mathbf{\Sigma}$ is the sample covariance

$$\hat{\mathbf{\Sigma}} := \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^\top,$$

and thus a natural choice of estimator of μ is

$$\hat{\mu} := \mathsf{esd}(\hat{\mathbf{\Sigma}}).$$

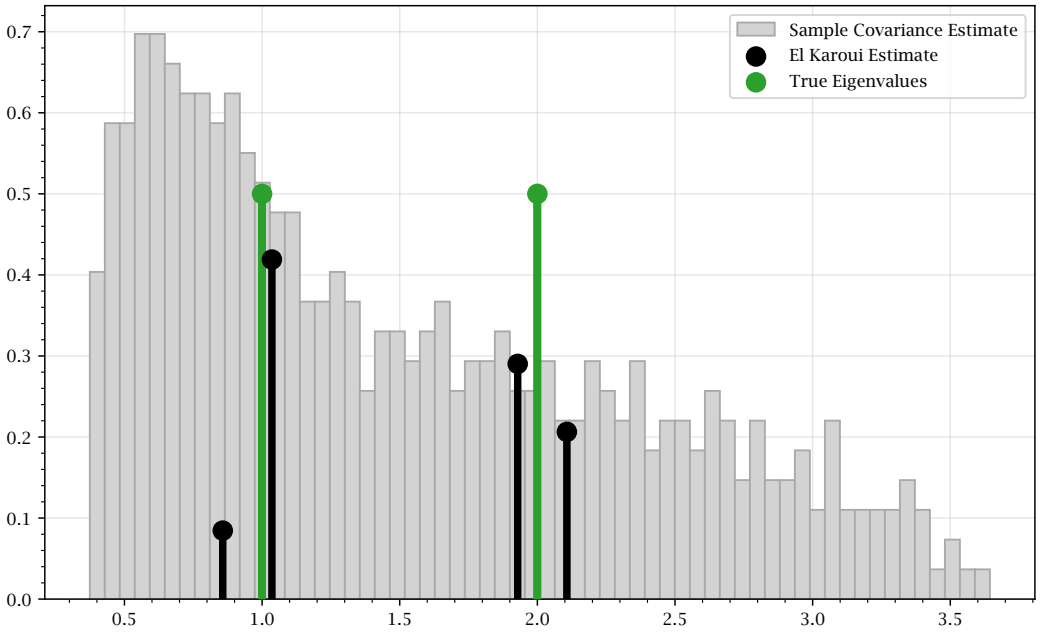


Figure 3.4: An example of the performance of the algorithm presented in Section 3.7. We take $d = 500, m = 2500$, and Σ to be the diagonal matrix half of whose entries are 1 and half of whose entries are 2. Thus its true empirical spectral distribution is $\mu = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$, shown in green. In light gray is the empirical spectral distribution of the sample covariance $\hat{\Sigma}$, nominally an estimate of μ but performing very poorly in this case and having output close to $(\frac{1}{2}\delta_1 + \frac{1}{2}\delta_2) \boxtimes \mu_{\mathrm{MP}(1/5)}$. In black is the estimate obtained from the algorithm, with a very conservative grid of 15 values of t_a and 50 values of z_b with imaginary part equal to 1. The construction and solution of the linear program requires less than a millisecond in the `MOSEK` solver accessed through `cvxpy`.

However, our tools show us that this estimator will be inconsistent, in the sense that it will not converge to μ as $d \to \infty$, and in fact we can characterize this inconsistency precisely. Introducing $g_i \sim \mathcal{N}(0, I_d)$, we have $\mathsf{Law}(x_i) = \mathsf{Law}(\Sigma^{1/2} g_i)$, and thus

$$\hat{\Sigma} \overset{\text{(law)}}{=} \Sigma^{1/2} \left(\frac{1}{m} \sum_{i=1}^{m} g_i g_i^\top \right) \Sigma^{1/2}.$$

Therefore, we have

$$\hat{\mu} \to \mu \boxtimes \mu_{\mathrm{MP}(c)},$$

a kind of smoothing of μ by multiplicative free convolution.

Two extremes are instructive to consider. First, note that as $c \to 0$ we have $\mu_{\mathrm{MP}(c)} \to \delta_1$, so indeed in this limit—which corresponds to $m \gg d$ where we have much more data than features we want to estimate—we do recover an accurate estimate of μ. Second, if $\Sigma^{(d)} = I_d$ so that $\mu = \delta_1$, then this limit is just $\mu_{\mathrm{MP}(c)}$, and the above recovers the Marchenko-Pastur limit theorem.

Let us see now how to use our knowledge of free probability to devise a *denoising* procedure for this estimate. We will use the following characterization of multiplicative free

convolution with the Marchenko-Pastur measure, which can be obtained from calculations with the S-transform.

Proposition 3.7.1. *For any compactly supported probability measure μ,*

$$G_{\mu \boxtimes \mu_{\mathsf{MP}(c)}}(z) = \int \frac{1}{z - x(1 - c + czG_{\mu \boxtimes \mu_{\mathsf{MP}(c)}}(z))} d\mu(x).$$

You may view this as an implicit equation for the Stieltjes transform of the multiplicative free convolution similar to, but more complicated than, the one we derived for the Stieltjes transform of the semicircle measure in Section 2.7. As a sanity check, note again that as $c \to 0$ the right-hand side tends to $G_\mu(z)$. If $\mu = \delta_1$, then this gives a self-consistency equation or implicit formula for the Marchenko-Pastur law $\mu_{\mathsf{MP}(c)}$, again in the spirit of the one we saw for the semicircle law.

Let us write $\boldsymbol{\lambda} := \boldsymbol{\lambda}(\hat{\boldsymbol{\Sigma}})$, the eigenvalues of the sample covariance. The idea of the algorithm is what we believe that the process generating these eigenvalues makes it so that

$$\mathsf{ed}(\boldsymbol{\lambda}) \approx \mu \boxtimes \mu_{\mathsf{MP}(d/m)},$$

where we substitute the empirical quantity d/m, to which our data gives us access, for the theoretical asymptotic quantity c. If this is true, then by Proposition 3.7.1, we should have

$$G_{\mathsf{ed}(\boldsymbol{\lambda})}(z) \approx \int \frac{1}{z - x(1 - \frac{d}{m} + \frac{d}{m} z G_{\mathsf{ed}(\boldsymbol{\lambda})}(z))} d\mu(x)$$

for all z.

Let us fix a grid of $t_1, \ldots, t_M \in \mathbb{R}$, and propose a discrete estimate for μ of the form

$$\hat{\mu} := \sum_{a=1}^{M} w_a \delta_{t_a},$$

for some $w_a \geq 0$, $\sum w_a = 1$. Further, let us take a grid of $z_1, \ldots, z_N \in \mathbb{C}$ (in practice it works well to choose these spaced along a horizontal line slightly above the real axis). Write

$$\hat{G}_b := G_{\mathsf{ed}(\boldsymbol{\lambda})}(z_b) = \frac{1}{d} \sum_{i=1}^{d} \frac{1}{z_b - \lambda_i},$$

noting that we may compute these from the z_b together with the λ_i, all quantities available to us. The Stieltjes transform equation, evaluated at these z_b, then becomes

$$\hat{G}_b \approx \sum_{a=1}^{M} \frac{1}{z_b - x_a(1 - \frac{d}{m} + \frac{d}{m} z_b \hat{G}_b)} w_a.$$

This is just asking that the w_a approximately satisfy a linear system: separating the real and imaginary parts, for suitable $\boldsymbol{A} \in \mathbb{R}^{2N \times M}$ and $\boldsymbol{b} \in \mathbb{R}^{2N}$, we just want

$$\boldsymbol{A}\boldsymbol{w} \approx \boldsymbol{b}.$$

Though both $\boldsymbol{A}$ and $\boldsymbol{b}$ depend on the $\widehat{G}_b$, this is not a problem, since those are just numbers now that we can compute from our data.

We may then solve any of a variety of convex optimization problems to do this, say of the form

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_p \\ \text{subject to} & \boldsymbol{w} \geq \boldsymbol{0}, \\ & \boldsymbol{1}^\top \boldsymbol{w} = 1. \end{array}$$

In particular taking $p = \infty$ is appealing, which leads to a linear program. As illustrated in Figure 3.4, the results are strikingly good in situations where the estimator $\mathsf{esd}(\widehat{\boldsymbol{\Sigma}})$ for μ is completely ineffective. See also [EK08] for some theoretical results guaranteeing strong approximations as $M, N \to \infty$ (though note that the algorithm presented there is a small variation on what is presented here).

3.8 Exercises

Exercise 3.8.1. *Define the* 2×2 *matrix*

$$\boldsymbol{A} := \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Let $t \sim \mathsf{Unif}([0, \pi])$ *and define the random rotation matrix*

$$\boldsymbol{U} := \begin{bmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{bmatrix}.$$

Finally, define $\boldsymbol{X}^{(2d)} := \boldsymbol{I}_d \otimes \boldsymbol{A} \in \mathbb{R}^{2d \times 2d}$ *and* $\boldsymbol{Y}^{(2d)} := \boldsymbol{I}_d \otimes (\boldsymbol{U}\boldsymbol{A}\boldsymbol{U}^\top) \in \mathbb{R}^{2d \times 2d}$ *random matrices.*

1. *Show that the sequences* $\boldsymbol{X}^{(2d)}$ *and* $\boldsymbol{Y}^{(2d)}$ *have converging empirical spectral moments (i.e., that* $\lim_{d \to \infty} \frac{1}{2d} \mathbb{E}\,\mathsf{Tr}\,\boldsymbol{X}^{(2d)^k}$ *exists for all* k *and likewise for* $\boldsymbol{Y}^{(2d)}$*) and that the pair of sequences is asymptotically free. (View the definition of asymptotic freeness as restricted to a sequence of matrices in only even dimensions.)*

   (**Hint:** *Boil this down to a statement about the* 2×2 *matrices* $\boldsymbol{A}$ *and* $\boldsymbol{U}$*.)*

2. *To what measure must the empirical spectral distribution of* $\boldsymbol{X}^{(2d)} + \boldsymbol{Y}^{(2d)}$ *then converge in expected moments? Why?*

   (**Hint:** *You do not need to calculate an additive free convolution by hand if you use a result we have seen.)*

3. *Show that the empirical spectral distribution of* $\boldsymbol{X}^{(2d)} + \boldsymbol{Y}^{(2d)}$ *almost surely consists of at most two atoms. Therefore, qualitatively, it will never resemble the measure you described in Part 2. For example, in a histogram of the eigenvalues, only at most two bins will ever be non-empty. Explain formally and precisely why this is not a contradiction to Part 2.*

Exercise 3.8.2. *We discussed p-regular graphs of large girth the chapter. In this problem, you will study p-regular graphs chosen uniformly at random numerically and observe that they share some but not all of the same properties.*

1. *Write code to generate a p-regular graph on d vertices (pd must be even) at random, as follows. View the d vertices as each having p "half-edges" attached to them, for a total of pd. A graph may be viewed as formed by gluing together half-edges in pairs to form full edges. As we have seen from the combinatorics of Gaussian moments, there are $(pd-1)!!$ possible perfect matchings among pd objects. Choose such a perfect matching uniformly at random (come up with and justify a way to perform this sampling). This forms a random p-regular* multigraph *G_0 on d vertices, since it is possible that you created self-loops or parallel edges in choosing your matching. Now, perform rejection sampling: repeat the procedure until you choose a matching that yields a simple graph G. Include this part of your code in your homework submission.*

   You do not need to prove it, but the resulting G is uniformly random among simple p-regular graphs on d vertices with labelled vertices.

2. *Write code to estimate $f(p,d) := \mathbb{P}[G_0$ is simple$]$ in the above procedure. For $p \in \{3,4\}$, estimate $f(p) := \lim_{d\to\infty} f(p,d)$ by taking d large. That is, for each p, for a sequence of growing d, report the fraction of trials giving G_0 simple out of a large total. Plot data to illustrate the convergence of your estimate as $d \to \infty$. (Optionally, if you are very patient, you may try $p = 5$. It helps to not generate an entire perfect matching before rejecting a G_0 that is not simple.)*

3. *Write $\boldsymbol{A}$ for the adjacency matrix of G formed above. For $p \in \{3,4\}$, confirm that, for large d, $\mathsf{esd}(\boldsymbol{A})$ is close to the Kesten-McKay measure with parameter 3 and 4 (respectively) as predicted in this chapter for regular graphs of large girth. How large of d is needed? Include convincing plots.*

4. *Let $T = T(G)$ be the number of triangles in G. Estimate $t(p,d) := \mathbb{E}T(G)$ for $p = 3$ and a sequence of growing d. Does our reasoning from the chapter apply to the random G? Try to identify what $\mathsf{Law}(T)$ converges to as $d \to \infty$. Include numerical evidence for your prediction of any kind you think is reasonable—histograms, experimental estimates of moments, etc.*

   *(***HINT:** *Compute $T(G)$ with matrix algebra, not for loops.)*

4 | Spiked Matrix Models

4.1 Motivation: Outliers in Real-World Spectra

In the previous section, we developed tools to understand how the "shape" or the "bulk" of the eigenvalues of a matrix is affected by noise. For example, we can understand the weak limit of $\boldsymbol{X} + \boldsymbol{W}$ where $\boldsymbol{W} \sim \mathsf{GOE}(d)$, or of $\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}\boldsymbol{G}^\top\boldsymbol{\Sigma}^{1/2}$ where $\boldsymbol{G} \sim \mathcal{N}(0,1)^{\otimes d\times m}$. But, in both cases, this weak limit only describes how the shape of the spectrum of $\boldsymbol{X}$ and $\boldsymbol{\Sigma}$ changes under such noising operations.

In reality, especially in statistical operations, we often are interested in understanding not this bulk behavior of all of the eigenvalues together, but in the behavior of *outlier* eigenvalues, which indicate, say in the case of a sample covariance matrix, "exceptional" correlations. These appear as single outliers in a histogram, or as "elbows" in the *scree plot* of λ_i as a function of i. An example with singular values on a real-world dataset is shown in Figure 3.4. The example shows nicely a useful intuition to have for real-world spectra of noisy matrices that include some statistical signal. Unlike what we have been able to handle with free probability, usually such consist of two distinct components: a bulk of mostly meaningless eigenvalues which arise due to noise, and a few informative outliers.

To understand how we should set our expectations in working with such data, we will take up the question of producing theoretical models that reproduce this phenomenon, and understanding how algorithms perform on those models.

You should see immediately that the methods from the previous chapter will not be useful. Free probability is (at the level of detail we have considered) a tool for establishing weak convergence and convergence in expected moments. And, as we saw in Section 2.5, that is in general not enough to get any information about individual extreme eigenvalues; intuitively, a single outlier contributes $\frac{1}{d}\delta_\lambda$ to the e.s.d. of a matrix, which cannot be "seen" by weak convergence. However, instead of the moment methods in Section 2.5, here we will pursue more sophisticated techniques for understanding these outliers.

4.2 Spiked Additive (Wigner) Model

The simplest model producing outlier eigenvalues is to simply add a rank-one perturbation to a matrix. Suppose $\boldsymbol{W} = \boldsymbol{W}^{(d)} \sim \mathsf{GOE}(d)$, $\boldsymbol{x} \in \mathbb{R}^d$ with $\|\boldsymbol{x}\| = 1$, and consider

$$\boldsymbol{Y} = \boldsymbol{Y}^{(d,\beta)} := \boldsymbol{W} + \beta\sqrt{d}\,\boldsymbol{x}\boldsymbol{x}^\top.$$

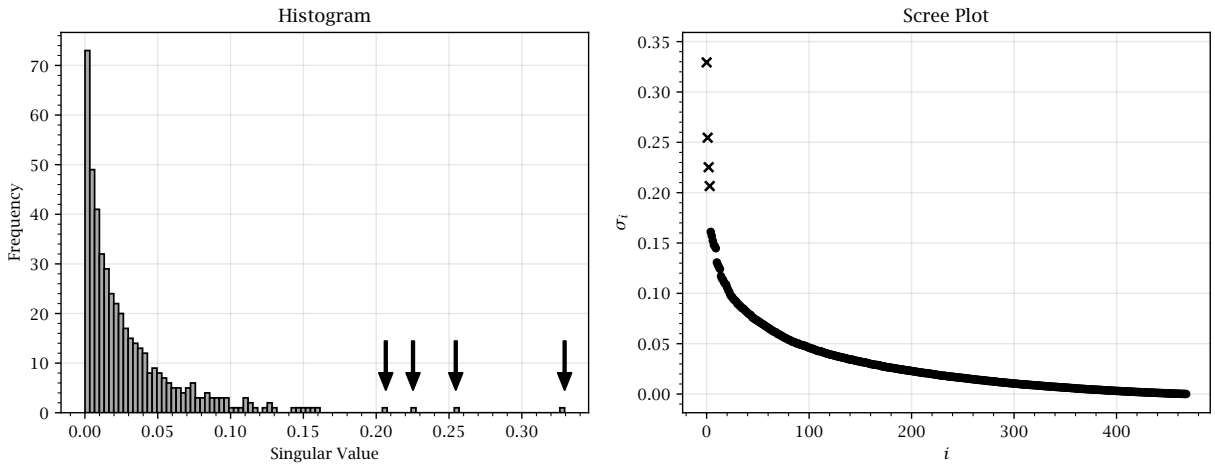


Figure 4.1: A histogram and scree plot of the singular values of a matrix of measurements of levels of interaction on Twitter between members of Congress (available here from the Stanford Large Network Dataset Collection). In either plot, four outlier values are clearly visible.

4.2.1 Basic Properties

Exercise 4.6.1 has you show that, as β grows, the top eigenvector of $\boldsymbol{Y}$ aligns more and more with $\boldsymbol{x}$. Further, provided $\boldsymbol{x}$ is deterministic or random independently of $\boldsymbol{W}$, then $\mathsf{Law}(\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{x}) = \mathcal{N}(0, 2\|\boldsymbol{x}\boldsymbol{x}^\top\|_F^2) = \mathcal{N}(0,2)$, with high probability, for any $\epsilon > 0$, we will have

$$\lambda_{\mathsf{max}}(\boldsymbol{Y}) \geq \boldsymbol{x}^\top \boldsymbol{Y}\boldsymbol{x} = \beta\sqrt{d} - \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{x} \geq (\beta - \epsilon)\sqrt{d}$$

and thus once $\beta > 2$ then $\boldsymbol{Y}$ with high probability will have its largest eigenvalue significantly larger than the typical value of $2\sqrt{d}$ for $\boldsymbol{Y} \sim \mathsf{GOE}(d)$.

Moreover, we may show that the other eigenvalues of $\boldsymbol{Y}$ are essentially the same as those of $\boldsymbol{W}$:

Proposition 4.2.1. *For all $2 \leq k \leq d$, $\lambda_k(\boldsymbol{W}) \leq \lambda_k(\boldsymbol{Y}) \leq \lambda_{k-1}(\boldsymbol{W})$.*

Proof. For the first inequality, just note that $\boldsymbol{Y} \succeq \boldsymbol{W}$ by construction (i.e., the perturbation we are adding is positive semidefinite). For the upper bound, by the Courant-Fischer min-max theorem, we have

$$\lambda_k(\boldsymbol{Y}) = \max_{\dim(V)=k} \min_{\substack{\boldsymbol{v}\in V \\ \|\boldsymbol{v}\|=1}} \boldsymbol{v}^\top \boldsymbol{Y}\boldsymbol{v}.$$

Any such V must contain a non-zero vector $\boldsymbol{v}$ orthogonal to $\boldsymbol{x}$ and the top $k-2$ eigenvectors of $\boldsymbol{W}$. For such $\boldsymbol{v}$, we have $\boldsymbol{v}^\top \boldsymbol{Y}\boldsymbol{v} = \boldsymbol{v}^\top \boldsymbol{W}\boldsymbol{v} \leq \lambda_{k-1}(\boldsymbol{W})$, and the result follows. □

Thus $\lambda_2(\boldsymbol{Y}), \ldots, \lambda_d(\boldsymbol{Y})$ are "sandwiched" among the eigenvalues of $\boldsymbol{W}$, and in particular with high probability lie in $[-2 - o(1), 2 + o(1)]$. A bit more thought also shows that these eigenvalues will moreover have the same empirical distribution of the semicircle. Thus it is *only* $\lambda_1(\boldsymbol{Y})$ that can be an outlier: $\mathsf{esd}(\boldsymbol{Y})$ will look either like the semicircle (as it does for $\boldsymbol{W}$, or equivalently when $\beta = 0$) or like the semicircle with a single outlier.

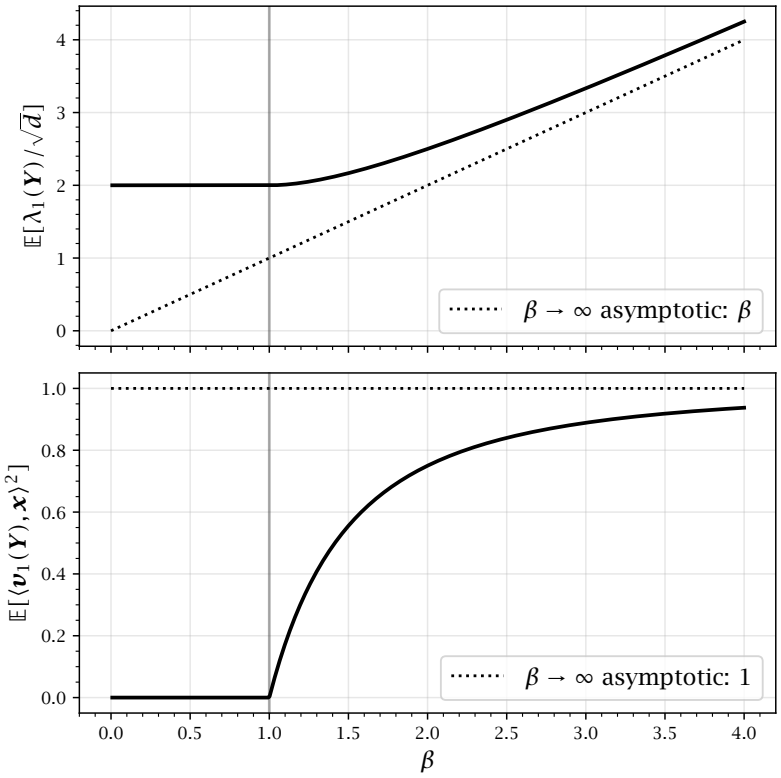


Figure 4.2: An illustration of the asymptotic results describing the phase transition in the spiked additive or Wigner model from Theorem 4.2.2.

So, our model indeed captures the behavior we wanted. Next comes the natural quantitative question: how large does β have to be for an outlier to appear? We have seen above that $\beta > 2$ suffices to create an outlier. Could it be that this is also *necessary*? Or perhaps any $\beta > 0$ suffices to "nudge" one of the top eigenvalues out of the bulk of the semicircle?

4.2.2 Phase Transition of Largest Eigenvalue

Surprisingly, neither guess is correct: there is a *critical* β required for an outlier to appear, but it is not 2 as the above argument suggests.

Theorem 4.2.2 ([FP07])**.** *Consider $\boldsymbol{Y} = \boldsymbol{Y}^{(d,\beta)}$ as above, for fixed $\beta > 0$ and $d \to \infty$ and any deterministic $\boldsymbol{x} = \boldsymbol{x}^{(d)}$ a unit vector. Write $\boldsymbol{v}_1(\boldsymbol{Y})$ for the eigenvector of $\lambda_1(\boldsymbol{Y})$. The following hold, with all convergences in probability.*

- *If $\beta < 1$, then:*

$$\frac{1}{\sqrt{d}}\lambda_1(\boldsymbol{Y}) \to 2,$$
$$\langle \boldsymbol{v}_1, \boldsymbol{x}\rangle^2 \to 0.$$

- *If $\beta > 1$, then:*

$$\frac{1}{\sqrt{d}}\lambda_1(\boldsymbol{Y}) \to \beta + \frac{1}{\beta} > 2,$$
$$\langle \boldsymbol{v}_1, \boldsymbol{x}\rangle^2 \to 1 - \frac{1}{\beta^2} > 0.$$

We may summarize the key features in the result as follows:

1. This model exhibits a *phase transition*: there is a critical value $\beta_* = 1$ such that the behaviors of the model and the impact of the perturbation when $\beta < \beta_*$ versus when $\beta > \beta_*$ look completely different.

2. Our previous idea about what value of β would suffice to create an outlier eigenvalue proposed that the mechanism for an outlier appearing was the $\boldsymbol{x}\boldsymbol{x}^\top$ perturbation completely "overpowering" $\boldsymbol{W}$ to forcibly make $\boldsymbol{x}^\top \boldsymbol{Y}\boldsymbol{x}$ large. What actually happens is subtler: when $\beta = 1 + \epsilon$, $\boldsymbol{x}$ "colludes" with the eigenvectors of $\boldsymbol{W}$ to create an outlier whose associated eigenvector is only very slightly correlated with $\boldsymbol{x}$.

3. Even when $\beta > 1$, $\lambda_1(\boldsymbol{Y}) \approx \beta + \frac{1}{\beta}$ is an inconsistent estimator for the parameter β of the model, one systematically biased to be larger than the truth.

4. Similarly, for any fixed β, $\boldsymbol{v}_1(\boldsymbol{Y})$ is an inconsistent estimator for the parameter $\boldsymbol{x}$, in the sense that $\|\boldsymbol{v}_1(\boldsymbol{Y}) - \boldsymbol{x}\| \not\to 0$. Indeed, if we view $\boldsymbol{x}$ as the "north pole" of $\mathbb{S}^{d-1}$, then $\boldsymbol{v}_1$ is concentrated around a "latitude" that moves closer to the pole as $\beta \to \infty$.

The inconsistency phenomena are similar to what we saw earlier in estimating $d \times d$ covariance matrices from $O(d)$ samples. There, the paucity of data or "information" in the model was captured by a small number of samples; here, it is captured by β not growing fast enough.

We will outline an argument using the resolvent and Stieltjes transform. We write

$$\widehat{\boldsymbol{W}} := \frac{1}{\sqrt{d}}\boldsymbol{W},$$
$$\widehat{\boldsymbol{Y}} := \frac{1}{\sqrt{d}}\boldsymbol{Y} = \widehat{\boldsymbol{W}} + \beta \boldsymbol{x}\boldsymbol{x}^\top.$$

The argument will depend on the following heuristic about the Stieltjes transform of $\widehat{\boldsymbol{W}}$, extending the ideas we encountered in Section 2.7 when sketching a proof of the semicircle limit theorem using the Stieltjes transform. There, we used concentration properties of the diagonal entries of $(z\boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1}$, whose average is the Stieltjes transform of $\mathsf{esd}(\widehat{\boldsymbol{W}})$. More is true, as we will discuss in more detail below: for any $z \in \mathbb{C} \setminus [-2, 2]$, as $d \to \infty$, this entire matrix acts as a multiple of the identity in small numbers of deterministic directions (in a similar fashion to the random projections we saw earlier). In particular, we expect, for fixed vectors $\boldsymbol{a}, \boldsymbol{b}$, that

$$\begin{aligned}\boldsymbol{a}^\top (z\boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1}\boldsymbol{b} &\approx \boldsymbol{a}^\top \left(\left(\frac{1}{d}\,\mathsf{Tr}(z\boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1}\right)\boldsymbol{I}_d\right)\boldsymbol{b} \\ &= \left(\frac{1}{d}\,\mathsf{Tr}(z\boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1}\right)\langle \boldsymbol{a}, \boldsymbol{b}\rangle \\ &\approx G_{\mu_{\mathsf{SC}}}(z)\langle \boldsymbol{a}, \boldsymbol{b}\rangle.\end{aligned}$$

We may at least verify this for the expectation: expanding the resolvent and then this quadratic form in the eigenspaces of $\widehat{\boldsymbol{W}} = \sum_{i=1}^{d} \lambda_i(\widehat{\boldsymbol{W}}) \boldsymbol{w}_i \boldsymbol{w}_i^\top$, we have

$$\mathbb{E}\boldsymbol{a}^\top (z\boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1}\boldsymbol{b} = \mathbb{E}\sum_{i=1}^{d} \frac{1}{z - \lambda_i(\widehat{\boldsymbol{W}})} \langle \boldsymbol{a}, \boldsymbol{w}_i\rangle\langle \boldsymbol{b}, \boldsymbol{w}_i\rangle$$

and now since the eigenvectors and eigenvalues of $\widehat{\boldsymbol{W}}$ are independent,

$$= \sum_{i=1}^{d} \mathbb{E}\frac{1}{z - \lambda_i(\widehat{\boldsymbol{W}})} \mathbb{E}\langle \boldsymbol{a}, \boldsymbol{w}_i\rangle\langle \boldsymbol{b}, \boldsymbol{w}_i\rangle$$

and the second expectation equals $\mathbb{E}\langle \boldsymbol{a}\boldsymbol{b}^\top, \boldsymbol{w}_i\boldsymbol{w}_i^\top\rangle = \langle \boldsymbol{a}\boldsymbol{b}^\top, \mathbb{E}\boldsymbol{w}_i\boldsymbol{w}_i^\top\rangle = \langle \boldsymbol{a}\boldsymbol{b}^\top, \frac{1}{d}\boldsymbol{I}_d\rangle = \frac{1}{d}\langle \boldsymbol{a}, \boldsymbol{b}\rangle$ since $\mathsf{Law}(\boldsymbol{w}_i) = \mathsf{Unif}(\mathbb{S}^{d-1})$, whereby

$$\begin{aligned} &= \langle \boldsymbol{a}, \boldsymbol{b}\rangle \cdot \mathbb{E}\frac{1}{d}\,\mathsf{Tr}(z\boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1} \\ &= \langle \boldsymbol{a}, \boldsymbol{b}\rangle \cdot \mathbb{E}G_{\mathsf{esd}(\widehat{\boldsymbol{W}})}(z) \end{aligned} \tag{4.2.1}$$

As $d \to \infty$, the remaining quantity converges to $G_{\mu_{\mathsf{SC}}}(z)$ by the argument in Section 2.7.

The rigorous version of what is alluded to above is called an *isotropic local law*, which amounts to a statement about concentration of $\boldsymbol{a}^\top(z\boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1}\boldsymbol{b}$ around its expectation, for a given $\boldsymbol{a}$ and $\boldsymbol{b}$ but uniformly over a large region of $z \in \mathbb{C}\setminus[-2,2]$. We continue the derivation taking the above proposal for granted, and discuss it further below in Section 4.2.3.

Heuristic Proof of Theorem 4.2.2. We continue writing $\widehat{\boldsymbol{W}} := \frac{1}{\sqrt{d}}\boldsymbol{W}$ and $\widehat{\boldsymbol{Y}} := \frac{1}{\sqrt{d}}\boldsymbol{Y}$. Write $\lambda_1 := \lambda_1(\widehat{\boldsymbol{Y}})$. Suppose that $\lambda_1 > 2$ is in fact an outlier eigenvalue, with eigenvector $\boldsymbol{v}_1$. This means

$$\lambda_1 \boldsymbol{v}_1 = \widehat{\boldsymbol{Y}}\boldsymbol{v}_1 = \beta\langle \boldsymbol{v}_1, \boldsymbol{x}\rangle \boldsymbol{x} + \widehat{\boldsymbol{W}}\boldsymbol{v}_1,$$

and we may reorganize to summon the resolvent of $\widehat{\boldsymbol{W}}$,

$$(\lambda_1 \boldsymbol{I} - \widehat{\boldsymbol{W}})\boldsymbol{v}_1 = \beta\langle \boldsymbol{v}_1, \boldsymbol{x}\rangle \boldsymbol{x},$$

whereby

$$\boldsymbol{v}_1 = \beta\langle \boldsymbol{v}_1, \boldsymbol{x}\rangle \cdot (\lambda_1 \boldsymbol{I} - \widehat{\boldsymbol{W}})^{-1}\boldsymbol{x},$$

where the inversion is justified since $\lambda_1 > 2$ is outside the typical support of the spectrum of $\widehat{\boldsymbol{W}}$.

We will use our above heuristic twice. First, take the inner product of either side with $\boldsymbol{x}$:

$$\langle \boldsymbol{v}_1, \boldsymbol{x}\rangle = \beta\langle \boldsymbol{v}_1, \boldsymbol{x}\rangle \cdot \boldsymbol{x}^\top(\lambda_1 \boldsymbol{I} - \widehat{\boldsymbol{W}})^{-1}\boldsymbol{x} \approx \beta\langle \boldsymbol{v}_1, \boldsymbol{x}\rangle G_{\mu_{\mathsf{SC}}}(\lambda_1).$$

Rearranging, we find

$$G_{\mu_{\mathsf{SC}}}(\lambda_1) = \frac{1}{\beta}.$$

We can check that $G_{\mu_{\mathsf{SC}}}(z) = \frac{1}{2}(z - \sqrt{z^2 - 4}) < 1$ whenever $z > 2$ (consult Figure 2.6), so we must have $\beta > 1$ for our calculation to be consistent with $\lambda_1 > 2$ being an outlier. And,

our earlier calculation $R_{\mu_{\mathrm{SC}}}(z) = G^{-1}_{\mu_{\mathrm{SC}}}(z) - \frac{1}{z} = z$ implies that $G^{-1}_{\mu_{\mathrm{SC}}}(z) = z + \frac{1}{z}$, and thus this implies

$$\lambda_1 = \beta + \frac{1}{\beta},$$

as claimed.

To study the top eigenvector, consider instead taking the squared norm of either side above. Then we have

$$1 = \|\boldsymbol{v}_1\|^2 = \beta^2 \langle \boldsymbol{v}_1, \boldsymbol{x} \rangle^2 \boldsymbol{x}^\top (\lambda_1 \boldsymbol{I} - \widehat{\boldsymbol{W}})^{-2} \boldsymbol{x}.$$

We should not use our heuristic above, since $(\lambda_1 \boldsymbol{I} - \widehat{\boldsymbol{W}})^{-1}$ no longer acts as a multiple of the identity on $(\lambda_1 \boldsymbol{I} - \widehat{\boldsymbol{W}})^{-1}\boldsymbol{x}$, which depends on $\widehat{\boldsymbol{W}}$. But, we may cleverly argue as follows:

$$\frac{d}{dz}(z\boldsymbol{I} - \widehat{\boldsymbol{W}})^{-1} = -(z\boldsymbol{I} - \widehat{\boldsymbol{W}})^{-2}.$$

Thus we expect

$$\boldsymbol{x}^\top (z\boldsymbol{I} - \widehat{\boldsymbol{W}})^{-2} \boldsymbol{x} = -\frac{d}{dz}\left[\boldsymbol{x}^\top (z\boldsymbol{I} - \widehat{\boldsymbol{W}})^{-1} \boldsymbol{x}\right] \approx -\frac{d}{dz} G_{\mu_{\mathrm{SC}}}(z) = -G'_{\mu_{\mathrm{SC}}}(z),$$

whereby, substituting our derivation of λ_1,

$$1 = -\beta^2 \langle \boldsymbol{v}_1, \boldsymbol{x} \rangle^2 G'_{\mu_{\mathrm{SC}}}(\lambda_1) = -\beta^2 \langle \boldsymbol{v}_1, \boldsymbol{x} \rangle^2 G'_{\mu_{\mathrm{SC}}}\left(\beta + \frac{1}{\beta}\right),$$

or

$$\langle \boldsymbol{v}_1, \boldsymbol{x} \rangle^2 = -\frac{1}{\beta^2 G'_{\mu_{\mathrm{SC}}}\left(\beta + \frac{1}{\beta}\right)}.$$

A calculation remains:

$$G'_{\mu_{\mathrm{SC}}}(z) = \frac{1}{2}\left(1 - \frac{z}{\sqrt{z^2 - 4}}\right),$$

and thus

$$\begin{aligned} G'_{\mu_{\mathrm{SC}}}\left(\beta + \frac{1}{\beta}\right) &= \frac{1}{2}\left(1 - \frac{\beta + \frac{1}{\beta}}{\sqrt{(\beta + \frac{1}{\beta})^2 - 4}}\right) \\ &= \frac{1}{2}\left(1 - \frac{\beta + \frac{1}{\beta}}{\sqrt{(\beta - \frac{1}{\beta})^2}}\right) \end{aligned}$$

and, choosing a sign of the square root that will make the final answer positive as it must be,

$$\begin{aligned} &= \frac{1}{2}\left(1 - \frac{\beta + \frac{1}{\beta}}{\beta - \frac{1}{\beta}}\right) \\ &= \frac{1}{2}\left(1 - \frac{\beta^2 + 1}{\beta^2 - 1}\right) \\ &= -\frac{1}{\beta^2 - 1}, \end{aligned}$$

which gives

$$\langle \boldsymbol{v}_1, \boldsymbol{x} \rangle^2 = \frac{1}{\beta^2 \cdot \frac{1}{\beta^2 - 1}} = \frac{\beta^2 - 1}{\beta^2} = 1 - \frac{1}{\beta^2}.$$

Note that, had we chosen the other sign for the square root, we would have gotten $G'_{\mu_{\mathrm{SC}}}(\beta + \frac{1}{\beta}) = \frac{\beta^2}{\beta^2 - 1} > 0$, giving a nonsense result of $\langle \boldsymbol{v}_1, \boldsymbol{x} \rangle^2 < 0$. □

Remark 4.2.3 (More spikes)**.** *The same argument also works for a model of the form*

$$\boldsymbol{Y} = \boldsymbol{W} + \sqrt{d} \sum_{i=1}^{k} \beta_i \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

for $\|\boldsymbol{x}_i\| = 1$*. The outcome is just a superposition of* k *copies of the case* $k = 1$*: there are a number of outlier eigenvalues equal to the number of* $\beta_i > 1$*, and the corresponding eigenvectors have the same correlations with the corresponding* $\boldsymbol{x}_i$ *as in Theorem 4.2.2.*

4.2.3 Isotropic Local Laws

To make the above argument precise requires a result called an *isotropic local* (semicircle, in our case) *law*. This is also similar to what would be required to make rigorous the Stieltjes transform argument for the semicircle limit theorem that we saw before in Section 2.7

An isotropic local law is just a statement of the form we alluded to in the intuition above, that, for any deterministic $\boldsymbol{a}, \boldsymbol{b}$ and any z quantitatively far from $[-2, 2]$,

$$\left| \boldsymbol{a}^\top (z \boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1} \boldsymbol{b} - G_{\mu_{\mathrm{SC}}}(z) \langle \boldsymbol{a}, \boldsymbol{b} \rangle \right|$$

is small with high probability. The results in Section 2.7 only asked for the diagonal entries of the resolvent, which corresponds to $\boldsymbol{a} = \boldsymbol{b} = \boldsymbol{e}_i$, while here we want to consider arbitrary vectors, but this can still be handled with similar methods based on Schur complement identities.

For us it is enough to look at $z > 2 + \epsilon$ real, or more generally z with $\mathsf{dist}(z, [-2, 2]) > \epsilon$. However, one may also study finer-grained properties of eigenvalues by establishing these results at distances $\mathsf{dist}(z, [-2, 2]) = \epsilon(n) = o(1)$ depending on n. This should be logical: after all, if we have access to $\mathsf{Im}(G_{\mathsf{esd}(\widehat{\boldsymbol{W}})}(z))$ for all such z, we have access to the convolution of $\mathsf{esd}(\widehat{\boldsymbol{W}})$ with $\mathsf{Cauchy}(\epsilon)$, as we observed in Section 2.7). See, e.g., [KY13] and references therein for further information.

For our purposes, concretely it is possible to show the following: for any $c, C, \delta > 0$, for n sufficiently large,

$$\mathbb{P}\Bigg[\left| \boldsymbol{a}^\top (z \boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1} \boldsymbol{b} - G_{\mu_{\mathrm{SC}}}(z) \langle \boldsymbol{a}, \boldsymbol{b} \rangle \right| \le n^{-1/2+\delta} \\ \text{for all } z \in \mathbb{C} \text{ with } \mathsf{dist}(z, [-2, 2]) > c, |z| < C \Bigg] \ge 1 - \frac{1}{n^{100}}.$$

The idea of the proof is to union bound over a grid of values of z, and for each z on the grid to decompose

$$\left| \boldsymbol{a}^\top (z \boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1} \boldsymbol{b} - G_{\mu_{\mathrm{SC}}}(z) \langle \boldsymbol{a}, \boldsymbol{b} \rangle \right| \\ \le \left| \boldsymbol{a}^\top (z \boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1} \boldsymbol{b} - \mathbb{E} \boldsymbol{a}^\top (z \boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1} \boldsymbol{b} \right| + \left| \mathbb{E} \boldsymbol{a}^\top (z \boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1} \boldsymbol{b} - G_{\mu_{\mathrm{SC}}}(z) \langle \boldsymbol{a}, \boldsymbol{b} \rangle \right|$$

where the second term, by our previous calculation in (4.2.1), is

$$\leq \left| \boldsymbol{a}^\top (z\boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1}\boldsymbol{b} - \mathbb{E}\boldsymbol{a}^\top (z\boldsymbol{I}_d - \widehat{\boldsymbol{W}})^{-1}\boldsymbol{b} \right| + \left| \mathbb{E}G_{\mathsf{esd}(\widehat{\boldsymbol{W}})}(z) - G_{\mu_{\mathsf{sc}}}(z) \right| \cdot |\langle \boldsymbol{a}, \boldsymbol{b} \rangle|.$$

The first term may be bounded by general concentration inequalities on functions of Gaussian random variables, which we will see later. Bounding the second term amounts to a kind of quantitative version of the proof of the semicircle limit theorem sketched in Section 2.7.

4.3 Spiked Covariance (Wishart) Model

Let us now consider a similar question in the more realistic setting of covariance estimation. We consider, as before, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ for $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(d)} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ positive semidefinite and with $\frac{d}{m} \to c$. But now, instead of $\boldsymbol{\Sigma}$ having eigenvalues converging to a non-trivial shape, we give it only a single exceptional outlier eigenvalue:

$$\boldsymbol{\Sigma} := \boldsymbol{I}_d + \beta \boldsymbol{x}\boldsymbol{x}^\top$$

for some $\|\boldsymbol{x}\| = 1$ and $\beta \in \mathbb{R}_{\geq 0}$. We construct the sample covariance

$$\hat{\boldsymbol{\Sigma}} := \frac{1}{m}\sum_{i=1}^m \boldsymbol{x}_i \boldsymbol{x}_i^\top,$$

and, as before, seek to estimate $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\Sigma}}$. We have already seen that this is not in general a good idea: even if $\beta = 0$, we will have $\mathsf{esd}(\hat{\boldsymbol{\Sigma}}) \to \mu_{\mathsf{MP}(c)} \neq \delta_1 = \mathsf{esd}(\boldsymbol{\Sigma})$. But, maybe we can still use this poor estimator to detect the outlier eigenvalue in $\boldsymbol{\Sigma}$?

We therefore ask the same question from the previous section of $\hat{\boldsymbol{\Sigma}}$: when does it have an outlier eigenvalue? And, when does its top eigenvector give a good estimate of $\boldsymbol{x}$? In fact, essentially the same phenomenon holds here as in the spiked additive model.

Theorem 4.3.1 ([BBAP05, Pau07])**.** *Consider $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}^{(d,\beta,c)}$ as above, for fixed $\beta, c > 0$ and $d \to \infty$ and any deterministic $\boldsymbol{x} = \boldsymbol{x}^{(d)}$ a unit vector. Write $\boldsymbol{v}_1(\hat{\boldsymbol{\Sigma}})$ for the unit eigenvector of $\lambda_1(\hat{\boldsymbol{\Sigma}})$. The following hold, with all convergences in probability:*

- *If $\beta < \sqrt{c}$, then:*

$$\lambda_1(\hat{\boldsymbol{\Sigma}}) \to (1+\sqrt{c})^2,$$
$$\langle \boldsymbol{v}_1(\hat{\boldsymbol{\Sigma}}), \boldsymbol{x}\rangle^2 \to 0.$$

- *If $\beta > \sqrt{c}$, then:*

$$\lambda_1(\hat{\boldsymbol{\Sigma}}) \to (1+\beta)\left(1+\frac{c}{\beta}\right) > (1+\sqrt{c})^2,$$
$$\langle \boldsymbol{v}_1(\hat{\boldsymbol{\Sigma}}), \boldsymbol{x}\rangle^2 \to \left(1-\frac{c}{\beta^2}\right)\left(1+\frac{c}{\beta}\right) > 0.$$

The proof is a more complicated version of the same argument using the resolvent that we have sketched for the additive model.

4.4 Application: Community Detection

We can use these tools to make predictions about a problems asking us to detect or estimate *community structure* in graphs.

4.4.1 Two Balanced Communities

One simple example is the *stochastic block model*. Consider a graph on d vertices, divided into two groups of $\frac{d}{2}$ each, say $S \sqcup T = [d]$. We then generate a random graph G with independent edges where:

$$\mathbb{P}[i \sim j] = \left\{ \begin{array}{ll} \frac{1}{2} + \beta & \text{if } i,j \in S \text{ or } i,j \in T, \\ \frac{1}{2} - \beta & \text{otherwise} \end{array} \right\}.$$

That is, edges are more frequent within communities than between them.

One way to get a hint of the matrix structure of this problem is to introduce $\boldsymbol{x} \in \{\pm 1\}^d$, the indicator of membership in the two groups. In that case,

$$\mathbb{P}[i \sim j] = \frac{1}{2} + \beta x_i x_j.$$

Thus, writing $\boldsymbol{A}$ for the $\{\pm 1\}$-valued adjacency matrix, we have

$$\mathbb{E}A_{ij} = \mathbb{P}[i \sim j] \cdot (+1) + \mathbb{P}[i \not\sim j] \cdot (-1) = 2\beta x_i x_j,$$

or equivalently

$$\mathbb{E}\boldsymbol{A} = 2\beta \boldsymbol{x}\boldsymbol{x}^\top.$$

(This is not correct on the diagonal, but let us omit these corrections, which do not affect the result.)

We employ a common device of extracting this "signal" by recentering $\boldsymbol{A}$:

$$\boldsymbol{A} = (\mathbb{E}\boldsymbol{A}) + (\underbrace{\boldsymbol{A} - \mathbb{E}\boldsymbol{A}}_{=:\boldsymbol{W}}) = \boldsymbol{W} + 2\beta \boldsymbol{x}\boldsymbol{x}^\top.$$

If $\beta = o(1)$, then, conditional on $\boldsymbol{x}$, $\boldsymbol{W}$ is nearly a Wigner matrix: its entries are independent, centered, and have variance $1 + o(1)$. They are not i.i.d., but it turns out that similar arguments to the i.i.d. case can still be carried out. Since $\|\boldsymbol{x}\| = \sqrt{d}$, we see that this model will have the scaling of the additive spiked matrix model when we take $\beta \sim 1/\sqrt{d}$. So, let us set

$$\begin{aligned} \widehat{\beta} &:= \sqrt{d}\beta, \\ \widehat{\boldsymbol{x}} &:= \frac{1}{\sqrt{d}}\boldsymbol{x} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}. \end{aligned}$$

With these definitions, we have

$$\boldsymbol{A} = 2\widehat{\beta}\sqrt{d}\,\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}^\top + \boldsymbol{W}.$$

Thus, the threshold corresponding to the spiked Wigner matrix transition is $\widehat{\beta}_* = 1/2$. Using a more general isotropic local law (for "nearly Wigner" matrices), you can show the following rigorously.

Theorem 4.4.1 (Spectral algorithm for stochastic block model)**.** *Suppose that $\hat{\beta} \geq 0$ is fixed and $\beta := \hat{\beta}/\sqrt{d}$ in the above model. Let $\boldsymbol{A} \in \{0, \pm 1\}^{d \times d}$ be the random adjacency matrix constructed as above. The following then hold, with all convergences in probability:*

- *If $\hat{\beta} < 1/2$ (including if $\hat{\beta} = 0$), then:*

$$\frac{1}{\sqrt{d}}\lambda_1(\boldsymbol{A}) \to 2,$$
$$\frac{1}{d}\langle \boldsymbol{v}_1(\boldsymbol{A}), \boldsymbol{x}\rangle^2 \to 0.$$

- *If $\hat{\beta} > 1/2$, then:*

$$\frac{1}{\sqrt{d}}\lambda_1(\boldsymbol{A}) \to 2\hat{\beta} + \frac{1}{2\hat{\beta}} > 2,$$
$$\frac{1}{d}\langle \boldsymbol{v}_1(\boldsymbol{A}), \boldsymbol{x}\rangle^2 \to 1 - \frac{1}{4\hat{\beta}^2} > 0.$$

From the results on eigenvalues, we can make a strong claim immediately about *detecting* or *hypothesis testing* for community structure: if $\hat{\beta} > 1/2$, then computing and thresholding $\lambda_1(\boldsymbol{A})$ with high probability distinguishes the model above from the *null model* of an *Erdős-Rényi random graph*, the case $\hat{\beta} = 0$ where each edge is just present with probability $1/2$ independently.

To make a result about estimating the community assignments $\boldsymbol{x}$, we must be a little bit more careful. Set $\boldsymbol{v} := \sqrt{d} \cdot \boldsymbol{v}_1(\boldsymbol{A})$, so that $\|\boldsymbol{v}\| = \|\boldsymbol{x}\| = \sqrt{d}$. The natural estimator of $\boldsymbol{x}$ is to take $\boldsymbol{y} := \pm \operatorname{sgn}(\boldsymbol{v})$, the entrywise sign. Note that $\boldsymbol{v}$ itself is only determined up to a global sign flip, so the sign ambiguity is inescapable. Indeed, $\boldsymbol{x}$ itself is only identified up to a sign flip, since $\boldsymbol{x}$ and $-\boldsymbol{x}$ give rise to the same (in law) random $\boldsymbol{A}$. So, let us choose this sign arbitrarily.

We may then analyze the performance of this as follows: we know that $|\langle \boldsymbol{v}, \boldsymbol{x}\rangle| \geq (1-\epsilon)d$ (here $\epsilon = \epsilon(\hat{\beta}) = 1 - \sqrt{1 - 1/4\hat{\beta}^2} \leq 1/4\hat{\beta}^2$ can be computed from the result). We also know

$$|\langle \boldsymbol{v}, \boldsymbol{y}\rangle| = |\langle \boldsymbol{v}, \operatorname{sgn}(\boldsymbol{v})\rangle| = \sum_{i=1}^{d} |v_i| = \max_{\boldsymbol{s} \in \{\pm 1\}^d} |\langle \boldsymbol{v}, \boldsymbol{s}\rangle| \geq |\langle \boldsymbol{v}, \boldsymbol{x}\rangle| \geq (1-\epsilon)d.$$

Since $\|\boldsymbol{v}\| = \|\boldsymbol{y}\| = \sqrt{d}$ we may convert these bounds into

$$\min\{\|\boldsymbol{v} - \boldsymbol{x}\|^2, \|\boldsymbol{v} + \boldsymbol{x}\|^2\} = 2d - 2\langle \boldsymbol{v}, \boldsymbol{x}\rangle \leq 2\epsilon d,$$
$$\min\{\|\boldsymbol{v} - \boldsymbol{y}\|^2, \|\boldsymbol{v} + \boldsymbol{y}\|^2\} = 2d - 2\langle \boldsymbol{v}, \boldsymbol{y}\rangle \leq 2\epsilon d.$$

Thus by triangle inequality

$$\min\{\|\boldsymbol{x} - \boldsymbol{y}\|^2, \|\boldsymbol{x} + \boldsymbol{y}\|^2\} \leq (2\sqrt{2\epsilon d})^2 = 8\epsilon d.$$

Reversing the previous manipulation,

$$(1 - 3\epsilon)d \leq |\langle \boldsymbol{x}, \boldsymbol{y}\rangle| = |\#\{i : x_i = y_i\} - \#\{i : x_i = -y_i\}|.$$

and finally this implies a direct bound on our estimation error, showing that either $\boldsymbol{y} = \mathsf{sgn}(\boldsymbol{v}_1(\boldsymbol{A}))$ or $-\boldsymbol{y} = -\mathsf{sgn}(\boldsymbol{v}_1(\boldsymbol{A}))$ is a good estimate of $\boldsymbol{x}$:

$$\max\left\{\#\{i : x_i = y_i\}, \#\{i : x_i = -y_i\}\right\} \geq \left(1 - \frac{3}{2}\epsilon\right) d \geq \left(1 - \frac{3}{8\widehat{\beta}^2}\right) d.$$

This simple argument shows that, past the slightly larger threshold $\widehat{\beta} > \sqrt{3/8} \approx 0.612$, the estimator $\mathsf{sgn}(\boldsymbol{v}_1(\boldsymbol{A}))$ achieves non-trivial estimation of the community structure of the graph. More careful analysis shows that in fact $\widehat{\beta} > 1/2$ already suffices.

4.4.2 More Communities, Different Sizes

We may also propose a more expressive model as follows: suppose we instead have k communities $S_1 \sqcup \cdots \sqcup S_k = [d]$, and $|S_i| \approx \alpha_i d$ for some $\alpha_1, \ldots, \alpha_k \geq 0$ with $\sum \alpha_i = 1$. We can also have a matrix of interaction strengths between the communities $\boldsymbol{B} \in [0,1]^{k \times k}$, such that, if $\sigma : [d] \to [k]$ returns the label of the community that i belongs to, then we sample edges independently with

$$\mathbb{P}[i \sim j] = B_{\sigma(i)\sigma(j)}.$$

For this to make sense as an undirected graph, $\boldsymbol{B}$ should be symmetric. The previous case of two balanced communities is $\alpha_1 = \alpha_2 = \frac{1}{2}$ and

$$\boldsymbol{B} = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix}^\top + a\begin{bmatrix} 1 \\ -1 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix}^\top.$$

We see that there is a matter of scaling to clarify, since previously the interesting transition occurred when $a \sim 1/\sqrt{d}$. To that end, let $\boldsymbol{A}$ at first just be the $\{0,1\}$-valued adjacency matrix. This will not be close to a Wigner matrix; for instance, the entries will not be centered. To figure out what to do about this, consider the statistics of an entry chosen uniformly at random: for $i, j \sim \mathsf{Unif}([d])$, we have

$$\begin{aligned} \mathbb{E}A_{ij} &= \sum_{a,b=1}^{k} \alpha_a \alpha_b B_{ab} = \boldsymbol{\alpha}^\top \boldsymbol{B} \boldsymbol{\alpha} =: \beta, \\ \mathsf{Var}\, A_{ij}^2 &= \mathbb{E}A_{ij} - (\mathbb{E}A_{ij})^2 = b(1-b) =: \sigma^2. \end{aligned}$$

Previously, we had $\beta = \frac{1}{2}$ and $\sigma^2 = \frac{1}{4}$. A natural centered and normalized $\boldsymbol{A}$, in particular one that is close to a Wigner matrix, is then

$$\widehat{\boldsymbol{A}} := \frac{1}{\sigma}(\boldsymbol{A} - \beta \mathbf{1}\mathbf{1}^\top) = \frac{1}{\sqrt{\beta(1-\beta)}}(\boldsymbol{A} - \beta \mathbf{1}\mathbf{1}^\top).$$

You can check that, in the previous case, this is precisely the $\{\pm 1\}$-valued adjacency matrix.

This also tells us what the natural "null model" is that we should compare against: the model where edges are drawn i.i.d. with the same probability, i.e., where $A_{ij} \overset{\text{iid}}{\sim} \mathsf{Ber}(\beta)$. This, with the same transformation as above, would lead to $\widehat{\boldsymbol{A}}$ truly being a Wigner matrix, with i.i.d. entries of mean 0 and variance 1.

Up to permutations, the expectation of $\boldsymbol{A}$ itself (not a random entry as above) will be a block matrix of the form

$$\mathbb{E}\boldsymbol{A} = \left[\begin{array}{c} B_{ab}\mathbf{1}_{\alpha_a d}\mathbf{1}_{\alpha_b d}^\top \end{array} \right]_{a,b\in[k]}.$$

Thus our decomposition of $\widehat{\boldsymbol{A}}$ will take the form

$$\widehat{\boldsymbol{A}} = \underbrace{\mathbb{E}\widehat{\boldsymbol{A}}}_{=:\boldsymbol{X}} + (\underbrace{\widehat{\boldsymbol{A}} - \mathbb{E}\widehat{\boldsymbol{A}}}_{=:\boldsymbol{W}}) = \boldsymbol{X} + \boldsymbol{W}.$$

Note that $\boldsymbol{X}$ is a determinstic matrix that is just a function of our parameters, in this case $\boldsymbol{\alpha}$ and $\boldsymbol{B}$. In the previous case, we had

$$\boldsymbol{X} = a \left[\begin{array}{c} 1 \\ -1 \end{array} \right] \left[\begin{array}{c} 1 \\ -1 \end{array} \right]^\top$$

having entries of order $O(1/\sqrt{d})$, and we will see that the same scaling is natural in this more general case.

We may compute

$$\boldsymbol{X} = \mathbb{E}\widehat{\boldsymbol{A}} = \left[\begin{array}{c} \frac{B_{ab}-\beta}{\sqrt{\beta(1-\beta)}}\mathbf{1}_{\alpha_a d}\mathbf{1}_{\alpha_b d}^\top \end{array} \right]_{a,b\in[k]} =: \frac{1}{\sqrt{d}}\left[\begin{array}{c} \hat{B}_{ab}\mathbf{1}_{\alpha_a d}\mathbf{1}_{\alpha_b d}^\top \end{array} \right]_{a,b\in[k]},$$

where we write $\widehat{\boldsymbol{B}}$ for the matrix with

$$\hat{B}_{ab} := \sqrt{d} \cdot \frac{B_{ab}-\beta}{\sqrt{\beta(1-\beta)}},$$

a kind of normalization of $\boldsymbol{B}$ with respect to the α_a's. We assume that $\widehat{\boldsymbol{B}}$ has constant entries as $d \to \infty$.

$\boldsymbol{W}$ above will again be approximately a Wigner matrix, so the behavior of this model will depend on the eigenvalues of $\boldsymbol{X}$. In words, the "signal part" $\boldsymbol{X}$ of the decomposition of $\widehat{\boldsymbol{A}}$ above is a block matrix with block (a,b) having size $\alpha_a d \times \alpha_b d$, and with entries all equal to $\hat{B}_{ab}$ on this block. What are the eigenvalues of such a matrix? First, suppose $\widehat{\boldsymbol{B}} = \sum_{i=1}^k \rho_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$ is the spectral decomposition, with $\boldsymbol{v}_i \in \mathbb{R}^k$. Then, let $\widetilde{\boldsymbol{v}}_i \in \mathbb{R}^d$ be a vector with k blocks, with block number a equal to $v_{ia}\mathbf{1}_{\alpha_a d}$, a kind of "unbalanced Kronecker product" of $\boldsymbol{v}_i$ with vectors of all ones of different sizes. We then have

$$\boldsymbol{X} = \frac{1}{\sqrt{d}} \sum_{i=1}^k \lambda_i \widetilde{\boldsymbol{v}}_i \widetilde{\boldsymbol{v}}_i^\top.$$

On the other hand, if the α_i are not constant, then this is no longer a spectral decomposition, since the $\widetilde{\boldsymbol{v}}_i$ are no longer orthogonal.

The above does tell us that $\mathsf{rank}(\boldsymbol{X}) \le k$, and that the column (equivalently, row) space of $\boldsymbol{X}$ is the space of vectors that are constant on blocks of the above sizes. Thus, consider a general $\boldsymbol{w} \in \mathbb{R}^k$, lifted in the above fashion to $\widetilde{\boldsymbol{w}} \in \mathbb{R}^d$ which has entries w_1 on the first $\alpha_1 d$ entries, entries w_2 on the next $\alpha_2 d$ entries, and so forth. $\boldsymbol{X}\widetilde{\boldsymbol{w}}$ then has the same block structure, where block a has entries $\frac{1}{\sqrt{d}}\sum_{b=1}^k \hat{B}_{ab} \cdot \alpha_b d \cdot w_b = \sqrt{d}(\widehat{\boldsymbol{B}}\boldsymbol{D}\boldsymbol{w})_a$ for $\boldsymbol{D} \in \mathbb{R}^{k\times k}$

the diagonal matrix of the α_a. Thus, the eigenvalues λ of $\boldsymbol{X}$ satisfy the equation, for some $\boldsymbol{w} \in \mathbb{R}^k$, that

$$\sqrt{d}\widehat{\boldsymbol{B}}\boldsymbol{D}\boldsymbol{w} = \lambda \boldsymbol{w}.$$

Multiplying on either side by $\boldsymbol{D}^{1/2}$, this is equivalently

$$(\sqrt{d}\boldsymbol{D}^{1/2}\widehat{\boldsymbol{B}}\boldsymbol{D}^{1/2})(\boldsymbol{D}^{1/2}\boldsymbol{w}) = \lambda(\boldsymbol{D}^{1/2}\boldsymbol{w}).$$

Therefore,

$$\boldsymbol{\lambda}(\boldsymbol{X}) = \sqrt{d}\boldsymbol{\lambda}(\boldsymbol{D}^{1/2}\widehat{\boldsymbol{B}}\boldsymbol{D}^{1/2}).$$

There is one last wrinkle, which is that, unlike in the previous case where we assumed $a \geq 0$, in this case the eigenvalues of a general $\boldsymbol{D}^{1/2}\widehat{\boldsymbol{B}}\boldsymbol{D}^{1/2}$ could be positive or negative. Fortunately, the same result as Theorem 4.2.2 holds both for more spikes (Remark 4.2.3) and for possibly negative spikes, with the same transition for negative outlier eigenvalues. This extended theory for spiked additive models then implies the following.

Theorem 4.4.2 (Spectral algorithm for detection in generalized stochastic block model)**.** *Let $\boldsymbol{A} \in \{0,1\}^{d\times d}_{\text{sym}}$ be a random matrix as constructed above. That is, there is $\sigma : [d] \to [k]$ for k constant as $d \to \infty$ such that $|\sigma^{-1}(a)| = \alpha_a d$ for some $\alpha_a \in (0,1)$ that are constant as $d \to \infty$, and*

$$A_{ij} \sim \text{Ber}\left(\beta + \frac{1}{\sqrt{d}} \cdot \sqrt{\beta(1-\beta)} \cdot \hat{B}_{ab}\right)$$

for some $\beta \in (0,1)$ and a matrix $\widehat{\boldsymbol{B}} \in \mathbb{R}^{k\times k}_{\text{sym}}$ that is also constant as $d \to \infty$. (These parameters may be computed from a stochastic block model on any number of communities with any interaction strengths, as we have done above.) Define

$$\widehat{\boldsymbol{A}} := \frac{1}{\sqrt{\beta(1-\beta)}}(\boldsymbol{A} - \beta \boldsymbol{1}\boldsymbol{1}^\top).$$

The following hold, with both convergences in probability:

- *If $\|\boldsymbol{D}^{1/2}\widehat{\boldsymbol{B}}\boldsymbol{D}^{1/2}\| < 1$ (including $\widehat{\boldsymbol{B}} = \boldsymbol{0}$), then*

$$\frac{1}{\sqrt{d}}\|\widehat{\boldsymbol{A}}\| \to 2.$$

- *If $\lambda := \|\boldsymbol{D}^{1/2}\widehat{\boldsymbol{B}}\boldsymbol{D}^{1/2}\| > 1$, then*

$$\frac{1}{\sqrt{d}}\|\widehat{\boldsymbol{A}}\| \to \lambda + \frac{1}{\lambda} > 2.$$

Thus, the spectral algorithm of examining $\boldsymbol{\lambda}(\widehat{\boldsymbol{A}})$ for outliers can detect the community structure in our graph if and only if

$$\|\boldsymbol{D}^{1/2}\widehat{\boldsymbol{B}}\boldsymbol{D}^{1/2}\| > 1.$$

This elegant characterization gives us a way to take a spectral property of our finitely many model parameters ($\boldsymbol{\alpha}$ which determines $\boldsymbol{D}$, and $\boldsymbol{B}$ which determines $\widehat{\boldsymbol{B}}$) and turn it into a characterization of when a spectral algorithm works or not as $d \to \infty$.

4.5 Application: Non-Gaussian Noise and Non-Linear PCA [PWBM16]

Theorem 4.2.2 holds with a noise distribution given by any Wigner matrix $\mathrm{Wig}(d, \mu)$ with, say, subgaussian entries, provided that μ has variance 1 to maintain the same scaling. That is, for all such choices of additive noise distribution, computing and thresholding $\lambda_1(\boldsymbol{Y})$ detects a spike if and only if $\beta > 1$.

Actually, whenever $\mu \neq \mathcal{N}(0,1)$, it is possible to do better—in this sense, Gaussian noise is the *hardest* noise distribution to handle of a given scale! This idea was proposed in the statistical physics literature by [KXZ16, LKZ15] and proved to work by [PWBM18].

4.5.1 Power of Entrywise Non-Linearities

Consider a larger class of algorithms, where, before computing $\lambda_1(\boldsymbol{Y})$, we apply an entrywise non-linear function to $\boldsymbol{Y}$. Say, we fix some $\phi : \mathbb{R} \to \mathbb{R}$ and compute

$$\phi(\boldsymbol{Y})_{ij} := \phi(Y_{ij}).$$

Could it be that $\lambda_1(\phi(\boldsymbol{Y}))$ is a more effective test statistic than $\lambda_1(\boldsymbol{Y})$? And, if so, how do we design ϕ?

We will make another assumption, that the entries of $\boldsymbol{x}$ are relatively "flat," say having $|x_i| \lesssim \mathrm{polylog}(d)/\sqrt{d}$ for all i with high probability. Consider, say, $\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ or $\boldsymbol{x} \sim \mathrm{Unif}(\{\pm 1/\sqrt{d}\}^d)$. We have

$$\boldsymbol{Y} = \boldsymbol{W} + \beta\sqrt{d}\,\boldsymbol{x}\boldsymbol{x}^\top,$$

and thus the entries are

$$Y_{ij} = W_{ij} + \beta\sqrt{d}\,x_i x_j,$$

where the second term is much smaller than the first, of order about $O(1/\sqrt{d})$ under the above models. When we apply ϕ, it is sensible to take a Taylor expansion. Doing this and manipulating a bit further, we have:

$$\begin{aligned}
\phi(\boldsymbol{Y})_{ij} &= \phi(W_{ij} + \beta\sqrt{d}x_i x_j) \\
&= \phi(W_{ij}) + \phi'(W_{ij})\beta\sqrt{d}x_i x_j + \Delta_{ij}^{(1)} \\
&= \phi(W_{ij}) + \mathbb{E}[\phi'(W_{ij})]\beta\sqrt{d}x_i x_j + \Delta_{ij}^{(1)} + \Delta_{ij}^{(2)}.
\end{aligned}$$

Here $\Delta_{ij}^{(1)}$ is the error in the Taylor expansion, which we expect (for most entries) to scale as $|\Delta_{ij}^{(1)}| \lesssim (\beta\sqrt{d}x_i x_j)^2 = O(1/d)$. Thus we will have $\|\boldsymbol{\Delta}^{(1)}\| = O(1)$. The other error terms are

$$\Delta_{ij}^{(2)} = (\phi'(W_{ij}) - \mathbb{E}[\phi'(W_{ij})])\beta\sqrt{d}x_i x_j.$$

Thus $\boldsymbol{\Delta}^{(2)}$ is a random matrix with independent centered entries of order $O(1/\sqrt{d})$, and so again we expect $\|\boldsymbol{\Delta}^{(2)}\| = O(1)$.

In particular, the above expansion means

$$\phi(\boldsymbol{Y}) = \phi(\boldsymbol{W}) + (\mathbb{E}_{w\sim\mu}[\phi'(w)]\beta)\sqrt{d}\,\boldsymbol{x}\boldsymbol{x}^\top + \boldsymbol{\Delta}^{(1)} + \boldsymbol{\Delta}^{(2)},$$

and the first two terms have operator norm $\Theta(\sqrt{d})$ while the latter two have operator norm $O(1)$. Thus, $\lambda_1(\phi(\boldsymbol{Y}))$ is roughly the top eigenvalue of the first two terms, which form another spiked matrix model. In this model, the noise distribution is $\mathsf{Law}(\phi(w))$ for $w \sim \mu$, while the effective amount of signal is

$$\beta \cdot \mathbb{E}_{w\sim\mu}[\phi'(w)] = \beta \int \phi'(w)\rho(w)\,dw.$$

Note that we should also assume $\mathbb{E}\phi(w) = 0$, or else the noise will not be centered.

Actually, the real effective amount of signal needs to be normalized by the standard deviation of the new noise distribution, which is no longer one but rather:

$$\sigma^2 := \mathbb{E}_{w\sim\mu}\phi(w)^2 = \int \phi(w)^2\rho(w)\,dw.$$

Thus, the "effective β" is in fact:

$$\beta_{\mathsf{eff}} = \beta_{\mathsf{eff}}(\phi) = \beta \cdot \frac{1}{\sigma} \cdot \mathbb{E}[\phi'(w)] = \beta \cdot \underbrace{\frac{\int \phi'(w)\rho(w)\,dw}{\sqrt{\int \phi(w)^2\rho(w)\,dw}}}_{\text{"gain" from } \phi},$$

where the latter factor is the "gain" in the effective amount of signal from applying the entrywise non-linearity ϕ. As a sanity check, merely scaling ϕ has no effect on this gain, as we should expect. Concretely, $\frac{1}{\sigma}\phi(\boldsymbol{Y})$ will then behave like a spiked matrix model with noise variance 1 and signal strength β_{eff}.

4.5.2 Optimizing the Non-Linearity

We now try to maximize β_{eff}, and see if the result can be made greater than 1, the threshold of detectability in a spiked Wigner model. This is a problem in the *calculus of variations*. Since we must take ϕ such that $\mathbb{E}\phi(w) = \int \phi(w)\rho(w)\,dw = 0$, and we may scale ϕ by any constant, let us assume that the numerator of the "gain factor" above is 1 and try to minimize the denominator. This gives the functional optimization problem:

$$\begin{array}{ll} \text{minimize} & \int \phi(w)^2\rho(w)\,dw \\ \text{subject to} & \int \phi(w)\rho(w)\,dw = 0, \\ & \int \phi'(w)\rho(w)\,dw = 1. \end{array}$$

Suppose $\phi(w)$ is optimal for this problem. Consider adding a small $\delta(w)$ to $\phi(w)$. In order for the resulting ϕ to still satisfy the constraints, we must have

$$\int \delta(w)\rho(w)\,dw = \int \delta'(w)\rho(w)\,dw = 0.$$

The second of these constraints may also be rewritten, by integrating by parts, as

$$\int \delta(w)\rho'(w)\,dw = 0.$$

For small $\delta(w)$, we may expand the objective function to leading order:

$$\int (\phi(w) + \delta(w))^2 \rho(w)\, dw \approx \int \phi(w)^2 \rho(w) + 2\int \phi(w)\delta(w)\rho(w)\, dw.$$

Thus the second term must be zero for all $\delta(w)$ satisfying the above two constraints. This can only happen if (thinking in linear-algebraic terms) $\phi(w)\rho(w)$ lies in the span of the two functions appearing in the constraints with δ,

$$\phi(w)\rho(w) = a\rho(w) + b\rho'(w)$$

for some $a, b \in \mathbb{R}$. In other words, we must have

$$\phi(w) = a + b\frac{\rho'(w)}{\rho(w)}.$$

Finally, to determine these constants we recall our constraints on ϕ, which are two equations that require

$$0 = \int \phi(w)\rho(w)\, dw = a\int \rho(w)\, dw + b\int \rho'(w)\, dw = a + \int w\rho(w)\, dw = a,$$
$$1 = \int \phi'(w)\rho(w)\, dw = -\int \phi(w)\rho'(w)\, dw = -b\int \frac{\rho'(w)^2}{\rho(w)}\, dw.$$

Let us define

$$F = F(\mu) := \int \frac{\rho'(w)^2}{\rho(w)}\, dw.$$

Then, the above says that $a = 0$ and $b = -1/F$, so the optimal ϕ is

$$\phi(w) = -\frac{1}{F}\frac{\rho'(w)}{\rho(w)}.$$

This achieves an objective function of

$$\int \phi(w)^2\rho(w)\, dw = \frac{1}{F^2}\int \frac{\rho'(w)^2}{\rho(w)}\, dw = \frac{1}{F}$$

in our minimization, and the gain it gives in the amount of signal is therefore

$$\beta_{\text{eff}} = \beta \cdot \frac{1}{\sqrt{1/F}} = \beta\sqrt{F}.$$

We thus reach the following elegant result.

Theorem 4.5.1 ([PWBM16])**.** *Consider $\boldsymbol{Y}$ as in Theorem 4.2.2, but with $\boldsymbol{W} \sim \mathrm{Wig}(d, \mu)$. Suppose that μ has density $\rho(w) > 0$, and let $\phi(w) = -\frac{1}{F(\mu)}\frac{\rho'(w)}{\rho(w)}$. Write $\beta_*(\mu) := 1/\sqrt{F(\mu)}$. Then, whenever $\beta > \beta_*(\mu)$ then thresholding $\lambda_1(\phi(\boldsymbol{Y}))$ distinguishes this from $\boldsymbol{Y} \sim \mathrm{Wig}(d, \mu)$, and this choice of ϕ leads to the optimal such threshold.*

4.5.3 Fisher Information, Fisher Score, and Denoising

A few remarks are in order. $F(\mu)$ is called the *Fisher information* of μ, and may be thought of as quantifying how much information an observation $y = x + w$ for $w \sim \mu$ carries about x. When $F(\mu)$ is larger, then *denoising* y to estimate x is easier, which causes the threshold of detectability to decrease above. The Fisher information has the following several equivalent forms.

Proposition 4.5.2. *For any μ with a smooth density ρ that is positive everywhere,*

$$\begin{aligned} F(\mu) &= \int \frac{\rho'(w)^2}{\rho(w)}\, dw \\ &= \int ((\log \rho)'(w))^2 \rho(w)\, dw \\ &= \mathop{\mathbb{E}}_{w\sim\mu} ((\log \rho)'(w))^2 \\ &= -\int (\log \rho)''(w) \rho(w)\, dw \\ &= -\mathop{\mathbb{E}}_{w\sim\mu} (\log \rho)''(w). \end{aligned}$$

In the above denoising setup, it may be helpful to consider the function $L(x \mid y) := \rho(y-x)$, the *likelihood* of x given the observation y. Then $\ell(x \mid y) := \log \rho(y-x)$ is the *log-likelihood* often encountered in statistics, and the above expressions involving $\log \rho$ may be viewed as measuring geometric properties of the log-likelihood when $x \approx 0$. This is sensible in our setting, because we are precisely interested in denoising very small signals $x = O(1/\sqrt{d})$ that have suffered very large amounts of additive noise $w = O(1)$. So, for instance, one of the forms above can be viewed as

$$F(\mu) = \mathop{\mathbb{E}}_{y\sim\mu} \left(\frac{\partial}{\partial x} \ell(x \mid y) \Big|_{x=0} \right)^2.$$

To understand this quantity, suppose we actually want to estimate x by maximizing the likelihood numerically. Knowing that x is small, we can start at an uninformative guess $x^{(0)} = 0$, and then take a "gradient step," which in this simple one-dimensional setting just boils down to following the derivative of the function; the greater the derivative, the more we expect the log-likelihood to increase in our first step. The above then measures the magnitude of this rate of increase, which is a proxy for how informative y is about x.

Our denoising function, $\phi(w) = -\rho'(w)/\rho(w)$ (omitting the normalizing constant for a moment) is called the *Fisher score*, and is just the derivative we encountered above: $\phi(y) = \frac{\partial}{\partial x} \ell(x \mid y)|_{x=0}$. Thus applying ϕ to Y_{ij} really is performing the naive one gradient step approach to approximately maximize the likelihood sketched above, starting from $x^{(0)} = 0$. In fact, when we include the constant $-1/F$, note that the last interpretation of F above is equivalently

$$F(\mu) = -\mathop{\mathbb{E}}_{w\sim\mu} \frac{\partial^2}{\partial x^2} \ell(x \mid y) \Big|_{x=0},$$

and thus overall the function

$$\phi(y) = -\frac{\frac{\partial}{\partial x} \ell(x \mid y) \Big|_{x=0}}{\mathbb{E}_{w\sim\mu} \frac{\partial^2}{\partial x^2} \ell(x \mid y) \Big|_{x=0}}$$

is similar to taking a *Newton step* to maximize the likelihood (though if we really did that we would not take the expectation in the denominator above).

4.5.4 Gaussian Noise is Hardest

Now let us return to the special role of the standard Gaussian with regard to the Fisher information. The following describes this, and implies, together with Theorem 4.5.1, that $\lambda_1(\phi(\boldsymbol{Y}))$ is a superior test statistic to $\lambda_1(\boldsymbol{Y})$ for detecting a spike precisely when $\mu \neq \mathcal{N}(0,1)$.

Proposition 4.5.3. *$F(\mathcal{N}(0,1)) = 1$, while whenever $\mu \neq \mathcal{N}(0,1)$ has mean zero and variance one and a strictly positive smooth density, then $F(\mu) > 1$.*

Proof. For $\mu = \mathcal{N}(0,1)$, the density is $\rho(w) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}w^2)$, so $(\log\rho)''(w) = -1$ and the result follows from Proposition 4.5.2.

For the other result, one could go through a calculus of variations calculation like we did before. But there is also a simplifying trick, as follows: consider

$$\begin{aligned}
0 &\leq \int \frac{1}{\rho(w)}(\rho'(w) + w\rho(w))^2\,dw \\
&= F(\mu) + 1 + 2\int w\rho'(w)\,dw \\
&= F(\mu) + 1 - 2\int w^2\rho(w)\,dw \\
&= F(\mu) + 1 - 2 \\
&= F(\mu) - 1.
\end{aligned}$$

This shows that $F(\mu) \geq 1$ for all μ satisfying the assumptions. Further, equality is achieved if and only if $\rho'(w) + w\rho(w) = 0$. We may write this as having, for all test functions f, that

$$0 = \int f(w)(\rho'(w) + w\rho(w)) = -\int f'(w)\rho(w) + \int wf(w)\rho(w),$$

or equivalently that

$$\mathop{\mathbb{E}}_{w\sim\mu} f'(w) = \mathop{\mathbb{E}}_{w\sim\mu} wf(w),$$

which Exercise 1.6.6 shows uniquely characterizes $\mu = \mathcal{N}(0,1)$. □

Further, one may show the following, though we will not have time to go into the proof.

Theorem 4.5.4 ([PWBM16])**.** *Write $\mathbb{P}_{d,\beta} := \mathsf{Law}(\boldsymbol{W} + \beta\sqrt{d}\boldsymbol{x}\boldsymbol{x}^\top)$ where $W_{ij} = W_{ji} \sim \mathcal{N}(0,1)$ for all $i \leq j$ (including the diagonal) and where $\boldsymbol{x} \sim \mathsf{Unif}(\mathbb{S}^{d-1})$. Then, for all $0 < \beta < 1$, there exists no sequence of functions $f^{(d)} : \mathbb{R}^{d\times d}_{\mathsf{sym}} \to \{0,1\}$ such that*

$$\begin{aligned}
\lim_{d\to\infty} \mathbb{P}_{d,0}[f^{(d)}(\boldsymbol{Y}) = 0] &= 1, \\
\lim_{d\to\infty} \mathbb{P}_{d,\beta}[f^{(d)}(\boldsymbol{Y}) = 1] &= 1.
\end{aligned}$$

Such a sequence of functions is often referred to as a sequence of tests *or* hypothesis tests *that* strongly distinguish *or* achieve strong detection between $\mathbb{P}_{d,0}$ *(which is just the law of a matrix of i.i.d. random entries distributed as $\mathcal{N}(0,1)$) and $\mathbb{P}_{d,\beta}$.*

This result is often summarized as strong detection (in the above sense) being *information-theoretically impossible* when $\beta < 1$.

We may therefore summarize our findings as follows:

1. The spiked matrix model is computationally hardest for Gaussian noise, among all "nice" centered distributions of additive noise having a given variance.

2. For a spiked matrix model with additive non-Gaussian noise, non-linear PCA with a suitable entrywise non-linearity (the test statistic $\lambda_1(\phi(\boldsymbol{Y}))$ performs strictly better than ordinary PCA (the test statistic $\lambda_1(\boldsymbol{Y})$).

4.6 Exercises

Exercise 4.6.1. *Write $\boldsymbol{v}_1(\boldsymbol{X})$ for the unit-norm eigenvector of $\lambda_1(\boldsymbol{X})$ for $\boldsymbol{X} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$. Whenever this notation is used below, you may assume that $\lambda_1(\boldsymbol{X})$ occurs with multiplicity 1 as an eigenvalue of $\boldsymbol{X}$.*

Suppose $\boldsymbol{M} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$, and $\boldsymbol{\Delta}$ has the same dimensions as $\boldsymbol{M}$ with $\|\boldsymbol{\Delta}\| < \lambda_1(\boldsymbol{M}) - \lambda_2(\boldsymbol{M})$ (the matrix norm without a subscript always denotes the operator norm). You will show the perturbation inequality

$$\langle \boldsymbol{v}_1(\boldsymbol{M}), \boldsymbol{v}_1(\boldsymbol{M}+\boldsymbol{\Delta})\rangle^2 \geq 1 - \left(\frac{\|\boldsymbol{\Delta}\|}{\lambda_1(\boldsymbol{M}) - \lambda_2(\boldsymbol{M}) - \|\boldsymbol{\Delta}\|}\right)^2.$$

Follow these steps, where we abbreviate $\boldsymbol{v} := \boldsymbol{v}_1(\boldsymbol{M})$ and $\widetilde{\boldsymbol{v}} := \boldsymbol{v}_1(\boldsymbol{M}+\boldsymbol{\Delta})$.

1. *Show that $\lambda_1(\boldsymbol{M}) - \lambda_i(\boldsymbol{M}+\boldsymbol{\Delta}) \geq \lambda_1(\boldsymbol{M}) - \lambda_2(\boldsymbol{M}) - \|\boldsymbol{\Delta}\|$ for all $i \geq 2$.*

   *(**Hint:** You may use the Courant-Fischer min-max theorem. Look it up and take a minute to internalize it if you are not familiar with this.)*

2. *Using Part 1, show that $\|\boldsymbol{\Delta}\boldsymbol{v}\| \geq (\lambda_1(\boldsymbol{M}) - \lambda_2(\boldsymbol{M}) - \|\boldsymbol{\Delta}\|) \cdot \|(\boldsymbol{I} - \widetilde{\boldsymbol{v}}\widetilde{\boldsymbol{v}}^\top)\boldsymbol{v}\|$.*

   *(**Hint:** Expand $\boldsymbol{v}$ in the orthonormal basis of eigenvectors of $\boldsymbol{M}+\boldsymbol{\Delta}$.)*

3. *Complete the proof.*

Also show the following application:

4. *Suppose that $\boldsymbol{W} \sim \mathsf{GOE}(d)$, and let $\boldsymbol{x} \in \mathbb{R}^d$ with $\|\boldsymbol{x}\| = 1$. Let $\lambda > 0$ and consider the matrix $\boldsymbol{Y} = \lambda\sqrt{d}\,\boldsymbol{x}\boldsymbol{x}^\top + \boldsymbol{W}$ (as in Section 4.2). Show that there is a function $f(\lambda) \in \mathbb{R}$ such that $f(\lambda) \to 1$ as $\lambda \to \infty$ and such that, for any fixed $\lambda > 0$, we have that*

$$\lim_{d\to\infty} \mathbb{P}[\langle \boldsymbol{v}_1(\boldsymbol{Y}), \boldsymbol{x}\rangle^2 \geq f(\lambda)] = 1.$$

   You may use the Wigner edge limit theorem (Theorem 2.5.2). More colloquially, this says that the top eigenvector of $\boldsymbol{Y}$ can achieve an arbitrarily good estimate of a rank one perturbation of $\boldsymbol{W}$ of sufficiently large magnitude λ.

5 | Non-Asymptotic Theory I: Concentration of Spectral Statistics

In the past three chapters on limit theorems, free probability, and spiked matrix models, we have been studying *asymptotics* in random matrix theory. That is, we have been making statements about the $d \to \infty$ limit without precisely specifying rates of convergence. These have been statements like:

$$\begin{aligned} \mathsf{esd}(\boldsymbol{W}) &\to \mu \text{ (weakly)}, \\ \lambda_1(\boldsymbol{W}) &\to c \text{ (in probability)}. \end{aligned}$$

In this chapter, we will develop some aspects of the *non-asymptotic theory* of random matrices. Here, we will be more concerned with results for finite d, giving explicit bounds on quantities like

$$\begin{aligned} E(d) &:= \mathbb{E}\lambda_1(\boldsymbol{W}), \\ T(d,s) &:= \mathbb{P}[|\lambda_1(\boldsymbol{W}) - \mathbb{E}\lambda_1(\boldsymbol{W})| > s]. \end{aligned}$$

It is already natural and useful to ask such questions about the classical models we have been studying, like Wigner and Wishart matrices, and we will start out with those examples.

But we will see that, once we take the non-asymptotic perspective, the range of questions it makes sense to ask expands dramatically. When we only care about asymptotics, we must ask questions about sequences of random matrices whose statistics of interest (empirical spectral distribution, norm, and so on) converge in some sense. The kinds of matrices we have been looking at have been constrained by this requirement that they must make sense as part of a *sequence* of growing matrices with $d \to \infty$.

If we wanted, say, to ask about Gaussian random matrices whose entries are correlated in some complicated way, we would then have to orchestrate a sequence of such correlated models converging to some interesting limit. Once we only care about a single random matrix of a single dimension $d \times d$, we can ask about all the same questions about far more general models that do not come from any natural such sequence. We will see a sequence of tools for answering these questions, building up to recent developments suggesting that, often, we can have the best of both worlds: similarly precise information to the classical limit theorems, for almost arbitrary models of random matrices, with explicit conclusions and error bounds for finite d.

5.1 Generic Models of Random Matrices

5.1.1 Independent, Differently-Distributed Entries

5.1.2 Gaussian Series

Definition 5.1.1. *Consider any deterministic symmetric matrices $\boldsymbol{A}_0, \boldsymbol{A}_1, \dots, \boldsymbol{A}_D \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$, and let $g_1, \dots, g_D \sim \mathcal{N}(0,1)$ independently. We call the following random matrix the* Gaussian series *associated to the sequence $(\boldsymbol{A}_i)_{i=0}^D$:*

$$\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{g}) := \boldsymbol{A}_0 + \sum_{i=1}^{D} g_i \boldsymbol{A}_i.$$

Clearly any such $\boldsymbol{X}$ is a symmetric random matrix with entries that have a multivariate Gaussian joint distribution. In fact, the converse is also true, so such Gaussian series are a way of writing the most general possible Gaussian random matrix.

Definition 5.1.2. *For any matrix $\boldsymbol{X}$, write $\mathsf{symvec}(\boldsymbol{X}) \in \mathbb{R}^{d(d+1)/2}$ for the vectorization of the entries on and above the diagonal of $\boldsymbol{X}$, and $\mathsf{vec}(\boldsymbol{X}) \in \mathbb{R}^{d^2}$ for the vectorization of* all *entries of $\boldsymbol{X}$ (which will include some repetitions because $\boldsymbol{X}$ is symmetric).*

Proposition 5.1.3. *Suppose that $\boldsymbol{X}$ is a random symmetric matrix having $\mathsf{Law}(\mathsf{symvec}(\boldsymbol{X})) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, there exist $\boldsymbol{A}_0, \boldsymbol{A}_1, \dots, \boldsymbol{A}_D \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ such that $\mathsf{Law}(\boldsymbol{X}) = \mathsf{Law}(\boldsymbol{A}_0 + \sum_{i=1}^D g_i \boldsymbol{A}_i)$. Moreover, it is possible to take $D \le \frac{d(d+1)}{2}$ in the latter representation.*

Proof. Fix $D = \frac{d(d+1)}{2}$. Note that $\boldsymbol{\mu} \in \mathbb{R}^D$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}_{\mathsf{sym}}$ is positive semidefinite. Consider expanding $\boldsymbol{\Sigma} = \sum_{i=1}^D \boldsymbol{a}_i \boldsymbol{a}_i^\top$ (say, if $(\rho_i, \widetilde{\boldsymbol{a}}_i)$ are the eigenpairs of $\boldsymbol{\Sigma}$, then we can take $\boldsymbol{a}_i = \sqrt{\rho_i}\widetilde{\boldsymbol{a}}_i$). Then, we have that $\mathsf{Law}(\mathsf{symvec}(\boldsymbol{X})) = \mathsf{Law}(\boldsymbol{\mu} + \sum_{i=1}^D g_i \boldsymbol{a}_i)$ for $g_i \sim \mathcal{N}(0,1)$ independently. Undoing the $\mathsf{symvec}(\cdot)$ operation (which is a bijection between symmetric matrices and D-dimensional vectors) then gives the Gaussian series representation. □

Remark 5.1.4. *While it is always possible to take $D = \frac{d(d+1)}{2}$, sometimes it is easier to write an "overparametrized" Gaussian series with D larger than this, which we will always allow in our future results about these models.*

In particular then, all of the Gaussian models we have seen before can be written as Gaussian series models. For instance, the GOE corresponds to taking $\frac{d(d+1)}{2}$ many $\boldsymbol{A}_i$, indexed $\boldsymbol{A}_{(ij)}$ for $1 \le i \le j \le d$, where $\boldsymbol{A}_{ii} = \sqrt{2}\boldsymbol{e}_i\boldsymbol{e}_i^\top$ and $\boldsymbol{A}_{ij} = \boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top$. But, we can also write dramatically more general such series. The following is an interesting class of examples.

Example 5.1.5 (Patterned Gaussian matrices)**.** *Consider a partition $\{(i,j) : 1 \le i \le j \le d\} = S_1 \sqcup \cdots \sqcup S_D$. Let $\boldsymbol{A}_i \in \{0,1\}^{d\times d}_{\mathsf{sym}}$ have $(\boldsymbol{A}_i)_{ab} = \mathbb{1}\{(a,b) \in S_i\}$ for each $a \le b$, for $i = 1, \dots, D$, and set $\boldsymbol{A}_0 = \boldsymbol{0}$. Then, the associated Gaussian series is a random matrix with each entry distributed as $\mathcal{N}(0,1)$, but some entries "glued" to take equal values according to the partition of the S_i. For instance, taking the partition into diagonal subsets you can write Gaussian* Toeplitz *or* circulant *matrices.*

Remark 5.1.6 (Rectangular and asymmetric matrices)**.** *We may embed the question of the singular values of square asymmetric or even rectangular Gaussian matrices into this same framework: if $\boldsymbol{B} \in \mathbb{R}^{a\times b}$ has jointly Gaussian entries, then we may take*

$$\boldsymbol{X} := \begin{bmatrix} \boldsymbol{0} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^{(a+b)\times(a+b)},$$

whose eigenvalues you may check will be $\pm\sigma_i(\boldsymbol{B})$ for $\sigma_i(\boldsymbol{B})$ the singular values (together with some zero eigenvalues. This $\boldsymbol{X}$ is symmetric and has jointly Gaussian entries, and thus may be written as a Gaussian series.

5.2 Principles of Concentration Inequalities

The relationship between random matrix theory and concentration inequalities has several different sides. Recall that a classical or scalar concentration inequality for some random variable Y gives a bound of the form

$$\mathbb{P}[|Y - \mathbb{E}Y| > s] \leq T(s) \tag{5.2.1}$$

More specifically, the classical concentration inequalities you have probably seen (Chernoff, Hoeffding, Bernstein) all concern the special case

$$Y = \sum_{i=1}^{N} X_i$$

for X_i i.i.d. or at least independent.

We have already seen in our discussion of the ϵ-net method for bounding random matrices in Chapter 1 that this can be a useful tool for bounding the spectral norm of a random matrix. Recall how that worked: we had

$$\|\boldsymbol{M}\| = \max_{\boldsymbol{v}\in\mathbb{S}^{d-1}} |\boldsymbol{v}^\top \boldsymbol{M}\boldsymbol{v}|,$$

ϵ-nets gave us a way to discretize this maximum and the union bound gave us a way to control $\mathbb{P}[\|\boldsymbol{M}\| > s]$ by controlling each $\mathbb{P}[|\boldsymbol{v}^\top \boldsymbol{M}\boldsymbol{v}| > s]$. Finally, if the entries of $\boldsymbol{M}$ have lots of independence, then $\boldsymbol{v}^\top \boldsymbol{M}\boldsymbol{v}$ can be expanded into an expression that standard scalar concentration inequalities apply to.

While previously we looked at the case $\boldsymbol{M} = \boldsymbol{G}\boldsymbol{G}^\top - \mathbb{E}\boldsymbol{G}\boldsymbol{G}^\top$, let us quickly recap here how such an argument would look in the simpler case $\boldsymbol{M} = \boldsymbol{W} \sim \mathsf{GOE}(d)$. We may choose an ϵ-net $\mathcal{X}$ of $\mathbb{S}^{d-1}$ with $|\mathcal{X}| \leq (1 + \frac{2}{\epsilon})^d$, whereby

$$\begin{aligned} \mathbb{P}[\|\boldsymbol{W}\| > s] &= \mathbb{P}[\max_{\boldsymbol{v}\in\mathbb{S}^{d-1}} |\boldsymbol{v}^\top \boldsymbol{W}\boldsymbol{v}| > s] \\ &\leq \mathbb{P}[\max_{\boldsymbol{v}\in\mathcal{X}} |\boldsymbol{v}^\top \boldsymbol{W}\boldsymbol{v}| > (1-2\epsilon)s] \end{aligned}$$

where a simple calculation shows that $\mathsf{Law}(\boldsymbol{v}^\top \boldsymbol{W} \boldsymbol{v}) = \mathcal{N}(0,2)$ for all $\boldsymbol{v}$, so

$$\begin{aligned}
&\leq |\mathcal{X}| \mathop{\mathbb{P}}_{g \sim \mathcal{N}(0,2)} [|g| > (1-2\epsilon)s] \\
&\leq \left(1 + \frac{2}{\epsilon}\right)^d \exp\left(-\frac{(1-2\epsilon)^2}{4} \cdot s^2\right) \\
&= \exp\left(\log\left(1 + \frac{2}{\epsilon}\right) \cdot d - \frac{(1-2\epsilon)^2}{4} \cdot s^2\right).
\end{aligned}$$

Let us be very quantitative here: this will show a strong tail bound once $s \geq (C+\delta)\sqrt{d}$ for any $\delta > 0$, where

$$C = \sqrt{\frac{4\log(1+\frac{2}{\epsilon})}{(1-2\epsilon)^2}},$$

which, even if we minimize over ϵ, gives $C \approx 4.27$.

Getting to the point, this approach gives us a weak kind of information: even though we know $\|\boldsymbol{W}\| \approx 2\sqrt{d}$ with high probability, the above calculation only establishes tail decay beyond a very large deviation to $4\sqrt{d}$. One way to see what the problem is is that this calculation conflates two matters: that of what the typical value of $\|\boldsymbol{W}\|$ is, and that of how spread out the distribution of $\|\boldsymbol{W}\|$ is around that typical value. It may seem that these two are hopelessly intertwined: for instance, when you write the formula for the variance, a simple measure of the latter, it involves the expectation, a measure of the former, $Var[X] = \mathbb{E}(X - \mathbb{E}X)^2$. Yet we will see that in fact these two issues can be disentangled, and that this is a much more fruitful way to understand functions of many random inputs (like $\boldsymbol{W} \mapsto \|\boldsymbol{W}\|$) than the previous direct approach.

You may think of the following as the two very high-level guiding principles of concentration inequalities:

- **Principle 1: Separation of Location and Concentration.** It is possible to understand the behavior of the deviation probabilities $\mathbb{P}[|X - \mathbb{E}X| > s]$ *without* understanding $\mathbb{E}X$ at all.
- **Principle 2: Concentration of Non-Linear Functions.** A function of many random variables $f(\boldsymbol{x}) = f(x_1, \ldots, x_N)$ concentrates around $\mathbb{E}f(\boldsymbol{x})$ provided two properties hold:
  (a) the x_i are only weakly dependent, and
  (b) f is only weakly sensitive to each of its N inputs.

The first principle is describing what we suggested above. The second describes when we can hope for concentration of general functions like $\boldsymbol{W} \mapsto \|\boldsymbol{W}\|$. Note that, at least directly, the Chernoff/Hoeffding/Bernstein family of inequalities you may be familiar with is not useful here. Indeed, the ϵ-net approach above is essentially a device for reducing concentration of the non-linear function $\|\boldsymbol{W}\|$ of $\boldsymbol{W}$ to concentration of the many linear functions $\langle \boldsymbol{W}, \boldsymbol{v}\boldsymbol{v}^\top \rangle$. In our case, this was possible because the convex function $\|\boldsymbol{W}\|$ is a supremum of many linear functions. It is possible to some extent to generalize this idea, but we will instead

examine more direct manifestations of the above for very general f, on which we place nearly no structural assumptions.

Principle 1 above suggests that, if we want to apply the general ideas of concentration inequalities to random matrices, we may develop separate tools for understanding the typical values or expectations of things like random matrix norms and the concentration of these quantities. We will see that the latter task is in some sense easier, in that it does not really have to do with the matrix structure of our questions. Instead we will consider $x_1, \dots, x_N$ random variables, usually independent, and $f(x_1, \dots, x_N) \in \mathbb{R}$ a function. If for instance we want to discuss properties of a Wigner matrix, we would take $N = \frac{d(d-1)}{2} + d = \frac{d(d+1)}{2}$ the number of scalar degrees of freedom of a symmetric matrix.

5.3 General-Purpose Variance Bounds

The simplest kinds of concentration inequalities bound the *variance* of a function of many random variables $\mathsf{Var}[f(x_1, \dots, x_N)]$. By Chebyshev's inequality, such bounds translate into concentration inequalities of the form

$$\mathbb{P}[|f(\boldsymbol{x}) - \mathbb{E}f(\boldsymbol{x})| \geq s] \leq \frac{\mathsf{Var}[f(\boldsymbol{x})]}{s^2}.$$

These are fundamentally limited by the rate $O(1/s^2)$ on the right-hand side: often this rate of decay with s is simply very far from the truth, and thus variance bounds do not suffice to characterize the actual tails of $f(\boldsymbol{x})$. Still, they can already be very useful, and are often easier to prove than stronger bounds, as we will see below.

It is instructive to recall what happens for f the sum function. Suppose that x_i are independent. We then have the elementary identity

$$\mathsf{Var}[f(\boldsymbol{x})] = \mathsf{Var}\left[\sum_{i=1}^N x_i\right] = \sum_{i=1}^N \mathsf{Var}[x_i]. \tag{5.3.1}$$

Thus the variance is controlled by a sum of scalar variances.

5.3.1 Efron-Stein and Bounded Differences

The fundamental result about concentration of measure at the level of the variance is that a simple variant of the same is true even for non-linear f. Let us adopt the notation $\boldsymbol{x}_{\sim i} := (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N) \in \mathbb{R}^{N-1}$ for a vector with the ith coordinate removed.

The following elementary observation will be useful for interpreting our result.

Proposition 5.3.1. *Let X be a random variable and Y be an independent copy of X. Then, $\mathsf{Var}[X] = \frac{1}{2}\mathbb{E}(X-Y)^2$.*

Proof. Expanding gives $\mathbb{E}(X-Y)^2 = \mathbb{E}X^2 + \mathbb{E}Y^2 - 2\mathbb{E}XY = 2\mathbb{E}X^2 - 2(\mathbb{E}X)^2$. □

Theorem 5.3.2 (Efron-Stein)**.** *For any $\boldsymbol{x} = (x_1, \dots, x_N)$ independent and $f : \mathbb{R}^N \to \mathbb{R}$,*

$$\mathsf{Var}[f(X_1, \dots, X_N)] \leq \sum_{i=1}^N \mathop{\mathbb{E}}_{\boldsymbol{x}_{\sim i}} \mathop{\mathsf{Var}}_{x_i}[f(\boldsymbol{x})],$$

where, more explicitly,

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_{\sim i}} \mathop{\mathsf{Var}}_{x_i}[f(\boldsymbol{x})] = \mathop{\mathbb{E}}_{x_1,\dots,x_{i-1},x_{i+1},\dots,x_N} \mathop{\mathbb{E}}_{x_i} (f(x_1,\dots,x_N) - \mathop{\mathbb{E}}_{x_i} f(x_1,\dots,x_N))^2,$$

with all coordinates but x_i being constant in the innermost expectation. Alternatively, by Proposition 5.3.1, this is equivalent to

$$\begin{aligned}&\mathsf{Var}[f(X_1,\dots,X_N)]\\ &\quad\le \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{y}} \sum_{i=1}^{N} (f(x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_N) - f(x_1,\dots,x_{i-1},y_i,x_{i+1},\dots,x_N))^2,\end{aligned}$$

where $\boldsymbol{x},\boldsymbol{y}$ are independent copies.

This is similar in spirit to (5.3.1): the variance of a general non-linear function of independent random variables is controlled by a sum of variance-like quantities describing the "variance in the ith coordinate" of f. Also, the inner terms in the second expression above may be viewed as the effect on the value of f of *resampling* the ith coordinate, which is a natural measurement of the sensitivity of f in the ith coordinate in the sense of our Principle 2.

Proof. The proof is quite simple using some basic ideas about martingales and conditional expectations, but let us give that proof in perhaps friendlier language. In general, for $f : \mathbb{R}^N \to \mathbb{R}$, write $\mathbb{E}_i f$ for the average over the ith coordinate with respect to the law of x_i:

$$(\mathbb{E}_i f)(\boldsymbol{x}) = \mathbb{E}_{x_i} f(x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_N).$$

This is a function only depending on the coordinates other than i of f. By Fubini's theorem, these operations commute with one another, and so we may also write $\mathbb{E}_S := \mathbb{E}_{i_1} \cdots \mathbb{E}_{i_k}$ for the operation averaging over a set of indices $S = \{i_1,\dots,i_k\}$.

Now, we may decompose

$$f = (f - \mathbb{E}_{\{1\}} f) + (\mathbb{E}_{\{1\}} f - \mathbb{E}_{\{1,2\}} f) + \cdots + (\mathbb{E}_{\{1,\dots,N-1\}} f - \mathbb{E}_{\{1,\dots,N\}} f) + \mathbb{E}_{\{1,\dots,N\}} f.$$

Since $\mathbb{E}_{\{1,\dots,N\}} f = \mathbb{E} f(\boldsymbol{x})$, we have

$$f - \mathbb{E} f(\boldsymbol{x}) = (f - \mathbb{E}_{\{1\}} f) + (\mathbb{E}_{\{1\}} f - \mathbb{E}_{\{1,2\}} f) + \cdots + (\mathbb{E}_{\{1,\dots,N-1\}} f - \mathbb{E}_{\{1,\dots,N\}} f).$$

Note that, for instance,

$$\begin{aligned}&\mathbb{E}(f(\boldsymbol{x}) - \mathbb{E}_{\{1\}} f(\boldsymbol{x}))(\mathbb{E}_{\{1\}} f(\boldsymbol{x}) - \mathbb{E}_{\{1,2\}} f(\boldsymbol{x}))\\ &\quad= \mathbb{E} f(\boldsymbol{x}) \mathbb{E}_{\{1\}} f(\boldsymbol{x}) - \mathbb{E} f(\boldsymbol{x}) \mathbb{E}_{\{1,2\}} f(\boldsymbol{x})\\ &\qquad\quad - \mathbb{E}\mathbb{E}_{\{1\}} f(\boldsymbol{x}) \mathbb{E}_{\{1\}} f(\boldsymbol{x}) + \mathbb{E}\mathbb{E}_{\{1\}} f(\boldsymbol{x}) \mathbb{E}_{\{1,2\}} f(\boldsymbol{x})\\ &\quad= \mathbb{E}\mathbb{E}_{\{1\}} f(\boldsymbol{x}) \mathbb{E}_{\{1\}} f(\boldsymbol{x}) - \mathbb{E}\mathbb{E}_{\{1,2\}} f(\boldsymbol{x}) \mathbb{E}_{\{1,2\}} f(\boldsymbol{x})\\ &\qquad\quad - \mathbb{E}\mathbb{E}_{\{1\}} f(\boldsymbol{x}) \mathbb{E}_{\{1\}} f(\boldsymbol{x}) + \mathbb{E}\mathbb{E}_{\{1,2\}} f(\boldsymbol{x}) \mathbb{E}_{\{1,2\}} f(\boldsymbol{x})\\ &\quad= 0,\end{aligned}$$

and similarly all cross-terms like this cancel. Thus,

$$\begin{aligned}\mathsf{Var}[f(\boldsymbol{x})] &= \mathbb{E}(f(\boldsymbol{x}) - \mathbb{E}f(\boldsymbol{x}))^2 \\ &= \mathbb{E}(f - \mathbb{E}_{\{1\}}f)^2 + \cdots + \mathbb{E}(\mathbb{E}_{\{1,\ldots,N-1\}}f - \mathbb{E}_{\{1,\ldots,N\}}f)^2.\end{aligned}$$

Finally, considering e.g. the last term we have

$$\mathbb{E}(\mathbb{E}_{\{1,\ldots,N-1\}}f - \mathbb{E}_{\{1,\ldots,N\}}f)^2 = \mathbb{E}(\mathbb{E}_{\{1,\ldots,N-1\}}(f - \mathbb{E}_{\{N\}}f))^2 \leq \mathbb{E}(f - \mathbb{E}_{\{N\}}f)^2$$

by Jensen's inequality, and working similarly on each term gives the result. □

We derive a simple consequence, which is like a version for the variance of the *bounded differences inequality* that you may have seen before.

Proposition 5.3.3. *If $X \in [a,b]$ almost surely, then $\mathsf{Var}[X] \leq \frac{1}{4}(b-a)^2$.*

Proof. You may show as a simple calculus exercise that $\mathsf{Var}[X] = \mathbb{E}(X - \mathbb{E}X)^2 = \min_c \mathbb{E}(X - c)^2$, with the minimum achieved at $c = \mathbb{E}X$. Taking $c = \frac{a+b}{2}$ then gives the result. □

Definition 5.3.4 (Maximum coordinate influences)**.** *Suppose $f : \Sigma^N \to \mathbb{R}$. For a given $\boldsymbol{x} \in \Sigma^N$, the* maximum influence of coordinate i *on f at $\boldsymbol{x}$ is*

$$D_i f(\boldsymbol{x}) := \sup_{y \in \Sigma} f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_N) - \inf_{y \in \Sigma} f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_N) \geq 0.$$

We also write

$$\Delta_i f := \sup_{\boldsymbol{x} \in \Sigma^N} \{D_i f(\boldsymbol{x})\},$$

and write $\boldsymbol{D}f(\boldsymbol{x}) := (D_1 f(\boldsymbol{x}), \cdots, D_N f(\boldsymbol{x}))$ and $\boldsymbol{\Delta} f := (\Delta_1 f, \ldots, \Delta_N f)$.

Corollary 5.3.5. *Let $f : \Sigma^N \to \mathbb{R}$. Then, in the setting of Theorem 5.3.2, we have*

$$\mathsf{Var}[f(x_1, \ldots, x_N)] \leq \frac{1}{4}\sum_{i=1}^{N} \mathbb{E}(D_i f(\boldsymbol{x}))^2 = \frac{1}{4}\mathop{\mathbb{E}}_{\boldsymbol{x}} \|\boldsymbol{D}f(\boldsymbol{x})\|^2.$$

In particular, if $|D_i f(\boldsymbol{x})| \leq \Delta_i$ for some constant Δ_i for all $\boldsymbol{x} \in \mathbb{R}^N$, then

$$\mathsf{Var}[f(x_1, \ldots, x_N)] \leq \frac{1}{4}\sum_{i=1}^{N} \Delta_i^2 = \frac{1}{4}\|\boldsymbol{\Delta} f\|^2.$$

This is another clear manifestation of Principle 2: the variance of any, perhaps very non-linear, function of independent random variables is controlled by the maximum that that function may be affected by changing each coordinate.

Remark 5.3.6. *In fact, all of this goes through just as well for $f : \Sigma_1 \times \cdots \times \Sigma_N \to \mathbb{R}$ for any measurable spaces Σ_i, provided that $x_i \in \Sigma_i$ are independent.*

5.3.2 Example: Variance of the Chromatic Number

The *chromatic number* of a graph G, denoted $\chi(G)$, is the minimum number of colors required to color its vertices so that any two adjacent vertices have different colors. Consider the Erdős-Rényi random graph G on d vertices, viewed as a $N = \binom{d}{2}$-dimensional binary random vector. It is easy to check that $D_i\chi(G) \le 1$ for all G: changing any one edge changes the number of colors needed by at most 1, since one vertex may always be given a new color to generate a valid coloring for the new graph. Unfortunately, this only shows $\mathsf{Var}[\chi(G)] \lesssim d^2$, a poor bound since $\chi(G) \le d$ always anyway.

To sharpen this, we can use Remark 5.3.6 and group some of the coordinates together: let X_1 contain all $d-1$ edges incident with vertex 1, X_2 contain all $d-2$ edges incident with vertex 2 but not vertex 1, and so forth until X_{d-1} contains just the outcome of the one edge between vertex $d-1$ and vertex d. We may view G as this $(d-1)$-dimensional coarse-grained random vector $(X_1, \dots, X_{d-1})$. By the same argument as above, $\chi(G)$ changes by at most 1 from a modification of any X_i. Thus, we immediately learn that $\mathsf{Var}[\chi(G)] \lesssim d$. Other combinatorial reasoning shows $\mathbb{E}[\chi(G)] \sim d/\log d$ (though, in accordance with Principle 1, we did not need to know this to establish our variance bound), so this indeed gives strong concentration of order $O(\sqrt{d})$ around this value.

The Efron-Stein family of results is a powerful tool in general for many optimization problems over graphs, because of this kind of limited sensitivity to individual edges or vertex neighborhoods.

5.3.3 Variance of λ_1 for Bounded Entries

Let us now see a random matrix application. We will show the following remarkable fact, in whose setting we see the advantage of working non-asymptotically: we can treat very general classes of random matrices that do not necessarily form a nice sequence approaching a legible limit.

Theorem 5.3.7. *Let $\boldsymbol{W} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ have independent entries with arbitrary distribution, such that $|W_{ij}| \le K$ almost surely for all $1 \le i \le j \le d$. Then, $\mathsf{Var}[\lambda_1(\boldsymbol{W})] \le 16K^2$.*

Note the startling generality: the result implies, for instance, that the largest eigenvalue of *any* random graph with independent edge values has variance at most 16; moreover, this is an exact bound with no error term in d. Moreover, the entries need not be centered, so we may apply this to models with a "signal" in their expectation like the stochastic block models we saw earlier.

The proof uses a useful trick to modify Corollary 5.3.5 in a way that is often useful, which we state separately first.

Definition 5.3.8 (Upper maximum coordinate influence)**.** *In the context of Definition 5.3.4, we also define*

$$D_i^+ f(\boldsymbol{x}) := \sup_{y\in\Sigma} f(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_N) - f(\boldsymbol{x}),$$

and write $\boldsymbol{D}^+ f(\boldsymbol{x}) := (D_1 f(\boldsymbol{x}), \dots, D_N f(\boldsymbol{x}))$.

Corollary 5.3.9. *In the context of Corollary 5.3.5, we also have*

$$\mathsf{Var}[f(\boldsymbol{x})] \le \sum_{i=1}^{N} \mathbb{E}(D_i^+ f(\boldsymbol{x}))^2 = \frac{1}{2}\|\boldsymbol{D}^+ f(\boldsymbol{x})\|^2.$$

Proof. Using our second form of the Efron-Stein inequality, writing $\boldsymbol{x}^{(i)}$ for the vector $\boldsymbol{x}$ with the ith coordinate replaced by y_i, we have

$$\mathsf{Var}[f(x_1, \ldots, X_N)] \le \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{y}} \sum_{i=1}^{N} (f(\boldsymbol{x}) - f(\boldsymbol{x}^{(i)}))^2,$$

where the laws of the random variables $f(\boldsymbol{x}) - f(\boldsymbol{x}^{(i)})$ are symmetric, whereby

$$\begin{aligned}
&= \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{y}} \sum_{i=1}^{N} (\max\{0, f(\boldsymbol{x}) - f(\boldsymbol{x}^{(i)})\} + \min\{0, f(\boldsymbol{x}) - f(\boldsymbol{x}^{(i)})\})^2 \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{y}} \sum_{i=1}^{N} (\max\{0, f(\boldsymbol{x}) - f(\boldsymbol{x}^{(i)})\})^2 \\
&\le \mathbb{E} \sum_{i=1}^{N} (D_i^+ f(\boldsymbol{x}))^2,
\end{aligned}$$

giving the result. □

Proof of Theorem 5.3.7. It suffices to control the $D_{ij}^+ \lambda_1(\boldsymbol{W})$, i.e., the upper maximum coordinate influences of each entry of a matrix on the largest eigenvalue. Let $\boldsymbol{W}$ and $\boldsymbol{W}'$ be two matrices that only differ in coordinate i, j and have all coordinates bounded by K, and let $\boldsymbol{v}_1$ be the top eigenvector of $\boldsymbol{W}$. We then have

$$\begin{aligned}
\lambda_1(\boldsymbol{W}) - \lambda_1(\boldsymbol{W}') &= \boldsymbol{v}_1^\top \boldsymbol{W} \boldsymbol{v}_1 - \max_{\|\boldsymbol{v}\|=1} \boldsymbol{v}^\top \boldsymbol{W}' \boldsymbol{v} \\
&\le \boldsymbol{v}_1^\top \boldsymbol{W} \boldsymbol{v}_1 - \boldsymbol{v}_1^\top \boldsymbol{W}' \boldsymbol{v}_1 \\
&= (1 + \mathbb{1}\{i = j\})(W_{ij} - W'_{ij}) v_{1i} v_{1j} \\
&\le 4K \cdot |v_{1i}| \cdot |v_{1j}|.
\end{aligned}$$

Thus, we have

$$D_{ij}^+ \lambda_1(\boldsymbol{W}) \le 4K \cdot |v_{1i}| \cdot |v_{1j}|.$$

Applying the Corollary above,

$$\mathsf{Var}[\lambda_1(\boldsymbol{W})] \le 16K^2 \mathbb{E} \sum_{1 \le i \le j \le d} |v_{1i}|^2 \cdot |v_{1j}|^2 \le 16K^2,$$

since $\sum_{i=1}^{d} |v_{1i}|^2 = 1$. □

5.3.4 Poincaré Inequalities

One annoyance of the previous results is that they strongly depend on the boundedness of the entries of $\boldsymbol{W}$. The theory of *Poincaré inequalities* gives a vast generalization of Efron-Stein types of results to many other distributions. Here, let us give one example that is accessible with the tools we have seen so far, which will at least let us treat Gaussian models, including the GOE.

Theorem 5.3.10 (Gaussian Poincaré inequality)**.** *For all C^1 functions $f : \mathbb{R}^N \to \mathbb{R}$,*

$$\operatorname*{Var}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_N)}[f(\boldsymbol{x})] \le \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_N)} \|\nabla f(\boldsymbol{x})\|^2.$$

A few remarks are in order. Note first the similarity to the version from Corollary 5.3.5 earlier, with the somewhat "harsher" notion of gradient $\boldsymbol{D}f(\boldsymbol{x})$ that measured the maximum possible effect on the value of f by changing each coordinate arbitrarily:

$$\mathsf{Var}[f(\boldsymbol{x})] \lesssim \mathop{\mathbb{E}}_{\boldsymbol{x}} \|\boldsymbol{D}f(\boldsymbol{x})\|^2.$$

In exchange for the more specific assumption of Gaussian inputs and a more delicate proof we will see below, we gain both a "gentler" measurement of sensitivity through the gradient and the possibility of working with unbounded distributions and values of $f(\boldsymbol{x})$. Second, there are many Poincaré inequalities, sometimes involving a different notion of gradient, and also allowing for a constant on the right-hand side that is of great interest to identify. The above result is sharp in terms of this constant, since the inequality is an equality for $N = 1$ and $f(x) = x$. This is often phrased as the *Poincaré constant* of the measure $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$ being equal to 1, the important qualitative aspect being that this value is independent of N.

Proof. Suppose that we could show the inequality for $N = 1$. Then, by Efron-Stein, we can use that result to prove the inequality for any N:

$$\begin{aligned}
\operatorname*{Var}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_N)}[f(\boldsymbol{x})] &\le \sum_{i=1}^{N} \mathop{\mathbb{E}}_{\boldsymbol{x}_{\sim i}} \operatorname*{Var}_{x_i}[f(\boldsymbol{x})] && \text{(Efron-Stein)} \\
&\le \sum_{i=1}^{N} \mathop{\mathbb{E}}_{\boldsymbol{x}_{\sim i}} \mathbb{E}_{x_i}(\partial_i f(\boldsymbol{x}))^2 && (N = 1 \text{ case}) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_N)} \|\nabla f(\boldsymbol{x})\|^2.
\end{aligned}$$

For the case $N = 1$, we use a very Gaussian-specific trick to introduce an opportunity to use Efron-Stein again. Introduce $z_1, \dots, z_k, z'_1, \dots, z'_k \sim \mathsf{Unif}(\{\pm 1\})$ i.i.d. random signs. We will be a bit heuristic with the central limit theorem below, but this argument is easy to make rigorous:

$$\begin{aligned}
\operatorname*{Var}_{x\sim\mathcal{N}(0,1)}[f(x)] &\approx \operatorname*{Var}_{z}\left[f\left(\frac{1}{\sqrt{k}}\sum_{i=1}^{k} z_i\right)\right] && \text{(CLT)} \\
&\le \frac{1}{2}\sum_{i=1}^{k} \mathop{\mathbb{E}}_{z,z'}\left(f\left(\frac{1}{\sqrt{k}}\sum_{j\neq i} z_j + \frac{z_i}{\sqrt{k}}\right) - f\left(\frac{1}{\sqrt{k}}\sum_{j\neq i} z_j + \frac{z'_i}{\sqrt{k}}\right)\right)^2 && \text{(Efron-Stein)}
\end{aligned}$$

Here, by just enumerating the outcomes of the pair (z_i, z_i') we see that

$$\begin{aligned}
&= \frac{1}{4}\sum_{i=1}^{k} \mathop{\mathbb{E}}_{z}\left(f\left(\frac{1}{\sqrt{k}}\sum_{j\neq i} z_j + \frac{1}{\sqrt{k}}\right) - f\left(\frac{1}{\sqrt{k}}\sum_{j\neq i} z_j - \frac{1}{\sqrt{k}}\right)\right)^2 \\
&\approx \frac{1}{4}\sum_{i=1}^{k} \mathop{\mathbb{E}}_{z}\left(\frac{2}{\sqrt{k}} f'\left(\frac{1}{\sqrt{k}}\sum_{j\neq i} z_j\right)\right)^2 && \text{(Taylor expansion)} \\
&= \frac{1}{k}\sum_{i=1}^{k} \mathop{\mathbb{E}}_{z} f'\left(\frac{1}{\sqrt{k}}\sum_{j\neq i} z_j\right)^2 \\
&\approx \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)} f'(x)^2, && \text{(CLT)}
\end{aligned}$$

completing the proof. □

The first part of the proof is often referred to as the *tensorization* of the Poincaré inequality: if it holds with the same constant over several probability measures, then it also holds with that constant over their product.

The following minor variant is easy to derive using that Lipschitz functions are almost everywhere differentiable.

Corollary 5.3.11. *Suppose that $f : \mathbb{R}^N \to \mathbb{R}$ is L-*Lipschitz*, meaning that $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \le L\|\boldsymbol{x} - \boldsymbol{y}\|$. Then,*

$$\mathop{\mathsf{Var}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_N)}[f(\boldsymbol{x})] \le L^2.$$

5.3.5 Variance of λ_1 for Gaussian Entries

We may now easily derive the following Gaussian-valued version of Theorem 5.3.7.

Theorem 5.3.12. *Let $\boldsymbol{W} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ have independent entries $W_{ij} = W_{ji} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$. Then,* $\mathsf{Var}[\lambda_1(\boldsymbol{W})] \le 2\max_{ij}\sigma_{ij}^2$.

Again, the result is remarkably general, applying uniformly to the GOE, to spiked matrix models, or to *heterogeneous* or *heteroskedastic* versions of those with different entry variances.

Proof. Let $\boldsymbol{g} \in \mathbb{R}^{d(d+1)/2}$ be a standard Gaussian vector, with entries indexed by pairs (i, j) with $1 \le i \le d$. We may view $\boldsymbol{W} = \boldsymbol{W}(\boldsymbol{g})$ as having entries $W_{ij} = \mu_{ij} + \sigma_{ij}g_{ij}$ for all $i \le j$. We are interested in this language in concentration of the function $f(\boldsymbol{g}) := \lambda_1(\boldsymbol{W}(\boldsymbol{g}))$. This satisfies:

$$|f(\boldsymbol{g}) - f(\boldsymbol{h})| = |\lambda_1(\boldsymbol{W}(\boldsymbol{g})) - \lambda_1(\boldsymbol{W}(\boldsymbol{h}))| \le \|\boldsymbol{W}(\boldsymbol{g}) - \boldsymbol{W}(\boldsymbol{h})\|_F \le \sqrt{2}\max_{i,j}\sigma_{ij}\|\boldsymbol{g} - \boldsymbol{h}\|,$$

with the $\sqrt{2}$ coming from each entry of $\boldsymbol{g}$ happening either once or twice in $\boldsymbol{W}(\boldsymbol{g})$. The result then follows from Corollary 5.3.11. □

The following is an simple corollary pertaining to scalar-valued probability.

Corollary 5.3.13. *Let $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_N)$. Then,* $\mathsf{Var}[\max_{i=1}^N x_i] \leq 2$.

Proof. Apply Theorem 5.3.12 to a diagonal matrix. □

This is a non-trivial concentration result, since, as you have probably seen,

$$\mathbb{E} \max_{i=1}^N x_i \sim \sqrt{2 \log N}.$$

5.4 General-Purpose Subgaussian Tail Bounds

Bounding the variance, say as $\mathsf{Var}[X] \leq \sigma^2$, can only give us tail bounds via Chebyshev's inequality of the form

$$\mathbb{P}[|X - \mathbb{E}X| > s] \lesssim \frac{1}{s^2}.$$

This rate of decay is often severely suboptimal; even as we saw with our epsilon net calculation earlier, what we really should hope for is *subgaussian* rates of the form

$$\mathbb{P}[|X - \mathbb{E}X| > s] \lesssim \exp\left(-\frac{s^2}{2\sigma^2}\right), \tag{5.4.1}$$

where the scaling with s is the same as the tails of $\mathcal{N}(0, \sigma^2)$. We will now see some machinery for proving such results, which are very common and useful in modern probability theory, even beyond random matrices.

5.4.1 Subgaussianity via Martingales

Recall that we saw in Corollary 5.3.5 of the Efron-Stein inequality that, if the influence on $f(\boldsymbol{x})$ of coordinate i is bounded as $|D_i f(\boldsymbol{x})| \leq \Delta_i$ then, whenever the $\boldsymbol{x}$ are independent random variables,

$$\mathsf{Var}[f(\boldsymbol{x})] \leq \frac{1}{4} \sum_{i=1}^N \Delta_i^2.$$

In fact, it turns out that this may be directly boosted to a subgaussian rate of tail decay.

Theorem 5.4.1 (McDiarmid)**.** *Under the above assumption,*

$$\mathbb{P}[|f(\boldsymbol{x}) - \mathbb{E}f(\boldsymbol{x})| > s] \leq 2 \exp\left(-\frac{s^2}{2 \cdot \frac{1}{4}\sum_{i=1}^N \Delta_i^2}\right),$$

where we use this form to emphasize that this amounts to taking $\sigma^2 = \frac{1}{4}\sum_{i=1}^N \Delta_i^2$ in (5.4.1), i.e., plugging our variance bound into the role of the variance parameter in the Gaussian tail bound.

Proof Sketch. We just give the main idea and some remarks. In fact, the idea is similar to the proof of the Efron-Stein inequality: we decompose

$$f = (f - \mathbb{E}_{\{1\}} f) + (\mathbb{E}_{\{1\}} f - \mathbb{E}_{\{1,2\}} f) + \cdots + (\mathbb{E}_{\{1,\ldots,N-1\}} f - \mathbb{E}_{\{1,\ldots,N\}} f) + \mathbb{E}_{\{1,\ldots,N\}} f,$$

and note that the sequence of partial sums starting from the end of this sum

$$\mathbb{E}_{\{1,\dots,N\}} f(\boldsymbol{x}),\ \mathbb{E}_{\{1,\dots,N-1\}} f(\boldsymbol{x}),\ \cdots,\ \mathbb{E}_{\{1\}} f(\boldsymbol{x}),\ f(\boldsymbol{x})$$

form a *martingale*, the so-called *Doob martingale* associated to the function f. One way to think of these is as a sequence of "forecasts" of the function $f(\boldsymbol{x})$ of unknown inputs, which is increasingly refined as we are given the outcomes of each coordinate.

A general philosophy in working with martingales is that they resemble random walks—many inequalities that hold for random walks also hold for martingales. The assumption of McDiarmid's inequality is that the increments of the martingale—analogous to the steps of the random walk—are bounded. Thus McDiarmid's inequality is a martingale analog of Hoeffding's inequality (really the general analog for arbitrary martingales is often called the *Azuma* or *Azuma-Hoeffding inequality*, of which McDiarmid's inequality is a consequence when applied to the special case of a Doob martingale). The proof of such results usually amounts to imitating the proof for random walks, with suitable adjustments. Here the proof is by a Chernoff bound: say for the upper tail,

$$\begin{aligned}
\mathbb{P}[f - \mathbb{E}f > s] &= \mathbb{P}[\exp(\lambda(f - \mathbb{E}f)) > \exp(\lambda s)] \\
&\le \exp(-\lambda s)\mathbb{E}\exp(\lambda(f - \mathbb{E}f)) \\
&= \exp(-\lambda s)\mathbb{E}_{x_N}\exp(\lambda(\mathbb{E}_{\le N-1}f - \mathbb{E}_{\le N}f)) \\
&\qquad \mathbb{E}_{x_{N-1}}\exp(\lambda(\mathbb{E}_{\le N-2}f - \mathbb{E}_{\le N-1}f)) \\
&\qquad \vdots \\
&\qquad \mathbb{E}_{x_1}\exp(\lambda(f - \mathbb{E}_{\le 1}f)),
\end{aligned}$$

and now from the innermost expectation (over x_1) to the outermost (over x_N) we iteratively apply the same reasoning as in Hoeffding's inequality. □

This inequality is useful, but has a few shortcomings. First, it cannot give us an analog of our main corollary of Efron-Stein, which was in terms of the coordinatewise influences in Corollary 5.3.5, stating that

$$\mathrm{Var}[f(\boldsymbol{x})] \lesssim \mathop{\mathbb{E}}_{\boldsymbol{x}} \|\boldsymbol{D}f(\boldsymbol{x})\|^2.$$

For direct comparison, the bounded differences inequality only tells us that we may have a subgaussian tail bound with the weaker parameter

$$\sigma^2 \lesssim \sum_{i=1}^{N} \left(\sup_{\boldsymbol{x}} |(D_i f(\boldsymbol{x}))|\right)^2.$$

The reason that this can be less useful should be intuitive: the former inequality is capturing how big $\|\boldsymbol{D}f(\boldsymbol{x})\|^2$ can be at any given $\boldsymbol{x}$ (even if we ignore the expectation and take a supremum), while the latter is capturing only how big *each coordinate* of $\boldsymbol{D}f(\boldsymbol{x})$ can be, not taking advantage of the fact that the coordinates might never all be large at once.

Recall in particular that our application of bounding the variance of $\lambda_1(\boldsymbol{W})$ for $\boldsymbol{W}$ with independent bounded entries in Section 5.3.3 crucially used the stronger variance bound, relating the coordinates of $\boldsymbol{D}\lambda_1(\boldsymbol{W})$ to the entries of the top eigenvector of $\boldsymbol{W}$. Accordingly, we will not get tight subgaussian tail bounds on random matrices from just McDiarmid's inequality. And, as in our passage from the direct application of Efron-Stein in Corollary 5.3.5, we would like to be able to work with unbounded distributions and functions as well.

5.4.2 Non-Tensorization of Subgaussianity

Given this situation, it is natural to try to imitate the Efron-Stein inequality, but controlling instead of the variance some quantity that corresponds to subgaussian tail bounds. The following are important definitions related to this matter.

Definition 5.4.2. *For a random variable X, we define the* moment *and* cumulant generating functions *as*

$$\phi_X(\lambda) := \mathbb{E}\exp(\lambda(X - \mathbb{E}X)),$$
$$\psi_X(\lambda) := \log \phi_X(\lambda).$$

*We say that X is σ^2-*subgaussian *$\psi_X(\lambda) \leq \frac{1}{2}\sigma^2\lambda^2$ for all $\lambda \in \mathbb{R}$.*

Proposition 5.4.3. *If X is σ^2-subgaussian, then it admits the tail bound*

$$\mathbb{P}[|X - \mathbb{E}X| > s] \leq 2\exp\left(-\frac{s^2}{2\sigma^2}\right).$$

Proof. The same as the proof of Chernoff's inequality. □

The key use to us of the Efron-Stein inequality was that it establishes the *tensorization of variance*: the variance of a non-linear function of independent random variables is controlled by individual variances of that function in each coordinate at a time. And, the inequality was tight for $f(\boldsymbol{x})$ a linear function, since the variances of independent random variables add:

$$\mathsf{Var}\left[\sum_{i=1}^N x_i\right] = \sum_{i=1}^N \mathsf{Var}[x_i].$$

Note that the cumulant generating function enjoys the same property:

$$\psi_{\sum_{i=1}^N x_i}(\lambda) = \sum_{i=1}^N \psi_{x_i}(\lambda).$$

In particular, this implies that subgaussianity is linear over independent random variables: if $x_1, \ldots, x_N$ are independent and x_i is σ_i^2-subgaussian, then $\sum_{i=1}^N x_i$ is $(\sum_{i=1}^N \sigma_i^2)$-subgaussian.

It is then very tempting to try to extend this to non-linear functions as the Efron-Stein inequality did for variances: we might expect that, for x_i independent,

$$\mathbb{E}\log\exp(\lambda(f(\boldsymbol{x}) - \mathbb{E}f(\boldsymbol{x}))) \overset{(?)}{\lesssim} \sum_{i=1}^N \mathop{\mathbb{E}}_{\boldsymbol{x}_{\sim i}} \log \mathop{\mathbb{E}}_{x_i} \exp(\lambda(f(\boldsymbol{x}) - \mathbb{E}f(\boldsymbol{x}))).$$

Such an inequality would mean that, if $f(\boldsymbol{x})$ is "subgaussian in each coordinate," then it is also subgaussian as a function of all of its random inputs at once. Sadly, such an inequality cannot hold.

Example 5.4.4. *Let $x_1, x_2, x_3 \sim \mathcal{N}(0,1)$ independently, and let $f(\boldsymbol{x}) = x_1 x_2 x_3$. On the right-hand side of our putative inequality, we have terms like*

$$\begin{aligned}\mathop{\mathbb{E}}_{x_2,x_3} \log \mathop{\mathbb{E}}_{x_1} \exp(\lambda x_1 x_2 x_3) &= \mathop{\mathbb{E}}_{x_2,x_3} \log \exp\left(\frac{1}{2}\lambda^2 x_2^2 x_3^2\right) \\ &= \frac{1}{2}\lambda^2 \mathop{\mathbb{E}}_{x_2,x_3} x_2^2 x_3^2 \\ &= \frac{1}{2}\lambda^2.\end{aligned}$$

Thus, f is 1-subgaussian in each coordinate, so we would expect $f(\boldsymbol{x})$ to be 3-subgaussian, or at least C-subgaussian for some $C > 0$. Yet this is not true: the left-hand side above is

$$\begin{aligned}\log \mathop{\mathbb{E}}_{x_1,x_2,x_3} \exp(\lambda x_1 x_2 x_3) &= \log \mathop{\mathbb{E}}_{x_2,x_3} \exp\left(\frac{1}{2}\lambda^2 x_2^2 x_3^2\right) \\ &= \log \mathop{\mathbb{E}}_{x_3} \frac{1}{\sqrt{1-\lambda^2 x_3^2}} \\ &= \infty,\end{aligned}$$

showing that the left-hand side expectation does not converge for any $\lambda \in \mathbb{R}$.

5.4.3 Logarithmic Sobolev Inequalities

For this reason, the issue of conveniently establishing subgaussian tail bounds for large classes of functions of independent random variables in high dimension is subtler than the same question for the variance. Fortunately, it is still possible through the technology of *logarithmic Sobolev inequalities (LSI)*, which we briefly sketch now.

These inequalities are based on the following substitute for the variance, called the *entropy* (and related to some extent to the Shannon and relative entropies, at least for discrete random variables).

Definition 5.4.5. *For $X > 0$, define* $\mathsf{Ent}[X] := \mathbb{E}[X \log X] - \mathbb{E}[X] \log \mathbb{E}[X]$.

Note that $\mathsf{Ent}[X]$ is the "gap" in the Jensen inequality for the convex function $f(x) = x \log x$ applied to the random variable X, just like $\mathsf{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ is the same kind of gap for $f(x) = x^2$. Thus for instance $\mathsf{Ent}[X] \geq 0$ with equality if and only if X is constant, and $\mathsf{Ent}[X]$ is another way of measuring the "spreadness" of a random variable's law.

The first key property of this quantity is that it satisfies an inequality precisely in the style of Efron-Stein.

Lemma 5.4.6. *For any $\boldsymbol{x} = (x_1, \ldots, x_N)$ independent and $f : \mathbb{R}^N \to \mathbb{R}$,*

$$\mathsf{Ent}[f(x_1, \ldots, x_N)] \leq \sum_{i=1}^{N} \mathop{\mathbb{E}}_{\boldsymbol{x}_{\sim i}} \mathop{\mathsf{Ent}}_{x_i}[f(\boldsymbol{x})].$$

It should not be clear what the relationship is between the entropy and subgaussianity, since entropy is an even milder measurement of spreadness than the variance. The key

observation is that the entropy appears naturally in the *derivative* of the moment generating function:

$$\phi_X'(\lambda) = \mathbb{E}X\exp(\lambda X) = \frac{1}{\lambda}\mathsf{Ent}[\exp(\lambda X)] + \frac{1}{\lambda}\phi_X(\lambda)\log\phi_X(\lambda).$$

Since subgaussianity asks for a bound on the moment generating function, it is natural that a bound on $\mathsf{Ent}[\exp(\lambda X)]$ should suffice. The following expresses what is needed.

Lemma 5.4.7 (Herbst)**.** *Suppose that, for all* $\lambda \in \mathbb{R}$, $\mathsf{Ent}[\exp(\lambda X)] \le \frac{1}{2}\sigma^2\lambda^2\mathbb{E}[\exp(\lambda X)]$*. Then,* X *is* σ^2*-subgaussian.*

Finally, we would like some family of inequalities like the Poincaré inequalites for the variance: inequalities that on the one hand tensorize, in the sense of generalizing easily to product measures, and on the other hand that control the subgaussianity of a wide range of non-linear functions of a random variable or vector.

Theorem 5.4.8 (Gaussian modified log-Sobolev inequality (MLSI))**.** *Let* $f : \mathbb{R}^N \to \mathbb{R}$ *be* C^1*. Then,*

$$\operatorname*{Ent}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_N)}[\exp(f(\boldsymbol{x}))] \le \frac{1}{2}\operatorname*{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_N)}\|\nabla f(\boldsymbol{x})\|^2\exp(f(\boldsymbol{x})).$$

The proof of the $N = 1$ case is beyond the scope of our discussion, but it is easy to show how to generalize to larger N using Lemma 5.4.6: we simply have, using the Lemma and then the $N = 1$ result,

$$\begin{aligned}\mathsf{Ent}[\exp(f(\boldsymbol{x}))] &\le \sum_{i=1}^N \operatorname*{\mathbb{E}}_{\boldsymbol{x}_{\sim i}}\operatorname*{\mathsf{Ent}}_{x_i}[f(\boldsymbol{x})] \\ &\le \frac{1}{2}\sum_{i=1}^N \mathbb{E}(\partial_i f(\boldsymbol{x}))^2\exp(f(\boldsymbol{x})) \\ &= \frac{1}{2}\mathbb{E}\|\nabla f(\boldsymbol{x})\|^2\exp(f(\boldsymbol{x})).\end{aligned}$$

This reflects the more general phenomenon of MLSI's *tensorizing*, just like Poincaré inequalities.

Finally, let us see how the MLSI implies concentration inequalities. The following is a crucially important theorem of modern probability, perhaps the most ubiquitous strong concentration inequality you will come across.

Theorem 5.4.9 (Gaussian Lipschitz concentration)**.** *Suppose* f *is* L*-Lipschitz and* $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N)$*. Then,* $f(\boldsymbol{x})$ *is* L^2*-subgaussian, whereby*

$$\mathbb{P}[|f(\boldsymbol{x}) - \mathbb{E}f(\boldsymbol{x})| > s] \le \exp\left(-\frac{s^2}{2L^2}\right).$$

Proof. By the Gaussian MLSI applied to the function $g(\boldsymbol{x}) = \exp(\lambda f(\boldsymbol{x}))$, we have

$$\mathsf{Ent}[\exp(\lambda f(\boldsymbol{x}))] \le \frac{1}{2}\mathbb{E}\|\nabla\lambda f(\boldsymbol{x})\|^2\exp(\lambda f(\boldsymbol{x})) \le \frac{1}{2}\lambda^2L^2\mathbb{E}\exp(\lambda f(\boldsymbol{x})).$$

But, this is precisely the inequality in the condition of Lemma 5.4.7, so we find that $f(\boldsymbol{x})$ is L^2-subgaussian. □

5.4.4 Subgaussianity of λ_1 for Gaussian Entries

Let us see some applications of Theorem 5.4.9. A simple one is that the result can recover the concentration of the norm of a Gaussian random vector that we derived much more painstakingly earlier:

Corollary 5.4.10 (Gaussian vector norm)**.** *Let $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_N)$. Then, $\|\boldsymbol{x}\|$ is 1-subgaussian. Consequently,*

$$\mathop{\mathbb{P}}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_N)} \left[\left| \|\boldsymbol{x}\| - \mathbb{E}\|\boldsymbol{x}\| \right| \geq s \right] \leq 2\exp\left(-\frac{s^2}{2}\right).$$

Proof. The function $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ is 1-Lipschitz, since $|\|\boldsymbol{x}\| - \|\boldsymbol{y}\|| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$ by triangle inequality, and the result follows by Theorem 5.4.9. □

Note that Exercise 2.8.1 also showed that $\mathbb{E}\|\boldsymbol{x}\| = (1 + o(1))\sqrt{d}$, making this a complete picture of the typical value of the norm.

We may also treat arbitrary matrices with independent Gaussian entries, as we did earlier for the variance.

Corollary 5.4.11 (λ_1 of independent Gaussian matrices)**.** *Let $\boldsymbol{W} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ have independent entries $W_{ij} = W_{ji} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$. Then, $\lambda_1(\boldsymbol{W})$ is $(2\max_{1\leq i\leq j\leq d} \sigma_{ij}^2)$-subgaussian. Consequently,*

$$\mathop{\mathbb{P}}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_N)} \left[\left| \lambda_1(\boldsymbol{W}) - \mathbb{E}\lambda_1(\boldsymbol{W}) \right| \geq s \right] \leq 2\exp\left(-\frac{s^2}{4\max \sigma_{ij}^2}\right).$$

Proof. From our earlier argument for Theorem 5.3.12, we may view $\boldsymbol{W} = \boldsymbol{W}(\boldsymbol{g})$ for $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_D)$ with $D = \frac{d(d+1)}{2}$, and then $f(\boldsymbol{g}) := \lambda_1(\boldsymbol{W}(\boldsymbol{g}))$ is $(\sqrt{2}\max_{1\leq i\leq j\leq d} \sigma_{ij})$-subgaussian. The result then follows by Theorem 5.4.9. □

Note in particular that this applies both to $\boldsymbol{W} \sim \mathsf{GOE}(d)$ and to $\boldsymbol{W} + \beta\sqrt{d}\boldsymbol{x}\boldsymbol{x}^\top$ for any $\beta, \boldsymbol{x}$ as in the spiked matrix model, to show that the largest eigenvalues of these are 2-subgaussian, compared to the typical values $\lambda_1 \asymp \sqrt{d}$ we derived earlier.

5.4.5 Subgaussianity of λ_1 for Gaussian Series

We may actually treat the concentration of the largest eigenvalue of a much more general model of a Gaussian random matrix. These will also play a central role in the rest of our study of matrix concentration.

The following result gives a concentration inequality for the top eigenvalue of *any* Gaussian series, and therefore any random matrix with jointly Gaussian entries. The concentration will be governed by the following parameter.

Definition 5.4.12. *For $\boldsymbol{X} = \boldsymbol{A}_0 + \sum_{i=1}^D g_i \boldsymbol{A}_i$ a Gaussian series, we define*

$$\begin{aligned}\sigma_*(\boldsymbol{X})^2 &:= \max_{\|\boldsymbol{v}\|=1} \mathbb{E}\left(\boldsymbol{v}^\top(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})\boldsymbol{v}\right)^2 \\ &:= \max_{\|\boldsymbol{v}\|=1} \sum_{i=1}^D (\boldsymbol{v}^\top \boldsymbol{A}_i \boldsymbol{v})^2 \\ &= \max_{\|\boldsymbol{v}\|=1} \left\langle \sum_{i=1}^D \boldsymbol{A}_i^{\otimes 2}, \boldsymbol{v}^{\otimes 4} \right\rangle,\end{aligned}$$

sometimes called the weak matrix variance statistic *or* parameter*, in contrast to a similar quantity we will see below in Definition 6.1.1.*

Remark 5.4.13. *From the last form, one may also write $\sigma_*(\boldsymbol{X}) = \|\boldsymbol{T}\|_{\mathsf{inj}}$, the* injective norm *of the symmetric 4-tensor $\boldsymbol{T}$ formed by symmetrizing $\sum_{i=1}^D \boldsymbol{A}_i^{\otimes 2}$ with respect to all permutations of the indices.*

The injective norm is difficult to compute in general [HL13], so the following upper bound can be more convenient.

Definition 5.4.14 (Matrix covariance)**.** *We write $\mathsf{Cov}(\boldsymbol{X}) := \mathsf{Cov}(\mathsf{vec}(\boldsymbol{X}))$, and*

$$v(\boldsymbol{X})^2 := \|\mathsf{Cov}(\boldsymbol{X})\| = \left\| \sum_{i=1}^D \mathsf{vec}(\boldsymbol{A}_i)\mathsf{vec}(\boldsymbol{A}_i)^\top \right\|.$$

Proposition 5.4.15. *If $\boldsymbol{\Sigma} = \mathsf{Cov}(\mathsf{symvec}(\boldsymbol{X}))$ is as in Proposition 5.1.3, then*

$$\sigma_*(\boldsymbol{X}) \le v(\boldsymbol{X}) \le \sqrt{2}\|\boldsymbol{\Sigma}\|.$$

Proof. For the first inequality, we have

$$\begin{aligned}\sigma_*(\boldsymbol{X})^2 &= \max_{\|\boldsymbol{v}\|=1} \left\langle \sum_{i=1}^D \boldsymbol{A}_i^{\otimes 2}, \boldsymbol{v}^{\otimes 4} \right\rangle \\ &= \max_{\|\boldsymbol{v}\|=1} \left\langle \sum_{i=1}^D \mathsf{vec}(\boldsymbol{A}_i)\mathsf{vec}(\boldsymbol{A}_i)^\top, \mathsf{vec}(\boldsymbol{v}\boldsymbol{v}^\top)\mathsf{vec}(\boldsymbol{v}\boldsymbol{v}^\top)^\top \right\rangle \\ &\le \max_{\|\boldsymbol{v}\|=1} \left\langle \sum_{i=1}^D \mathsf{vec}(\boldsymbol{A}_i)\mathsf{vec}(\boldsymbol{A}_i)^\top, \boldsymbol{v}\boldsymbol{v}^\top \right\rangle \\ &= v(\boldsymbol{X})^2.\end{aligned}$$

For the second inequality, note that $\mathsf{Cov}(\boldsymbol{X})$ is a submatrix of the matrix

$$\begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} & \boldsymbol{\Sigma} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \otimes \boldsymbol{\Sigma},$$

whose operator norm is $2\|\boldsymbol{\Sigma}\|$. □

Theorem 5.4.16. *Let $\boldsymbol{X}$ be the Gaussian series of $(\boldsymbol{A}_i)_{i=0}^D$. Then, $\lambda_1(\boldsymbol{X})$ and $\|\boldsymbol{X}\|$ are both $\sigma_*(\boldsymbol{X})^2$-subgaussian, and hence also $v(\boldsymbol{X})^2$-subgaussian and $2\|\boldsymbol{\Sigma}\|^2$-subgaussian.*

Proof. By the same bounds as we have seen before, we may bound

$$\begin{aligned}
|\lambda_1(\boldsymbol{X}(\boldsymbol{g})) - \lambda_1(\boldsymbol{X}(\boldsymbol{h}))| &\leq \|\boldsymbol{X}(\boldsymbol{g}) - \boldsymbol{X}(\boldsymbol{h})\| \\
&= \left\| \sum_{i=1}^{D} (g_i - h_i) \boldsymbol{A}_i \right\| \\
&= \max_{\|\boldsymbol{v}\|=1} \left| \boldsymbol{v}^\top \left(\sum_{i=1}^{D} (g_i - h_i) \boldsymbol{A}_i \right) \boldsymbol{v} \right| \\
&= \max_{\|\boldsymbol{v}\|=1} \left| \sum_{i=1}^{D} (g_i - h_i)(\boldsymbol{v}^\top \boldsymbol{A}_i \boldsymbol{v}) \right| \\
&\leq \max_{\|\boldsymbol{v}\|=1} \left(\sum_{i=1}^{D} (\boldsymbol{v}^\top \boldsymbol{A}_i \boldsymbol{v})^2 \right)^{1/2} \cdot \|\boldsymbol{g} - \boldsymbol{h}\| \qquad \text{(Cauchy-Schwarz)} \\
&= \sigma_*(\boldsymbol{X}) \cdot \|\boldsymbol{g} - \boldsymbol{h}\|,
\end{aligned}$$

which gives the result by Theorem 5.4.9. The argument for $\|\boldsymbol{X}\|$ is essentially the same, except instead of the inequality $|\lambda_1(\boldsymbol{X}) - \lambda_1(\boldsymbol{Y})| \leq \|\boldsymbol{X} - \boldsymbol{Y}\|$, we use that $|\,\|\boldsymbol{X}\| - \|\boldsymbol{Y}\|\,| \leq \|\boldsymbol{X} - \boldsymbol{Y}\|$. □

Note that, using the weaker result of $v(\boldsymbol{X})^2$-subgaussianity, the case of independent Gaussian entries from before just corresponds to $\boldsymbol{\Sigma} = \mathsf{Cov}(\mathsf{symvec}(\boldsymbol{X}))$ being a diagonal matrix of the σ_{ij}^2, so we easily recover Corollary 5.4.11. At the other extreme, we can consider the matrix $\boldsymbol{X} = g\mathbf{1}_d\mathbf{1}_d^\top$, where $g \sim \mathcal{N}(0,1)$ and thus $\boldsymbol{X}$ is just a constant matrix with all entries equal to g. We have $\lambda_1(\boldsymbol{X}) = \max\{0, dg\}$, and thus is at best d^2-subgaussian, since $\mathsf{Law}(dg) = \mathcal{N}(0, d^2)$ (and the truncation at zero only affects the lower tail, not the upper tail). And indeed, in this case we have $\mathsf{Cov}(\boldsymbol{X}) = \mathbf{1}_{d^2}\mathbf{1}_{d^2}^\top$, so $v(\boldsymbol{X})^2 = d^2$ and the above result is again sharp.

6 | Non-Asymptotic Theory II: Typical Spectral Statistics

6.1 Non-Commutative Khintchine Inequality

We now switch gears slightly and focus on matrix norms rather than maximum eigenvalues. Theorem 5.4.16 gives a remarkably general result showing that $\|\boldsymbol{X}\|$ for any Gaussian random matrix $\boldsymbol{X}$ concentrates around its $\mathbb{E}\|\boldsymbol{X}\|$. The mean remains mysterious: clearly for different choices of $\boldsymbol{A}_i$ in a Gaussian series we may engineer a great variety of matrix structures with eigenvalues on different scales and so forth. Yet, even more surprising than our general treatment of concentration, the following result nearly pins down the *location* of the mean of any Gaussian random matrix up to constants.

Definition 6.1.1. *For a random matrix $\boldsymbol{X}$, define*

$$\sigma(\boldsymbol{X})^2 := \|\mathbb{E}(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})^2\|,$$

sometimes called the matrix variance statistic *or* parameter *(in contrast to its relative in Definition 6.1.1). For $\boldsymbol{X}$ the Gaussian series of $\boldsymbol{A}_0, \boldsymbol{A}_1, \dots, \boldsymbol{A}_D$, this is equivalently*

$$\sigma(\boldsymbol{X})^2 = \|\sum_{i=1}^{D} \boldsymbol{A}_i^2\|.$$

Note that $\sigma(\boldsymbol{X})$ is something that we can compute in terms of the deterministic matrices $\boldsymbol{A}_i$ that specify the Gaussian series model of $\boldsymbol{X}$. Thus it is supremely remarkable that this quantity nearly captures the scaling of the expected norm of $\boldsymbol{X}$ up to constants:

Theorem 6.1.2 (Non-commutative Khintchine (NCK) inequality [LP86, LPP91])**.** *There are absolute constants $c, C > 0$ such that, for any centered $d \times d$ Gaussian random matrix $\boldsymbol{X}$ (i.e., any Gaussian series with $\boldsymbol{A}_0 = \boldsymbol{0}$), we have*

$$c \cdot \sigma(\boldsymbol{X}) \le \mathbb{E}\|\boldsymbol{X}\| \le C\sqrt{\log d} \cdot \sigma(\boldsymbol{X}).$$

Concretely, one may take $c = 1/\sqrt{2}$ and $C = \sqrt{2e}$.

Let us first check that, unfortunately, both the upper and lower bounds can be tight in different cases, so the $\sqrt{\log d}$ cannot be removed in general, though it can *sometimes*, an issue we will revisit later.

Example 6.1.3 (Diagonal matrix)**.** *Consider $\boldsymbol{X}$ diagonal with $X_{ii} \sim \mathcal{N}(0,1)$ i.i.d. Then, we have $\mathbb{E}\boldsymbol{X}^2 = \boldsymbol{I}$, so $\sigma(\boldsymbol{X}) = 1$. On the other hand, $\mathbb{E}\|\boldsymbol{X}\| = \mathbb{E}\max_{i=1}^d |X_{ii}| \asymp \sqrt{\log d}$, so in this case the upper bound above is tight (up to constants).*

Example 6.1.4 (GOE matrix)**.** *Consider $\boldsymbol{X} \sim \mathsf{GOE}^{(0)}(d)$. A simple calculation by expanding the matrix multiplication shows that $\mathbb{E}\boldsymbol{X}^2 = (d-1)\boldsymbol{I}_d$, so $\sigma(\boldsymbol{X}) = \sqrt{d-1}$. And, we have $\mathbb{E}\|\boldsymbol{X}\| = (2+o(1))\sqrt{d}$ as we have seen, so in this case the lower bound above is tight (up to constants).*

Also, let us compare this with the concentration inequality we proved for $\|\boldsymbol{X}\|$. Recall by Theorem 5.4.16 that $\|\boldsymbol{X}\|$ is $\sigma_*(\boldsymbol{X})^2$-subgaussian, for the parameter $\sigma_*(\boldsymbol{X})$ as in Definition 5.4.12. In fact we may show that these fluctuations at least do not swamp the expectation:

Proposition 6.1.5. *We have:*

$$\mathsf{Var}[\|\boldsymbol{X}\|] \lesssim \sigma_*(\boldsymbol{X})^2 \lesssim (\mathbb{E}\|\boldsymbol{X}\|)^2.$$

Proof. The first result is immediate from Theorem 5.4.16. For the second result, we recall the definition

$$\sigma_*(\boldsymbol{X}) = \max_{\|\boldsymbol{v}\|=1} \left(\sum_{i=1}^D (\boldsymbol{v}^\top \boldsymbol{A}_i \boldsymbol{v})^2 \right)^{1/2}.$$

On the other hand,

$$\sum_{i=1}^D (\boldsymbol{v}^\top \boldsymbol{A}_i \boldsymbol{v})^2 = \mathsf{Var}\left[\sum_{i=1}^D \boldsymbol{v}^\top (g_i \boldsymbol{A}_i) \boldsymbol{v} \right] = \mathsf{Var}[\boldsymbol{v}^\top \boldsymbol{X} \boldsymbol{v}],$$

for any fixed $\boldsymbol{v}$. In particular then, we have

$$\left(\sum_{i=1}^D (\boldsymbol{v}^\top \boldsymbol{A}_i \boldsymbol{v})^2 \right)^{1/2} = \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}\left| \boldsymbol{v}^\top \boldsymbol{X} \boldsymbol{v} \right|,$$

since $\sqrt{2/\pi} = \mathbb{E}_{g\sim\mathcal{N}(0,1)}|g|$. Thus, by Jensen's inequality over the maximum,

$$\sigma_*(\boldsymbol{X}) \le \max_{\|\boldsymbol{v}\|=1} \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}\left| \boldsymbol{v}^\top \boldsymbol{X} \boldsymbol{v} \right| \le \sqrt{\frac{\pi}{2}} \cdot \mathbb{E} \max_{\|\boldsymbol{v}\|=1} \left| \boldsymbol{v}^\top \boldsymbol{X} \boldsymbol{v} \right| = \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}\|\boldsymbol{X}\|,$$

completing the proof. □

Let us prove Theorem 6.2.4. We will consider the upper and lower bounds separately. The lower bound is much simpler using our above observation.

Proof of Theorem 6.2.4: Lower Bound. First, note that by Jensen's inequality we have

$$\mathbb{E}\|\boldsymbol{X}\|^2 = \mathbb{E}\|\boldsymbol{X}^2\| \ge \|\mathbb{E}\boldsymbol{X}^2\| = \sigma(\boldsymbol{X})^2.$$

On the other hand, using Proposition 6.1.5 above,

$$\mathbb{E}\|\boldsymbol{X}\|^2 = (\mathbb{E}\|\boldsymbol{X}\|)^2 + \mathsf{Var}[\|\boldsymbol{X}\|] \lesssim (\mathbb{E}\|\boldsymbol{X}\|)^2.$$

Putting the two inequalities together,

$$(\mathbb{E}\|\boldsymbol{X}\|)^2 \gtrsim \sigma(\boldsymbol{X}),$$

and the result follows. □

The real content of the result is in the upper bound. For that, we will follow the "trace method" as we have discussed in Section 2.5. This prescribes that we estimate the expectations of traces of high powers of $\boldsymbol{X}$, which is achieved by the following.

Lemma 6.1.6 (Non-commutative Khintchine trace estimate)**.**

$$\left(\mathbb{E}\,\mathsf{Tr}\,\boldsymbol{X}^{2k}\right)^{\frac{1}{2k}} \leq \sqrt{2k-1}\cdot(\mathsf{Tr}(\mathbb{E}\boldsymbol{X}^2)^k)^{\frac{1}{2k}}.$$

The following is a simple but important preliminary, whose short proof we omit. See [Tro18, vH17] for the details.

Proposition 6.1.7. *Let $\boldsymbol{A},\boldsymbol{B}\in\mathbb{R}^{d\times d}_{\mathsf{sym}}$ and $0\leq k\leq \ell$. Then, $\mathsf{Tr}(\boldsymbol{A}\boldsymbol{B}^{\ell}\boldsymbol{A}\boldsymbol{B}^{2k-\ell})\leq \mathsf{Tr}(\boldsymbol{A}^2\boldsymbol{B}^{2k})$.*

Proof of Lemma 6.1.6. We expand the moments we are interested in as follows:

$$\begin{aligned}
M_{2k} &:= \mathbb{E}\,\mathsf{Tr}\,\boldsymbol{X}^{2k}\\
&= \mathbb{E}\,\mathsf{Tr}\left(\sum_{i=1}^{D} g_i\boldsymbol{A}_i\right)^{2k}\\
&= \sum_{i=1}^{D}\mathbb{E}g_i\,\mathsf{Tr}\left(\boldsymbol{A}_i\left(\sum_{i=1}^{D} g_i\boldsymbol{A}_i\right)^{2k-1}\right)
\end{aligned}$$

Now, using Gaussian integration by parts on each term gives

$$= \sum_{i=1}^{D}\sum_{\ell=0}^{2k-2}\mathbb{E}\,\mathsf{Tr}\left(\boldsymbol{A}_i\left(\sum_{i=1}^{D} g_i\boldsymbol{A}_i\right)^{\ell}\boldsymbol{A}_i\left(\sum_{i=1}^{D} g_i\boldsymbol{A}_i\right)^{2k-2-\ell}\right)$$

to which we may apply Proposition 6.1.7,

$$\begin{aligned}
&\leq \sum_{i=1}^{D}\sum_{\ell=0}^{2k-2}\mathbb{E}\,\mathsf{Tr}\left(\boldsymbol{A}_i^2\left(\sum_{i=1}^{D} g_i\boldsymbol{A}_i\right)^{2k-2}\right)\\
&= (2k-1)\,\mathbb{E}\,\mathsf{Tr}\left(\left(\sum_{i=1}^{D}\boldsymbol{A}_i^2\right)\left(\sum_{i=1}^{D} g_i\boldsymbol{A}_i\right)^{2k-2}\right)\\
&= (2k-1)\,\mathbb{E}\left\langle\sum_{i=1}^{D}\boldsymbol{A}_i^2,\boldsymbol{X}^{2k-2}\right\rangle.
\end{aligned}$$

Let us define

$$\boldsymbol{V} := \sum_{i=1}^{D}\boldsymbol{A}_i^2.$$

By von Neumann's trace inequality followed by Hölder's inequality, we may continue

$$\begin{aligned}
M_{2k} &\le (2k-1)\mathbb{E}\langle \boldsymbol{\lambda}(\boldsymbol{V}), \boldsymbol{\lambda}(\boldsymbol{X}^{2k-2})\rangle && \text{(von Neumann)}\\
&\le (2k-1)\|\boldsymbol{\lambda}(\boldsymbol{V})\|_k \mathbb{E}\|\boldsymbol{\lambda}(\boldsymbol{X}^{2k-2})\|_{\frac{k}{k-1}} && \text{(Hölder)}\\
&= (2k-1)\operatorname{Tr}(\boldsymbol{V}^k)^{1/k}\mathbb{E}\operatorname{Tr}(\boldsymbol{X}^{(2k-2)\cdot\frac{k}{k-1}})^{\frac{k-1}{k}}\\
&= (2k-1)\operatorname{Tr}(\boldsymbol{V}^k)^{1/k}\mathbb{E}\operatorname{Tr}(\boldsymbol{X}^{2k})^{1-\frac{1}{k}}\\
&\le (2k-1)\operatorname{Tr}(\boldsymbol{V}^k)^{1/k}(\mathbb{E}\operatorname{Tr}\boldsymbol{X}^{2k})^{1-\frac{1}{k}} && \text{(Jensen)}\\
&= (2k-1)\operatorname{Tr}(\boldsymbol{V}^k)^{1/k}M_{2k}^{1-\frac{1}{k}}.
\end{aligned}$$

Finally we have reached a closed inequality for M_{2k}, which, upon rearranging, reads

$$M_{2k}^{\frac{1}{2k}} \le \sqrt{2k-1}\operatorname{Tr}(\boldsymbol{V}^k)^{\frac{1}{2k}},$$

completing the proof. □

With the Lemma proved, we can give the proof of the upper bound of Theorem 6.2.4 easily.

Proof of Theorem 6.2.4: Upper Bound. We have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{X}\| &\le \mathbb{E}(\operatorname{Tr}\boldsymbol{X}^{2k})^{\frac{1}{2k}}\\
&\le (\mathbb{E}\operatorname{Tr}\boldsymbol{X}^{2k})^{\frac{1}{2k}} && \text{(Jensen)}\\
&\le \sqrt{2k}\cdot(\operatorname{Tr}(\mathbb{E}\boldsymbol{X}^2)^k)^{\frac{1}{2k}} && \text{(Lemma 6.1.6)}\\
&\le \sqrt{2k}\cdot(d\cdot\sigma^{2k})^{\frac{1}{2k}}\\
&= d^{\frac{1}{2k}}\sqrt{2k}\cdot\sigma.
\end{aligned}$$

Optimizing over k gives that the best choice is $k \asymp \log d$, and plugging this in gives the result. □

6.2 Sums of Independent Random Matrices

We will next see how it is possible to derive very general matrix concentration inequalities for sums of *arbitrary* independent random matrices,

$$\boldsymbol{X} = \sum_i \boldsymbol{H}_i \text{ for } \boldsymbol{H}_i \text{ random and independent,}$$

starting from the NCK inequality. This should be surprising: an entire random matrix $\boldsymbol{H}_i$ seems to allow for much more general structure than the $g_i\boldsymbol{A}_i$ for deterministic $\boldsymbol{A}_i$ that appear in NCK. Yet, these extensions actually follow from NCK by two elementary tricks.

6.2.1 Trick 1: Gaussian to Rademacher

The first trick is truly trivial-looking: the NCK inequality with Gaussian random variables implies the same, with a slightly larger constant, for *Rademacher* random variables having distribution $\mathsf{Unif}(\{\pm 1\})$.

Lemma 6.2.1. *Suppose $\boldsymbol{A}_i \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ are deterministic. We have*

$$\mathop{\mathbb{E}}_{z_i\sim\mathsf{Unif}(\{\pm 1\})} \left\|\sum_i z_i \boldsymbol{A}_i\right\| \le \sqrt{\frac{\pi}{2}} \mathop{\mathbb{E}}_{g_i\sim\mathcal{N}(0,1)} \left\|\sum_i g_i \boldsymbol{A}_i\right\|.$$

Proof. Introduce $g_i \sim \mathcal{N}(0,1)$ and $z_i \sim \mathsf{Unif}(\{\pm 1\})$ all independent of one another. Since $\mathsf{Law}(g_i) = \mathsf{Law}(z_i|g_i|)$, we may rewrite and use Jensen's inequality:

$$\begin{aligned}\mathop{\mathbb{E}}_{g_i}\left\|\sum_i g_i \boldsymbol{A}_i\right\| &= \mathop{\mathbb{E}}_{z_i,g_i}\left\|\sum_i z_i|g_i|\boldsymbol{A}_i\right\| \\ &\ge \mathop{\mathbb{E}}_{z_i}\left\|\sum_i \mathbb{E}[|g_i|]\cdot z_i\boldsymbol{A}_i\right\| \\ &= \sqrt{\frac{2}{\pi}}\mathop{\mathbb{E}}_{z_i\sim\mathsf{Unif}(\{\pm 1\})}\left\|\sum_i z_i\boldsymbol{A}_i\right\|,\end{aligned}$$

and the result follows by rearranging. We have used the elementary integral calculation $\mathbb{E}[|g_i|] = \sqrt{2/\pi}$. □

We call a random matrix of the form $\sum_i z_i \boldsymbol{A}_i$ a *Rademacher series*. As a consequence, a version of the NCK inequality holds for Rademacher series. In fact, this was the original form of the NCK inequality studied in [LP86, LPP91]; the conveniences of the Gaussian form were only realized later.

Corollary 6.2.2 (Rademacher NCK inequality)**.** *There are absolute constants $c, C > 0$ such that, for any centered $d \times d$ Rademacher series $\boldsymbol{X} = \sum_{i=1}^D z_i \boldsymbol{A}_i$ with $z_i \sim \mathsf{Unif}(\{\pm 1\})$, we have*

$$c\cdot\sigma(\boldsymbol{X}) \le \mathbb{E}\|\boldsymbol{X}\| \le C\sqrt{\log d}\cdot\sigma(\boldsymbol{X}).$$

6.2.2 Trick 2: Symmetrization

It seems like we have not made much progress. The next trick is really the key: any sum of independent random matrices, merely so long as they are *centered*, can be controlled by a random Rademacher series: $\sum_i z_i \boldsymbol{A}_i$ where the $\boldsymbol{A}_i$ are random but independent of z_i. As we will see, we may then apply NCK conditionally on the $\boldsymbol{A}_i$ and obtain non-trivial bounds.

Lemma 6.2.3. *Suppose $\boldsymbol{H}_i \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ are arbitrary independent random matrices. Introduce $z_i \sim \mathsf{Unif}(\{\pm 1\})$ independent of one another and of the $\boldsymbol{H}_i$. Then,*

$$\mathbb{E}\left\|\sum_i (\boldsymbol{H}_i - \mathbb{E}\boldsymbol{H}_i)\right\| \le 2\mathbb{E}\left\|\sum_i z_i \boldsymbol{H}_i\right\|.$$

Proof. Introduce $(\boldsymbol{H}_i')$ an independent copy of the collection $(\boldsymbol{H}_i)$. We then have

$$\begin{aligned}\mathbb{E}\left\|\sum_i(\boldsymbol{H}_i-\mathbb{E}\boldsymbol{H}_i)\right\| &= \mathop{\mathbb{E}}_{\boldsymbol{H}_i}\left\|\mathop{\mathbb{E}}_{\boldsymbol{H}_i'}\sum_i(\boldsymbol{H}_i-\boldsymbol{H}_i')\right\| \\ &\le \mathbb{E}\left\|\sum_i(\boldsymbol{H}_i-\boldsymbol{H}_i')\right\| \qquad \text{(Jensen)}\end{aligned}$$

and now, since $\boldsymbol{H}_i-\boldsymbol{H}_i'$ has a symmetric law, we have

$$\begin{aligned}&= \mathbb{E}\left\|\sum_i z_i(\boldsymbol{H}_i-\boldsymbol{H}_i')\right\| \\ &\le \mathbb{E}\left\|\sum_i z_i\boldsymbol{H}_i\right\| + \mathbb{E}\left\|\sum_i z_i\boldsymbol{H}_i'\right\| \\ &= 2\mathbb{E}\left\|\sum_i z_i\boldsymbol{H}_i\right\|,\end{aligned}$$

giving the result. □

As a corollary, we can derive the following inequality, which we may view as a version of the NCK inequality for general sums of independent random matrices.

Corollary 6.2.4 (Independent sum NCK)**.** *There is an absolute constant $C>0$ such that, for any independent random $\boldsymbol{H}_i\in\mathbb{R}^{d\times d}_{\mathsf{sym}}$,*

$$\mathbb{E}\left\|\sum_i(\boldsymbol{H}_i-\mathbb{E}\boldsymbol{H}_i)\right\| \le C\sqrt{\log d}\,\mathbb{E}\left\|\sum_i\boldsymbol{H}_i^2\right\|^{1/2}.$$

6.2.3 Matrix Chernoff Bound

Theorem 6.2.5 (Matrix Chernoff)**.** *There are absolute constants $C_1,C_2>0$ such that, for any $\boldsymbol{H}_i\in\mathbb{R}^{d\times d}_{\mathsf{sym}}$ random and satisfying $\boldsymbol{H}_i\succeq\mathbf{0}$ almost surely, and for any $\epsilon>0$, we have*

$$\begin{aligned}\mathbb{E}\left\|\sum_i\boldsymbol{H}_i\right\| &\le \left(\left\|\sum_i\mathbb{E}\boldsymbol{H}_i\right\|^{1/2} + C_1\sqrt{\log d}\,\left(\mathbb{E}\max_i\|\boldsymbol{H}_i\|\right)^{1/2}\right)^2 \\ &\le (1+\epsilon)\left\|\sum_i\mathbb{E}\boldsymbol{H}_i\right\| + \left(1+\frac{1}{\epsilon}\right)C_2\log d\,\mathbb{E}\max_i\|\boldsymbol{H}_i\|.\end{aligned}$$

Often in applications, $\|\boldsymbol{H}_i\|$ is uniformly bounded almost surely, in which case the second term above is just $O(\log d)$. In that case, the inequality is usually effective just once there are sufficiently many of the $\boldsymbol{H}_i$ that the first term overwhelms this error term. Also, sometimes it is fine to take $\epsilon=1$ and just use a bound ignoring constant factors,

$$\mathbb{E}\left\|\sum_i\boldsymbol{H}_i\right\| \lesssim \left\|\sum_i\mathbb{E}\boldsymbol{H}_i\right\| + \log d\,\mathbb{E}\max_i\|\boldsymbol{H}_i\|,$$

but sometimes, especially when the number of terms in the sum is large, it can be useful to take ϵ small in which case the above can become a quite sharp estimate.

Proof. First, note that we may bound by triangle inequality

$$\mathbb{E}\left\|\sum_i \boldsymbol{H}_i\right\| \le \left\|\sum_i \mathbb{E}\boldsymbol{H}_i\right\| + \mathbb{E}\left\|\sum_i (\boldsymbol{H}_i - \mathbb{E}\boldsymbol{H}_i)\right\|.$$

On the second term, we use the independent sum NCK:

$$\begin{aligned}
\mathbb{E}\left\|\sum_i (\boldsymbol{H}_i - \mathbb{E}\boldsymbol{H}_i)\right\| &\lesssim \sqrt{\log d}\,\mathbb{E}\left\|\sum_i \boldsymbol{H}_i^2\right\|^{1/2} && \text{(independent sum NCK)} \\
&\le \sqrt{\log d}\,\mathbb{E}\left[\left(\max_i \|\boldsymbol{H}_i\|\right)^{1/2}\left\|\sum_i \boldsymbol{H}_i\right\|^{1/2}\right] \\
&\le \sqrt{\log d}\left(\mathbb{E}\max_i \|\boldsymbol{H}_i\|\right)^{1/2}\left(\mathbb{E}\left\|\sum_i \boldsymbol{H}_i\right\|\right)^{1/2}. && \text{(Cauchy-Schwarz)}
\end{aligned}$$

Thus we find that, for some absolute $C > 0$,

$$\mathbb{E}\left\|\sum_i \boldsymbol{H}_i\right\| \le \left\|\sum_i \mathbb{E}\boldsymbol{H}_i\right\| + C\sqrt{\log d}\left(\mathbb{E}\max_i \|\boldsymbol{H}_i\|\right)^{1/2}\left(\mathbb{E}\left\|\sum_i \boldsymbol{H}_i\right\|\right)^{1/2}.$$

But now, this is a quadratic inequality for the norm we are interested in. Solving it gives that, for another constant C',

$$\mathbb{E}\left\|\sum_i \boldsymbol{H}_i\right\| \le \left(\left\|\sum_i \mathbb{E}\boldsymbol{H}_i\right\|^{1/2} + C'\sqrt{\log d}\left(\mathbb{E}\max_i \|\boldsymbol{H}_i\|\right)^{1/2}\left(\mathbb{E}\left\|\sum_i \boldsymbol{H}_i\right\|\right)^{1/2}\right)^2,$$

which is the first form of the result. The second follows from using a weighted arithmetic-geometric mean inequality, a simple but handy scalar bound:

$$\begin{aligned}
(x+y)^2 &= x^2 + y^2 + 2xy \\
&= x^2 + y^2 + 2\left(\sqrt{\epsilon}x\right)\left(\frac{y}{\sqrt{\epsilon}}\right) \\
&\le x^2 + y^2 + \left(\sqrt{\epsilon}x\right)^2 + \left(\frac{y}{\sqrt{\epsilon}}\right)^2 \\
&= (1+\epsilon)x^2 + \left(1 + \frac{1}{\epsilon}\right)y^2,
\end{aligned}$$

completing the proof. □

6.2.4 Matrix Bernstein Bound

Theorem 6.2.6 (Matrix Bernstein)**.** *There are absolute constants $C_1, C_2 > 0$ such that, for any $\boldsymbol{H}_i \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$ random with $\mathbb{E}\boldsymbol{H}_i = \mathbf{0}$,*

$$\mathbb{E}\left\|\sum_i \boldsymbol{H}_i\right\| \le C_1\sqrt{\log d}\left\|\sum_i \mathbb{E}\boldsymbol{H}_i^2\right\|^{1/2} + C_2 \log d\,\left(\mathbb{E}\max_i \|\boldsymbol{H}_i\|^2\right)^{1/2}.$$

If we write $\boldsymbol{X} := \sum_i \boldsymbol{H}_i$ and $\|\boldsymbol{H}_i\| \le K$ almost surely for all i, then we may write the above in a form similar to the upper bound of the NCK inequality, with an extra error term:

$$\mathbb{E}\left\|\sum_i \boldsymbol{H}_i\right\| \lesssim \sqrt{\log d} \cdot \sigma(\boldsymbol{X}) + \log d \cdot K.$$

In this sense, which we will discuss further a little bit later, a fruitful way to view the matrix Bernstein inequality is as a universal NCK inequality*, extending the Gaussian NCK setting to general independent sums.*

As with matrix Chernoff, quite often in applications we have $\|\boldsymbol{H}_i\|$ uniformly bounded almost surely, in which case the second term is just $O(\log d)$.

Note also that, if we write $\boldsymbol{X} := \sum_i \boldsymbol{H}_i$, then $\sigma(\boldsymbol{X}) = \|\mathbb{E}\boldsymbol{X}^2\|^{1/2} = \|\sum_i \mathbb{E}\boldsymbol{H}_i^2\|^{1/2}$ is precisely what appears in the first term above, and the entire first term is $\sqrt{\log d}\,\sigma(\boldsymbol{X})$, just as in the NCK inequality. Thus the matrix Bernstein inequality is a sort of "universal NCK" inequality for sums of independent random matrices.

Proof. The result follows directly by composing the independent sum NCK with the matrix Chernoff bound, as follows:

$$\begin{aligned}
\mathbb{E}\left\|\sum_i \boldsymbol{H}_i\right\| &\lesssim \sqrt{\log d}\,\mathbb{E}\left\|\sum_i \boldsymbol{H}_i^2\right\|^{1/2} && \text{(independent sum NCK)}\\
&\lesssim \sqrt{\log d}\left(\mathbb{E}\left\|\sum_i \boldsymbol{H}_i^2\right\|\right)^{1/2} && \text{(Jensen)}\\
&\lesssim \sqrt{\log d}\left(\left\|\sum_i \mathbb{E}\boldsymbol{H}_i^2\right\|^{1/2} + \sqrt{\log d}\,\mathbb{E}\max_i \|\boldsymbol{H}_i\|\right) && \text{(matrix Chernoff)}
\end{aligned}$$

which gives the result. □

6.3 Application: Covariance Estimation Revisited

Consider the matter of estimating the covariance of a probability distribution from samples that we have seen several times before. Consider, though, a much more general model than the Gaussian one that we discussed: suppose $\boldsymbol{v}_1, \dots, \boldsymbol{v}_m \in \mathbb{R}^d$ are i.i.d. random vectors, with $\mathbb{E}\boldsymbol{v}_i = \boldsymbol{0}$ and $\mathrm{Cov}[\boldsymbol{v}_i] = \mathbb{E}\boldsymbol{v}_i\boldsymbol{v}_i^\top = \boldsymbol{\Sigma}$. We consider the sample covariance,

$$\hat{\boldsymbol{\Sigma}} := \frac{1}{m}\sum_{i=1}^m \boldsymbol{v}_i\boldsymbol{v}_i^\top.$$

How well does $\hat{\boldsymbol{\Sigma}}$ approximate $\boldsymbol{\Sigma}$? We will study this general question under just the general condition that the $\boldsymbol{v}_i$ are bounded: for some $K > 0$, almost surely we have

$$\|\boldsymbol{v}_i\| \le K.$$

We will study the norm of the difference,

$$\hat{\Sigma} - \Sigma = \sum_{i=1}^{m} \underbrace{\frac{1}{m}(v_i v_i^\top - \Sigma)}_{H_i},$$

which the above shows may be written as a sum of independent (indeed, in this case, i.i.d.) centered matrices $H_i = \frac{1}{m}(v_i v_i^\top - \Sigma) = \frac{1}{m}(v_i v_i^\top - \mathbb{E} v_i v_i^\top)$. Let us gather the information needed to apply matrix Bernstein, namely a uniform bound on the norms of the H_i and the first two moments of H_i:

$$\begin{aligned}
\|H_i\| &\le \frac{1}{m}(\|v_i\|^2 + \mathbb{E}\|v_i\|^2) \\
&\le \frac{2K}{m} \text{ almost surely,} \\
\mathbb{E} H_i &= 0, \\
\mathbb{E} H_i^2 &= \frac{1}{m^2}\mathbb{E}(v_i v_i^\top - \mathbb{E} v_i v_i^\top)^2 \\
&= \frac{1}{m^2}\mathbb{E}(v_i v_i^\top)^2 - (\mathbb{E} v_i v_i^\top)^2 \\
&\preceq \frac{1}{m^2}\mathbb{E}\|v_i\|^2 v_i v_i^\top \\
&\preceq \frac{K^2}{m^2}\Sigma.
\end{aligned}$$

Plugging this information into matrix Bernstein, we find

$$\begin{aligned}
\mathbb{E}\|\hat{\Sigma} - \Sigma\| &= \mathbb{E}\left\|\sum_i H_i\right\| \\
&\lesssim \sqrt{\log d}\left\|\sum_i \mathbb{E} H_i^2\right\|^{1/2} + \log d \left(\mathbb{E}\max_i \|H_i\|^2\right)^{1/2} \\
&\lesssim \sqrt{\log d} \cdot \sqrt{m \cdot \frac{K^2}{m^2} \cdot \|\Sigma\|} + \log d \cdot \frac{K^2}{m}.
\end{aligned}$$

A more natural quantity is the relative error in our estimate, which then takes the form

$$\frac{\mathbb{E}\|\hat{\Sigma} - \Sigma\|}{\|\Sigma\|} \lesssim \cdot \sqrt{\frac{K^2 \log d}{m\|\Sigma\|}} + \cdot \frac{K^2 \log d}{m\|\Sigma\|}.$$

Thus $\hat{\Sigma}$ will be a good estimate (with shrinking relative error) provided that

$$m \gg \frac{K^2 \log d}{\|\Sigma\|}.$$

If we scale the v_i such that $K = O(1)$, then the typical scaling to expect from v_i that are not exceptionally constrained is that $\|\Sigma\| = \Theta(1/d)$. For instance, if $v_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}(1))$, then $K = 1$, while $\Sigma = \mathbb{E} v_i v_i^\top = \frac{1}{d} I_d$. Thus the result says that, generically, we expect the sample complexity of estimating the sample covariance of a "nice" bounded distribution of random vectors to be roughly $d \log d$.

6.4 Application: Randomized Numerical Linear Algebra

6.4.1 Randomized Sparse Matrix Approximation

We next consider a more sophisticated use of matrix Bernstein appearing in computational applications. Suppose that we have a matrix $\boldsymbol{Y} \in \mathbb{R}^{d\times d}_{\mathsf{sym}}$, and we want to approximate $\boldsymbol{Y}$ by some $\widehat{\boldsymbol{Y}}$ which satisfies (1) that $\|\widehat{\boldsymbol{Y}} - \boldsymbol{Y}\|$ is small, but also (2) that $\widehat{\boldsymbol{Y}}$ is *sparse* (in its entries).

We propose the following general class of methods to produce such an estimate, within which we will then try to find the best choice. Define

$$\boldsymbol{E}_{ij} := \left\{ \begin{array}{ll} \boldsymbol{e}_i \boldsymbol{e}_i^\top & \text{if } i = j, \\ \boldsymbol{e}_i \boldsymbol{e}_j^\top + \boldsymbol{e}_j \boldsymbol{e}_i^\top & \text{if } i \neq j \end{array} \right\}.$$

We may expand $\boldsymbol{Y}$ in terms of these matrices and its entries as

$$\boldsymbol{Y} = \sum_{1\leq i\leq j\leq d} y_{ij} \boldsymbol{E}_{ij}.$$

Now, we will design a very sparse estimator of $\boldsymbol{Y}$, indeed, a matrix with just one or two non-zero entries. Suppose that $(p_{ij})_{1\leq i\leq j\leq d}$ is a probability distribution: $p_{ij} \geq 0$ and $\sum p_{ij} = 1$. We define a random matrix $\boldsymbol{G}_1$ by enumerating the values it takes and specifying their probabilities:

$$\boldsymbol{G}_1 = \frac{1}{p_{ij}} y_{ij} \boldsymbol{E}_{ij} \text{ with probability } p_{ij}.$$

We have engineered things such that

$$\mathbb{E}\boldsymbol{G}_1 = \sum_{1\leq i\leq j\leq d} p_{ij} \cdot \frac{1}{p_{ij}} y_{ij} \boldsymbol{E}_{ij} = \boldsymbol{Y}.$$

Thus $\boldsymbol{G}_1$ is a very sparse and *unbiased* estimator of $\boldsymbol{Y}$. Of course, its variance will be huge, and $\|\boldsymbol{G}_1 - \boldsymbol{Y}\|$ will be very large with high probability for most choices of $\boldsymbol{Y}$.

Still, we have made progress: given an unbiased randomized estimator, we may perform a simple form of *variance reduction* by averaging together several runs of the estimator. That is, let $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_m$ be i.i.d. copies of $\boldsymbol{G}_1$, and set

$$\widehat{\boldsymbol{Y}} := \frac{1}{m} \sum_{a=1}^{m} \boldsymbol{G}_a.$$

Again, we have $\mathbb{E}\widehat{\boldsymbol{Y}} = \boldsymbol{Y}$, and indeed as $m \to \infty$ by the law of large numbers $\widehat{\boldsymbol{Y}}$ will converge to $\boldsymbol{Y}$. The question is: quantitatively and non-asymptotically, what governs how large we must take m for this strategy to work well?

Proceeding as in the case of covariance estimation, we have

$$\widehat{\boldsymbol{Y}} - \boldsymbol{Y} = \sum_{a=1}^{m} \frac{1}{m} (\boldsymbol{G}_a - \boldsymbol{Y}) = \sum_{a=1}^{m} \underbrace{\frac{1}{m} (\boldsymbol{G}_a - \mathbb{E}\boldsymbol{G}_a)}_{\boldsymbol{H}_a}.$$

We may now proceed just as before in computing the basic quantities concerning $\boldsymbol{H}_a$ that we need in order to apply matrix Bernstein. We need two things: a uniform bound on $\|\boldsymbol{H}_a\|$ (which will control the $\mathbb{E}\max\|\boldsymbol{H}_a\|^2$ term sufficiently for our purposes, as in our analysis of covariance estimation) and control of $\mathbb{E}\boldsymbol{H}_a^2$. We have:

$$\begin{aligned}\|\boldsymbol{H}_a\| &\leq \frac{1}{m}(\|\boldsymbol{G}_a\| + \|\mathbb{E}\boldsymbol{G}_a\|) \\ &\leq \frac{1}{m}(\|\boldsymbol{G}_a\| + \mathbb{E}\|\boldsymbol{G}_a\|) && \text{(Jensen)} \\ &\leq \frac{4}{m}\max_{i,j}\frac{|y_{ij}|}{p_{ij}},\end{aligned}$$

using that $\|\boldsymbol{E}_{ij}\| \leq 2$. And,

$$\begin{aligned}\mathbb{E}\boldsymbol{H}_a^2 &= \frac{1}{m^2}\mathbb{E}(\boldsymbol{G}_a - \mathbb{E}\boldsymbol{G}_a)^2 \\ &= \frac{1}{m^2}(\mathbb{E}\boldsymbol{G}_a^2 - (\mathbb{E}\boldsymbol{G}_a)^2) \\ &\preceq \frac{1}{m^2}\mathbb{E}\boldsymbol{G}_a^2 \\ &= \frac{1}{m^2}\sum_{1\leq i\leq j\leq d} p_{ij}\cdot\frac{1}{p_{ij}^2}y_{ij}^2\boldsymbol{E}_{ij}^2\end{aligned}$$

and now, since $\boldsymbol{E}_{ii}^2 = \boldsymbol{E}_{ii}$ while $\boldsymbol{E}_{ij}^2 = \boldsymbol{E}_{ii} + \boldsymbol{E}_{jj}$, the resulting matrix is diagonal,

$$= \frac{1}{m^2}\boldsymbol{D},$$

where

$$D_{ii} = \sum_{j=1}^{d}\frac{y_{ij}^2}{p_{ij}}.$$

Thus, for the actual parameter appearing in matrix Bernstein, as the above does not depend on a,

$$\left\|\sum_{a=1}^{m}\mathbb{E}\boldsymbol{H}_a^2\right\| \leq \left\|m\cdot\frac{1}{m^2}\boldsymbol{D}\right\| = \frac{1}{m}\max_{i=1}^{d}\sum_{j=1}^{d}\frac{y_{ij}^2}{p_{ij}}.$$

Putting the pieces together, matrix Bernstein then implies

$$\begin{aligned}\mathbb{E}\|\widehat{\boldsymbol{Y}} - \boldsymbol{Y}\| &= \mathbb{E}\left\|\sum_{a=1}^{m}\boldsymbol{H}_a\right\| \\ &\lesssim \sqrt{\log d}\left\|\sum_{a=1}^{m}\mathbb{E}\boldsymbol{H}_a^2\right\|^{1/2} + \log d\left(\mathbb{E}\max_{a=1}^{m}\|\boldsymbol{H}_a\|^2\right)^{1/2} \\ &= \Bigg(\frac{\log d}{m}\underbrace{\max_{i=1}^{d}\sum_{j=1}^{d}\frac{y_{ij}^2}{p_{ij}}}_{①}\Bigg)^{1/2} + \frac{\log d}{m}\underbrace{\max_{i,j}\frac{|y_{ij}|}{p_{ij}}}_{②}.\end{aligned}$$

We now have a concrete bound on the error in our sparse approximation in terms of the p_{ij}, which we can try to optimize over this choice. The second expression (2) in p_{ij} above is clearly minimized when we set

$$p_{ij} := p_{ij}^{(2)} = \frac{|y_{ij}|}{\|\boldsymbol{Y}\|_{\ell^1}},$$

where the denominator is

$$\|\boldsymbol{Y}\|_{\ell^1} := \sum_{k,\ell} |y_{k,\ell}|.$$

Note that we have the general inequality

$$\|\boldsymbol{Y}\|_{\ell^1} \le d\|\boldsymbol{Y}\|_F. \tag{6.4.1}$$

So, when we choose $p_{ij} := p_{ij}^{(2)}$, we have

$$(2) = \max_{i,j} \frac{|y_{ij}|}{p_{ij}^{(1)}} = \|\boldsymbol{Y}\|_{\ell^1} \le d\|\boldsymbol{Y}\|_F.$$

To make the two terms comparable, let us consider how to make $\|\boldsymbol{Y}\|_F$ also appear in the first expression. We could do this by setting

$$p_{ij} := p_{ij}^{(1)} = \frac{y_{ij}^2}{\|\boldsymbol{Y}\|_F^2}.$$

Indeed, this achieves

$$(1) = \max_{i=1}^{d} \sum_{j=1}^{d} \frac{y_{ij}^2}{p_{ij}^{(1)}} = d\|\boldsymbol{Y}\|_F^2.$$

To compromise between these choices, what we will actually do is to take

$$p_{ij} := \frac{p_{ij}^{(1)}}{2} + \frac{p_{ij}^{(2)}}{2} = \frac{1}{2}\frac{y_{ij}^2}{\|\boldsymbol{Y}\|_F^2} + \frac{1}{2}\frac{|y_{ij}|}{\|\boldsymbol{Y}\|_{\ell^1}}.$$

In particular, note that up to a constant (of 2) we have

$$p_{ij} \gtrsim p_{ij}^{(1)},$$
$$p_{ij} \gtrsim p_{ij}^{(2)}.$$

Since p_{ij} always occurs in the denominator of our bound, this means we can have, up to a constant, the best behavior in each term. We find, also using the general inequality (6.4.1),

$$\begin{aligned}
\mathbb{E}\|\widehat{\boldsymbol{Y}} - \boldsymbol{Y}\| &\lesssim \left(\frac{\log d}{m} \max_{i=1}^{d} \sum_{j=1}^{d} \frac{y_{ij}^2}{p_{ij}^{(1)}}\right)^{1/2} + \frac{\log d}{m} \max_{i,j} \frac{|y_{ij}|}{p_{ij}^{(2)}} \\
&= \left(\frac{d\log d}{m}\|\boldsymbol{Y}\|_F^2\right)^{1/2} + \frac{\log d}{m} \max_{i,j} \|\boldsymbol{Y}\|_{\ell^1} \\
&\le \left(\sqrt{\frac{d\log d}{m}} + \frac{d\log d}{m}\right) \|\boldsymbol{Y}\|_F
\end{aligned}$$

and, provided we choice $m \geq d \log d$, we will then have

$$\lesssim \sqrt{\frac{d \log d}{m}} \|\boldsymbol{Y}\|_F^2.$$

Now, as for covariance estimation, let us consider the relative error to $\|\boldsymbol{Y}\|$. The following quantity will arise.

Definition 6.4.1 (Stable rank)**.** *The* stable rank *of* $\boldsymbol{Y}$ *is*

$$\mathsf{srank}(\boldsymbol{Y}) := \frac{\|\boldsymbol{Y}\|_F^2}{\|\boldsymbol{Y}\|^2} = \sum_{i=1}^{d} \frac{\lambda_i(\boldsymbol{Y})^2}{\max_{j=1}^{d} \lambda_j(\boldsymbol{Y})^2}.$$

The above expression makes the name intuitive: we always have

$$1 \leq \mathsf{srank}(\boldsymbol{Y}) \leq \mathsf{rank}(\boldsymbol{Y}),$$

and the stable rank is like a weighted rank that gives smaller weight to small non-zero eigenvalues. We then have

$$\frac{\mathbb{E}\|\widehat{\boldsymbol{Y}} - \boldsymbol{Y}\|}{\|\boldsymbol{Y}\|} \lesssim \sqrt{\frac{d \log d \cdot \mathsf{srank}(\boldsymbol{Y})}{m}}.$$

In particular, we may obtain a very sparse approximation ($m \ll d^2$) with low relative error (the above being $\ll 1$) once

$$\mathsf{srank}(\boldsymbol{Y}) \ll \frac{d}{\log d}.$$

Thus we have found an expression of the general principle that **matrices of low (stable) rank admit good sparse approximations**.

6.4.2 Sketch of Randomized Matrix Multiplication

We briefly mention how essentially the same ideas allow for a randomized approximation to matrix multiplication, whose practical relevance might be clearer. Given matrices $\boldsymbol{Y} \in \mathbb{R}^{m \times d}$, $\boldsymbol{Z} \in \mathbb{R}^{d \times n}$ whose product $\boldsymbol{Y}\boldsymbol{Z}$ we want to compute, consider expanding

$$\boldsymbol{Y} = \begin{bmatrix} | & | & & | \\ \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_d \\ | & | & & | \end{bmatrix},$$

$$\boldsymbol{Z} = \begin{bmatrix} - & \boldsymbol{z}_1 & - \\ - & \boldsymbol{z}_2 & - \\ & \vdots & \\ - & \boldsymbol{z}_d & - \end{bmatrix}.$$

Then, we have

$$\boldsymbol{Y}\boldsymbol{Z} = \sum_{i=1}^{d} \boldsymbol{y}_i \boldsymbol{z}_i^\top.$$

This suggests a similar "empirical" approach: introduce a discrete probability distribution $(p_i)_{i=1}^d$, and define $\boldsymbol{G}_1, \dots, \boldsymbol{G}_m$ i.i.d. random matrices taking values

$$\frac{1}{p_i} \boldsymbol{y}_i \boldsymbol{z}_i^\top \text{ with probability } p_i.$$

We may again consider an estimator $\frac{1}{m} \sum_{a=1}^m \boldsymbol{G}_a$, and a direct analog of the above analysis goes through. See [Tro15] for details.

6.5 Application: Connectivity of Random Graphs

We next give two applications with a quite different flavor, involving graphs and combinatorial optimization. As a warmup, we consider the question: how dense does a random graph have to be in order to be connected with high probability? We specifically consider the following very simple distribution of random graph:

Definition 6.5.1 (Erdős-Rényi graph)**.** *We write $G \sim \mathcal{G}(d, p)$ for the graph on vertex set $[d]$ where $i \sim j$ independently with probability $p \in (0, 1)$ for each $i < j$.*

Over this class of graphs, we ask: how large does p have to be in order for G to typically be connected?

It is at first unclear what this has to do with random matrix theory. While we have seen before that there is a bridge between random matrices and random graphs through the adjacency matrix, it is still not obvious how to relate spectral properties of the adjacency matrix to connectedness of the associated graph. In fact, a different matrix is more useful for this purpose.

Definition 6.5.2 (Graph Laplacian)**.** *Given a graph G, we write $\boldsymbol{A} = \boldsymbol{A}_G$ for its adjacency matrix, $\boldsymbol{D} = \boldsymbol{D}_G$ for the diagonal matrix with $D_{ii} = \deg(i)$ (the degree of i, or the number of neighbors it has in G), and define the* graph Laplacian *as*

$$\boldsymbol{L} = \boldsymbol{L}_G := \boldsymbol{A}_G - \boldsymbol{D}_G.$$

The following is the key observation to understanding many properties of the graph Laplacian:

Proposition 6.5.3. *$\boldsymbol{L}$ satisfies the following properties:*

$$\boldsymbol{L} = \sum_{i \sim j} (\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top,$$

and, for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$\boldsymbol{x}^\top \boldsymbol{L} \boldsymbol{x} = \sum_{i \sim j} (x_i - x_j)^2.$$

You may check that the classical Laplacian Δf of a smooth function $f : \mathbb{R}^d \to \mathbb{R}^d$ likewise satisfies:

$$\int_{\mathbb{R}^d} f(\boldsymbol{x})(\Delta f)(\boldsymbol{x})\, d\boldsymbol{x} = -\int_{\mathbb{R}^d} \|\nabla f(\boldsymbol{x})\|^2\, d\boldsymbol{x}$$

by integrating by parts. Thus, $\boldsymbol{x}^\top \boldsymbol{L}\boldsymbol{x}$ plays the role of the "average norm-squared of the gradient" in discrete settings over graphs (for instance in the Poincaré inequalities we saw earlier when considered for measures over graphs).

Note that, in particular, we have $\boldsymbol{L} \succeq \boldsymbol{0}$, and $\boldsymbol{x}^\top \boldsymbol{L}\boldsymbol{x} = 0$ only if whenever $x_i \neq x_j$, then i and j are not connected in G. The next result then follows immediately.

Corollary 6.5.4. *Let $S_1, \dots, S_k$ be the connected components of G, and $\mathbf{1}_{S_i}$ the associated indicator vectors of their vertices in $[d]$. Then, the kernel of $\boldsymbol{L}$ is precisely the span of $\mathbf{1}_{S_1}, \dots, \mathbf{1}_{S_k}$. In particular, we have*

$$0 = \lambda_d(\boldsymbol{L}) = \cdots = \lambda_{d-k+1}(\boldsymbol{L}) < \lambda_{d-k}(\boldsymbol{L}).$$

Thus $\lambda_d(\boldsymbol{L}) = \boldsymbol{0}$ for all graph Laplacians $\boldsymbol{L}$, and $\lambda_{d-1}(\boldsymbol{L}) > 0$ if and only if G is connected. In this case, the eigenvector of $\lambda_d(\boldsymbol{L})$ is the all-ones vector $\mathbf{1}$.

This gives us a way to show that a graph is connected using only linear algebra (and thus random matrix theory for random graphs), by showing that $\lambda_{d-1}(\boldsymbol{L}) > 0$. Further, note that if $G \sim \mathcal{G}(d,p)$, then $\boldsymbol{L}_G$ is precisely a sum of independent random matrices of the kind that our tools let us handle. Indeed, it is actually quite similar to the matrix series treated in the NCK inequality and its ilk, though with coefficients that are neither Gaussian nor Rademacher. Rather, for $b_{ij} \sim \mathsf{Ber}(p)$ (i.e., equal to 1 with probability p and 0 with probability $1-p$) independent over all $1 \leq i < j \leq d$, we have

$$\boldsymbol{L}_G \overset{\text{(law)}}{=} \sum_{1 \leq i < j \leq d} b_{ij} (\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top.$$

This is just a way of expressing that every edge is present independent with probability p via the Laplacian matrix. Now it is clear that we can use the matrix Bernstein inequality and learn something about connectedness. Implementing this strategy, we will show the following.

Theorem 6.5.5. *There is a constant $C > 0$ such that, if $p = p(d) \geq C\frac{\log d}{d}$ and $G \sim \mathcal{G}(d, p(d))$, then $\mathbb{P}[G \text{is connected}] \to 1$.*

In fact, this result is sharp up to the value of the constant C: combinatorial analysis implies the following sharper result.

Theorem 6.5.6. *In the setting of Theorem 6.5.5, the following hold for any $\epsilon > 0$:*

1. *If $p(d) \geq (1+\epsilon)\frac{\log d}{d}$, then $\mathbb{P}[G \text{is connected}] \to 1$.*
2. *If $p(d) \leq (1-\epsilon)\frac{\log d}{d}$, then $\mathbb{P}[G \text{is connected}] \to 0$.*

The idea of this proof does not have to do with random matrix theory, but let us mention the main idea: it turns out that whether G is connected or not is determined by whether

G has any isolated vertices or not. At a heuristic level, calculating the expected number of these vertices gives, for small p,

$$\begin{aligned}\mathbb{E}\#\{\text{isolated vertices}\} &= \mathbb{E}\sum_{i=1}^{d} \mathbb{1}\{i \text{ is isolated}\} \\ &= \sum_{i=1}^{d} \mathbb{P}[i \text{ is isolated}] \\ &= d \cdot (1-p)^{d-1} \\ &\approx \exp\left(\log d - p(d-1)\right),\end{aligned}$$

which suggests that the transition in whether these vertices appear or not indeed occurs around $p \approx \frac{\log d}{d}$.

Proof of Theorem 6.5.5. Let us abbreviate $\boldsymbol{v}_{ij} := \boldsymbol{e}_i - b\boldsymbol{e}_j$. Note that $\boldsymbol{L}_G$ is not centered, so we will not be able to use the matrix Bernstein inequality until we address this. We compute

$$\begin{aligned}\mathbb{E}\boldsymbol{L}_G &= \sum_{1\le i<j\le d} \mathbb{E}[b_{ij}]\boldsymbol{v}_{ij}\boldsymbol{v}_{ij}^\top \\ &= p\sum_{1\le i<j\le d} \boldsymbol{v}_{ij}\boldsymbol{v}_{ij}^\top \\ &= p\boldsymbol{L}_{K_d} \\ &= p(\boldsymbol{I}_d - \boldsymbol{1}\boldsymbol{1}^\top),\end{aligned}$$

where we have identified the Laplacian of the complete graph $\boldsymbol{L}_{K_d}$ as appearing in this calculation. In particular, we have $\lambda_d(\mathbb{E}\boldsymbol{L}_G) = 0$ while $\lambda_i(\mathbb{E}\boldsymbol{L}_G) = pd$ for all $1 \le i \le d-1$. Now, the Courant-Fischer min-max theorem implies that

$$\lambda_{d-1}(\boldsymbol{L}_G) \ge \lambda_{d-1}(\mathbb{E}\boldsymbol{L}_G) - \|\boldsymbol{L}_G - \mathbb{E}\boldsymbol{L}_G\| = pd - \|\boldsymbol{L}_G - \mathbb{E}\boldsymbol{L}_G\|.$$

So, it suffices to establish that $\|\boldsymbol{L}_G - \mathbb{E}\boldsymbol{L}_G\| < pd$ with high probability under our assumption. We consider just the expectation here, and leave the small remaining concentration argument as an exercise.

We have

$$\boldsymbol{L}_G - \mathbb{E}\boldsymbol{L}_G = \sum_{1\le i<j\le d} \underbrace{(b_{ij} - \mathbb{E}[b_{ij}])\boldsymbol{v}_{ij}\boldsymbol{v}_{ij}^\top}_{\boldsymbol{H}_{ij}}.$$

As before, we gather the two pieces of information about the $\boldsymbol{H}_{ij}$ that matrix Bernstein demands:

$$\begin{aligned}\|\boldsymbol{H}_{ij}\| &\le \|\boldsymbol{v}_{ij}\|^2 = 2, \\ \mathbb{E}\boldsymbol{H}_{ij}^2 &= \mathsf{Var}[b_{ij}]\|\boldsymbol{v}_{ij}\|^2\boldsymbol{v}_{ij}\boldsymbol{v}_{ij}^\top = 2p(1-p)\boldsymbol{v}_{ij}\boldsymbol{v}_{ij}^\top, \\ \left\|\sum_{i<j}\mathbb{E}\boldsymbol{H}_{ij}^2\right\|^{1/2} &= \|2p(1-p)\boldsymbol{L}_{K_d}\|^{1/2} \lesssim \sqrt{pd}.\end{aligned}$$

Thus, by matrix Bernstein, for some $C > 0$,

$$\mathbb{E}\|\boldsymbol{L}_G - \mathbb{E}\boldsymbol{L}_G\| \le C(\sqrt{\log d} \cdot \sqrt{pd} + \log d)$$

and if $p \ge C' \frac{\log d}{d}$, then

$$\le (C\sqrt{C'} + C) \log d,$$

which for sufficiently large C' will indeed be smaller than $pd \ge C' \log d$ (since the above scales as $\sqrt{C'}$). □

6.6 Application: Spectral Sparsification [SS11]

We now consider a similar application to sparsifying matrices from before, but for graphs. Given G, we want to construct $\hat{G}$ which is "close to" G (in a sense we will clarify below), while having a small number of edges.

What properties of G we want $\hat{G}$ to share depends on applications, but they will usually be some sort of combinatorial graph-theoretic notions. Many computations with graphs, for instance, depend on the sizes of *cuts* in G: we define for any $S \subseteq V(G)$

$$\mathsf{cut}_G(S) := \#\{i, j \in V(G) : i \sim j, i \in S, j \notin S\}.$$

Thus $\mathsf{cut}_G(S)$ is the number of edges that cross the partition of the vertices of G into S and its complement. So, cuts are important for solving problems like community detection that we saw before: one may formulate a version of that as, for instance, the *minimum bisection* problem:

$$\begin{array}{ll} \text{minize} & \mathsf{cut}_G(S) \\ \text{subject to} & |S| = |V(G)|/2. \end{array}$$

Note that we must constrain the size of S, since the minimum cut over all S is of course trivially $\mathsf{cut}_G(\emptyset) = 0$.

Once again, it seems that this is not a linear algebra problem at all, so it is unclear how our study of sparsifying matrices can help us. Fortunately, the sizes of cuts are captured by the graph Laplacian: if we associate to S an indicator vector $\boldsymbol{x} \in \{\pm 1\}^d$ with

$$x_i = \left\{ \begin{array}{ll} +1 & \text{if } i \in S \\ -1 & \text{if } i \notin S \end{array} \right\},$$

then we have

$$\boldsymbol{x}^\top \boldsymbol{L}_G \boldsymbol{x} = \sum_{i \sim j} (x_i - x_j)^2 = \sum_{i \sim j} \left\{ \begin{array}{ll} 0 & \text{if } i, j \in S \text{ or } i, j \notin S, \\ 4 & \text{otherwise} \end{array} \right\} = 4 \cdot \mathsf{cut}_G(S).$$

In this sense, sparsifying G while preserving the sizes of cuts amounts to sparsifying the Laplacian matrix $\boldsymbol{L}_G$.

There are a few wrinkles we must address. Firstly, clearly we cannot simply remove edges from G to obtain $\hat{G}$, since then all cuts will get uniformly smaller. Instead, we must consider

$\widehat{G}$ a *weighted* graph, with associated notions of weighted cuts and weighted Laplacians. In general, suppose $\widehat{G}$ is a weighted graph with an associated weight function $w : [d]^2 \to \mathbb{R}_{\geq 0}$. We make the convention that if $\{i, j\}$ is not an edge in $\widehat{G}$, then we set $w(i, j) = 0$. The natural definition of the Laplacian is then

$$(\boldsymbol{L}_{\widehat{G}})_{ij} := \left\{ \begin{array}{ll} \sum_{j \sim i} w(i,j) & \text{if } i = j, \\ -w(i,j) & \text{if } i \neq j \end{array} \right\},$$

which is the same as

$$\boldsymbol{L}_{\widehat{G}} = \sum_{i,j} w(i,j) \boldsymbol{v}_{ij} \boldsymbol{v}_{ij}^\top.$$

Correspondingly, if we define weighted cuts in $\widehat{G}$ as

$$\mathsf{cut}_{\widehat{G}}(S) := \sum_{i \in S, j \notin S} w(i,j),$$

then we still have

$$\mathsf{cut}_{\widehat{G}}(S) = \frac{1}{4} \boldsymbol{x}^\top \boldsymbol{L}_{\widehat{G}} \boldsymbol{x}.$$

More important, we cannot just use our previous approach to sparsify $\boldsymbol{L}_G$. That would give us a sparsified $\boldsymbol{L}_{\widehat{G}}$ satisfying a guarantee of the form

$$\frac{\|\boldsymbol{L}_G - \boldsymbol{L}_{\widehat{G}}\|}{\|\boldsymbol{L}_G\|} \leq \epsilon.$$

What does such a guarantee tell us about the approximation of cut sizes? We may infer

$$\begin{aligned} |\mathsf{cut}_{\widehat{G}}(S) - \mathsf{cut}_G(S)| &= \frac{1}{4} \left| \boldsymbol{x}^\top \boldsymbol{L}_{\widehat{G}} \boldsymbol{x} - \boldsymbol{x}^\top \boldsymbol{L} \boldsymbol{x} \right| \\ &\leq \frac{1}{4} \|\boldsymbol{L}_{\widehat{G}} - \boldsymbol{L}_G\| \cdot \|\boldsymbol{x}\|^2 \\ &\leq \frac{d}{4} \epsilon \|\boldsymbol{L}_G\|. \end{aligned}$$

But on the other hand, we have

$$\|\boldsymbol{L}_G\| \geq \max_{\boldsymbol{x} \in \{\pm 1/\sqrt{d}\}^d} \boldsymbol{x}^\top \boldsymbol{L}_G \boldsymbol{x} = \frac{4}{d} \mathsf{maxcut}(G),$$

where we define

$$\mathsf{maxcut}(G) = \max_{S \subseteq V(G)} \mathsf{cut}_G(S).$$

So, the above guarantee is no better than

$$|\mathsf{cut}_{\widehat{G}}(S) - \mathsf{cut}_G(S)| \leq \epsilon \cdot \mathsf{maxcut}(G).$$

This might be fine if we only care about querying the sizes of *large* cuts in G and receiving good estimates. However, as mentioned above in the context of community detection, in some important situations it is actually the *small* cuts in G that we care about.

The above calculations suggest that we should instead aim for a guarantee of the form

$$|\mathsf{cut}_{\hat{G}}(S) - \mathsf{cut}_G(S)| \leq \epsilon \cdot \mathsf{cut}_G(S),$$

or equivalently

$$(1-\epsilon)\mathsf{cut}_G(S) \leq \mathsf{cut}_{\hat{G}}(S) \leq (1+\epsilon)\mathsf{cut}_G(S),$$

for all $S \subseteq V(G)$. By our observations about quadratic forms, this would be implied by the following form of *spectral approximation* of the Laplacian:

$$(1-\epsilon)\boldsymbol{L}_G \preceq \boldsymbol{L}_{\hat{G}} \preceq (1+\epsilon)\boldsymbol{L}_G. \tag{6.6.1}$$

(Indeed, this is a stronger guarantee, ensuring that *any* quadratic form with the Laplacian, or from our previous intuition any "squared gradient norm" query, is preserved to within a factor $1 \pm \epsilon$.)

We will show the following remarkable result:

Theorem 6.6.1 ([SS11])**.** *For any graph G and any $\epsilon > 0$, there is a weighted $\hat{G}$ with at most $O(d \log d/\epsilon^2)$ edges such that (6.6.1), which may be found efficiently by a randomized algorithm.*

Proof. For the sake of brevity, we will avoid dealing carefully with the issue that graph Laplacians are not invertible. Really, whenever we write $\boldsymbol{L}^{-1}$ below, this should be replaced with the Moore-Penrose pseudoinverse, and this will cause some minor changes in the calculations as well. However, the main ideas are precisely as we outline below.

The first issue to address is that (6.6.1) is not a statement about operator norm bounds, whereby matrix Bernstein is not directly helpful. We may remedy this by rewriting it as

$$(1-\epsilon)\boldsymbol{I}_d \preceq \boldsymbol{L}_G^{-1/2}\boldsymbol{L}_{\hat{G}}\boldsymbol{L}_G^{-1/2} \preceq (1+\epsilon)\boldsymbol{I}_d,$$

or equivalently

$$||\boldsymbol{L}_G^{-1/2}(\boldsymbol{L}_{\hat{G}} - \boldsymbol{L}_G)\boldsymbol{L}_G^{-1/2}|| = ||\boldsymbol{L}_G^{-1/2}\boldsymbol{L}_{\hat{G}}\boldsymbol{L}_G^{-1/2} - \boldsymbol{I}_d|| \leq \epsilon.$$

This is the kind of result matrix Bernstein can show, provided that $\boldsymbol{L}_{\hat{G}}$ is a sum of independent random matrices with $\mathbb{E}\boldsymbol{L}_{\hat{G}} = \boldsymbol{L}_G$.

We proceed as we did for sparsifying matrices (or, more precisely, as we proposed for randomized matrix multiplication). We have

$$\boldsymbol{L}_G = \sum_{i \sim j} \boldsymbol{v}_{ij}\boldsymbol{v}_{ij}^\top.$$

So, let us choose a probability distribution p_{ij} on the edges of G (i.e., on pairs with $i \sim j$), and define $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_m$ i.i.d. random matrices taking value $\frac{1}{p_{ij}}\boldsymbol{v}_{ij}\boldsymbol{v}_{ij}^\top$ with probability p_{ij}. We then have $\mathbb{E}\boldsymbol{F}_a = \boldsymbol{L}_G$, and so we may set

$$\boldsymbol{L}_{\hat{G}} := \frac{1}{m}\sum_{a=1}^{m} \boldsymbol{F}_a.$$

Note that this indeed gives the Laplacian of a weighted graph, whose weights we may calculate by grouping the $\boldsymbol{F}_a$ that take each value $\boldsymbol{v}_{ij}$ and adding together their contributions

$\frac{1}{p_{ij}m}$ to the total weight of that edge. And, this underlying weighted graph $\hat{G}$ has at most m edges.

Thus, we may write

$$\mathbb{E}\left\|\boldsymbol{L}_G^{-1/2}(\boldsymbol{L}_{\hat{G}}-\boldsymbol{L}_G)\boldsymbol{L}_G^{-1/2}\right\| = \mathbb{E}\left\|\sum_{a=1}^{m}\underbrace{\frac{1}{m}\boldsymbol{L}_G^{-1/2}(\boldsymbol{F}_a-\mathbb{E}\boldsymbol{F}_a)\boldsymbol{L}_G^{-1/2}}_{\boldsymbol{H}_a}\right\|.$$

It will be helpful to define

$$\widehat{\boldsymbol{F}}_a := \boldsymbol{L}_G^{-1/2}\boldsymbol{F}_a\boldsymbol{L}_G^{-1/2},$$

so that

$$\boldsymbol{H}_a = \frac{1}{m}(\widehat{\boldsymbol{F}}_a - \mathbb{E}\widehat{\boldsymbol{F}}_a) = \frac{1}{m}(\widehat{\boldsymbol{F}}_a - \boldsymbol{I}_d).$$

These matrices take the values $\frac{1}{p_{ij}}(\boldsymbol{L}_G^{-1/2}\boldsymbol{v}_{ij})(\boldsymbol{L}_G^{-1/2}\boldsymbol{v}_{ij})^\top$ with probability p_{ij}. Let us also define

$$\boldsymbol{w}_{ij} := \boldsymbol{L}_G^{-1/2}\boldsymbol{v}_{ij}.$$

The elegant choice of the weights p_{ij} is just to make $\|\widehat{\boldsymbol{F}}_a\| = \frac{1}{p_{ij}}\|\boldsymbol{w}_{ij}\|^2$ always take the same value. This amounts to taking

$$p_{ij} \propto \|\boldsymbol{w}_{ij}\|^2 = \boldsymbol{v}_{ij}^\top\boldsymbol{L}_G^{-1}\boldsymbol{v}_{ij}.$$

Further, the normalizing constant is very simple: the sum of these weights is

$$\begin{aligned}
\sum_{i\sim j}\|\boldsymbol{w}_{ij}\|^2 &= \sum_{i\sim j}\boldsymbol{v}_{ij}^\top\boldsymbol{L}_G^{-1}\boldsymbol{v}_{ij}\\
&= \sum_{i\sim j}\langle\boldsymbol{L}_G^{-1}, \boldsymbol{v}_{ij}\boldsymbol{v}_{ij}^\top\rangle\\
&= \left\langle\boldsymbol{L}_G^{-1}, \sum_{i\sim j}\boldsymbol{v}_{ij}\boldsymbol{v}_{ij}^\top\right\rangle\\
&= \langle\boldsymbol{L}_G^{-1}, \boldsymbol{L}_G\rangle\\
&= \mathsf{Tr}(\boldsymbol{L}_G^{-1}\boldsymbol{L}_G)\\
&= \mathsf{Tr}(\boldsymbol{I}_d)\\
&= d.
\end{aligned}$$

(Note that this calculation will change a little bit and depend on the number of connected components in G if we try to be more careful about inverting Laplacians.) Thus we simply have $p_{ij} = \|\boldsymbol{w}_{ij}\|^2/d$, and $\|\widehat{\boldsymbol{F}}_a\| = d$ always. In particular, if we define the unit vectors

$$\widehat{\boldsymbol{w}}_{ij} := \frac{\boldsymbol{w}_{ij}}{\|\boldsymbol{w}_{ij}\|},$$

then we will have $\widehat{\boldsymbol{F}}_a = d\widehat{\boldsymbol{w}}_{ij}\widehat{\boldsymbol{w}}_{ij}^\top$ with probability $\|\boldsymbol{w}\|_{ij}^2/d$. It is then clear that $\widehat{\boldsymbol{F}}_a^2 = d\widehat{\boldsymbol{F}}_a$, since $\widehat{\boldsymbol{F}}_a$ equals d times a (rank one) projection matrix.

With this choice made, we prepare to apply matrix Bernstein with the same calculations as always:

$$\begin{aligned}
\|\boldsymbol{H}_a\| &= \frac{1}{m}\|\widehat{\boldsymbol{F}}_a - \mathbb{E}\widehat{\boldsymbol{F}}_a\| \\
&\leq \frac{2d}{m}, \\
\mathbb{E}\boldsymbol{H}_a^2 &\preceq \frac{1}{m^2}\mathbb{E}\widehat{\boldsymbol{F}}_a^2 \\
&= \frac{d}{m^2}\mathbb{E}\widehat{\boldsymbol{F}}_a \\
&= \frac{d}{m^2}\boldsymbol{I}_d, \\
\left\|\sum_{a=1}^{m}\mathbb{E}\boldsymbol{H}_a^2\right\|^{1/2} &= \sqrt{\frac{d}{m}}.
\end{aligned}$$

Thus, matrix Bernstein implies

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{L}_G^{-1/2}(\boldsymbol{L}_{\hat{G}} - \boldsymbol{L}_G)\boldsymbol{L}_G^{-1/2}\| &\lesssim \sqrt{\log d}\cdot\sqrt{\frac{d}{m}} + \log d\cdot\frac{d}{m} \\
&= \sqrt{\frac{d\log d}{m}} + \frac{d\log d}{m}
\end{aligned}$$

and if $m \gtrsim d\log d/\epsilon^2$, we can ensure

$$\lesssim \epsilon + \epsilon^2,$$

and choosing the constants appropriately gives the result. □

6.6.1 Effective Resistance Interpretation

The choice of weights

$$p_{ij} \propto \|\boldsymbol{w}_{ij}\|^2 = (\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{L}_G^{-1}(\boldsymbol{e}_i - \boldsymbol{e}_j) =: R_{ij}$$

works beautifully with the algebraic details of matrix Bernstein in the above proof. But what is its intuitive meaning? In what sense are the edges with larger p_{ij} that our sparsification is more likely to include and weigh highly in $\hat{G}$ the ones that are "important" to the structure of G?

This question has an even more beautiful answer: the number R_{ij} is the *effective resistance* between i and j in G. Consider a physical network of resistors, each with one unit of resistance (say one Ohm), built according to G. Put a voltage difference of one unit (one volt) across i and j: set an electric potential to equal $V(i) = 1$ and $V(j) = 0$ (say by attaching the poles of a one-volt battery to i and j). This will cause some amount of current I (in amperes) to flow out of i and into j. At every other vertex, by Kirchoff's current rule the total amount of current entering will equal the total amount exiting, but i will be a source

of an amount I of current, and j will be a sink of the same amount. Now, view the entire circuit as a black-box apparatus between i and j. This object has an "effective resistance" describing the amount of resistance it "looks like" it has, which is just:

$$R_{ij} := \frac{V(i) - V(j)}{I} = \frac{1}{I}$$

by Ohm's rule ("$V = IR$").

How does this quantity behave? If the only path between i and j is across the edge that joins them, then all current from i to j must flow along that edge, and you may calculate that in fact $R_{ij} = 1$: the effective resistance is just the resistance of the single resistor joining i and j. However, if there are more paths from i to j, then the above black-box apparatus essentially behaves as several resistors in parallel. As you likely know (and as is intuitive since current has several "choices" of how to flow), resistors placed in parallel have lower resistance than any individual one of the resistors, and thus $R_{ij} < 1$. So, the choice of weights proportional to effective resistance precisely makes it more likely to include the "bottlenecks" of G in $\widehat{G}$ while ignoring the edges $\{i, j\}$ for which there are many redundant paths connecting i and j.

Let us briefly sketch the simple argument that this definition in fact coincides with our definition of the p_{ij}, which is not algebraically obvious. We will still use a bit of physical reasoning.

Theorem 6.6.2. *The effective resistance R_{ij} described above equals $(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{L}_G^{-1}(\boldsymbol{e}_i - \boldsymbol{e}_j)$.*

Proof. The setup above implicitly assigns a potential $V(x)$ to every $x \in [d]$ a vertex of G. These satisfy $V(i) = 1$ and $V(j) = 0$ by our choice. For every adjacent pair $\{x, y\}$, there is also an amount of current $I(x, y)$ flowing along the resistor between x and y. This is a directed quantity, so that $I(y, x) = -I(x, y)$.

Kirchoff's rule implies that every vertex other than i and j is neither a source or a sink of current: if $x \notin \{i, j\}$, then

$$\sum_{y \sim x} I(x, y) = 0.$$

On the other hand, by Ohm's rule, since the resistor between x and y has resistance of one unit, $I(x, y) = V(y) - V(x)$. Thus,

$$\sum_{y \sim x} (V(y) - V(x)) = 0,$$

which we may rewrite as

$$V(x) = \frac{1}{\deg(x)} \sum_{y \sim x} V(y).$$

In words, at every vertex x other than i and j, the potential at x is the average of the potentials of the neighbors of x. This is called V being a *harmonic function* on all vertices other than i and j.

Let us write $\boldsymbol{v} \in \mathbb{R}^d$ for the vector of potentials at each vertex. The above may be written as $(\boldsymbol{D}\boldsymbol{v})_x = (\boldsymbol{A}\boldsymbol{v})_x$ for all $x \notin \{i, j\}$ for $\boldsymbol{D}$ the diagonal matrix of degrees in G and $\boldsymbol{A}$ the adjacency matrix. In other words, $(\boldsymbol{L}\boldsymbol{v})_x = 0$ for all $x \notin \{i, j\}$ for $\boldsymbol{L}$ the Laplacian. (This is

parallel to the definition of harmonic functions you might have seen in analysis as functions whose Laplacian is zero aside from some boundary conditions.)

On the other hand, letting $R = R_{ij}$ be the effective resistance, then the amount of current flowing out of i or into j is $1/R$. Rewriting this similarly to the above gives that $(\boldsymbol{L}\boldsymbol{v})_i = 1/R$ and $(\boldsymbol{L}\boldsymbol{v})_j = -1/R$. In summary then,

$$\boldsymbol{L}\boldsymbol{v} = \frac{1}{R}(\boldsymbol{e}_i - \boldsymbol{e}_j),$$

or

$$\boldsymbol{v} = \frac{1}{R}\boldsymbol{L}^{-1}(\boldsymbol{e}_i - \boldsymbol{e}_j).$$

Finally, we have

$$1 = v_i - v_j = \frac{1}{R}(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{L}^{-1}(\boldsymbol{e}_i - \boldsymbol{e}_j),$$

and rearranging gives the result. □

This is only the beginning of a deep theory relating electrical networks to properties of graphs and random walks on graphs. A good place to learn more about this is Chapter 2 of the textbook [LP17].

Bibliography

[ABBN04] Shiri Artstein, Keith M Ball, Franck Barthe, and Assaf Naor. On the rate of convergence in the entropic central limit theorem. *Probability theory and related fields*, 129(3):381–390, 2004.

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

[Ach01] Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281, 2001.

[Ans99] Michael Anshelevich. The linearization of the central limit operator in free probability theory. *Probability theory and related fields*, 115:401–416, 1999.

[ARV09] Sanjeev Arora, Satish Rao, and Umesh Vazirani. Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM (JACM)*, 56(2):1–37, 2009.

[Bar86] Andrew R Barron. Entropy and the central limit theorem. *The Annals of probability*, pages 336–342, 1986.

[BBAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[Bou85] Jean Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel Journal of Mathematics*, 52:46–52, 1985.

[EK08] Noureddine El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. 2008.

[ES63] Paul Erdos and Horst Sachs. Reguläre graphen gegebener taillenweite mit minimaler knotenzahl. *Wiss. Z. Martin-Luther-Univ. Halle-Wittenberg Math.-Natur. Reihe*, 12(251-257):22, 1963.

[FP07] Delphine Féral and Sandrine Péché. The largest eigenvalue of rank one deformation of large Wigner matrices. *Communications in Mathematical Physics*, 272(1):185–228, 2007.

[HL13] Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):1–39, 2013.

[HLW06] Shlomo Hoory, Nathan Linial, and Avi Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

[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.

[KS07] Leonid Koralov and Yakov G Sinai. *Theory of probability and random processes*. Springer Science & Business Media, 2007.

[KXZ16] Florent Krzakala, Jiaming Xu, and Lenka Zdeborová. Mutual information in rank-one matrix estimation. In *2016 IEEE Information Theory Workshop (ITW)*, pages 71–75. IEEE, 2016.

[KY13] Antti Knowles and Jun Yin. The isotropic semicircle law and deformation of wigner matrices. *Communications on Pure and Applied Mathematics*, 66(11):1663–1749, 2013.

[Lin59] Ju V Linnik. An information-theoretic proof of the central limit theorem with Lindeberg conditions. *Theory of Probability & Its Applications*, 4(3):288–299, 1959.

[LKZ15] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In *53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton 2015)*, pages 680–687. IEEE, 2015.

[LM23] Iris Stephanie Arenas Longoria and James A Mingo. Freely independent coin tosses, standard Young tableaux, and the Kesten–McKay law. *American Mathematical Monthly*, 130(1):35–48, 2023.

[LP86] Françoise Lust-Piquard. Inégalités de khintchine dans c^p $(1 < p < \infty)$. *CR Acad. Sci. Paris*, 303:289–292, 1986.

[LP17] Russell Lyons and Yuval Peres. *Probability on trees and networks*, volume 42. Cambridge University Press, 2017.

[LPP91] Françoise Lust-Piquard and Gilles Pisier. Non commutative khintchine and paley inequalities. *Arkiv för matematik*, 29(1):241–260, 1991.

[LPS88] Alexander Lubotzky, Ralph Phillips, and Peter Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988.

[LS21] Nati Linial and Michael Simkin. A randomized construction of high girth regular graphs. *Random Structures & Algorithms*, 58(2):345–369, 2021.

[LSS13] Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood: approximating kernel expansions in loglinear time. In *Proceedings of the international conference on machine learning*, volume 85, 2013.

[Ott23] Sébastien Ott. A note on the renormalization group approach to the central limit theorem. *arXiv preprint arXiv:2303.13905*, 2023.

[Pau07] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007.

[PB17] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International conference on machine learning*, pages 2798–2806. PMLR, 2017.

[PB20] Marc Potters and Jean-Philippe Bouchaud. *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists*. Cambridge University Press, 2020.

[PWBM16] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of pca for spiked random matrices and synchronization. *arXiv preprint arXiv:1609.05573*, 2016.

[PWBM18] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of PCA I: Spiked random matrix models. *Annals of Statistics*, 46(5):2416–2451, 2018.

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

[RR07] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[RV13] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[Sha13] Cosma Shalizi. Advanced data analysis from an elementary point of view, 2013.

[SS11] Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

[Tro15] Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[Tro18] Joel A Tropp. Second-order matrix concentration inequalities. *Applied and Computational Harmonic Analysis*, 44(3):700–736, 2018.

[vH17] Ramon van Handel. Structured random matrices. *Convexity and concentration*, pages 107–156, 2017.

[Wor99] Nicholas C Wormald. Models of random regular graphs. *London Mathematical Society Lecture Note Series*, pages 239–298, 1999.