Random Matrix Theory in Data Science and Statistics

Dmitriy (Tim) Kunisky

Fall 2024 (Last Updated: September 12, 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 (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 1.2 (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 1.3. *The* operator norm *of a matrix* $\boldsymbol{Y}$ *is* $\|\boldsymbol{Y}\| := \sigma_1(\boldsymbol{Y})$.

Proposition 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 1.5 (Variational form of operator nom)**.** $\|\boldsymbol{Y}\| = \sup_{\boldsymbol{v}\neq\boldsymbol{0}} \|\boldsymbol{Y}\boldsymbol{v}\|/\|\boldsymbol{v}\|$.

Theorem 1.6 (Eckart-Young-Mirsky for operator norm)**.** *For any* $\boldsymbol{Y} \in \mathbb{R}^{m\times n}$ *and* $1 \le d < m$,

$$\left\{ \begin{array}{ll} \textit{minimize} & \|\boldsymbol{Y} - \boldsymbol{Z}\| \\ \textit{subject to} & \boldsymbol{Z} \in \mathbb{R}^{m\times n}, \\ & \mathsf{rank}(\boldsymbol{Z}) \le d \end{array} \right\} = \sigma_{d+1}(\boldsymbol{Y}), \tag{1.1}$$

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:

$$\|\boldsymbol{Y} - \boldsymbol{Z}\| = \left\| \sum_{i=d+1}^{\min\{m,n\}} \sigma_i(\boldsymbol{Y}) \boldsymbol{u}_i \boldsymbol{v}_i^\top \right\| = \sigma_{d+1}, \tag{1.2}$$

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:

$$\begin{aligned} \|(\boldsymbol{Y} - \boldsymbol{Z})\boldsymbol{v}\| = \|\boldsymbol{Y}\boldsymbol{v}\| &= \left\| \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \right) \left(\sum_{j=1}^{d+1} \alpha_j \boldsymbol{v}_j \right) \right\| \\ &= \left\| \sum_{i=1}^{\min\{m,n\}} \sum_{j=1}^{d+1} \sigma_i \alpha_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle \boldsymbol{u}_i \right\| \\ &= \left\| \sum_{j=1}^{d+1} \sigma_j \alpha_j \boldsymbol{u}_j \right\| && \text{(orthonormality of } \boldsymbol{v}_j\text{)} \\ &= \sqrt{\sum_{j=1}^{d+1} \sigma_j^2 \alpha_j^2} && \text{(orthonormality of } \boldsymbol{u}_j\text{)} \\ &\ge \sigma_{d+1} \sqrt{\sum_{j=1}^{d+1} \alpha_j^2} \\ &= \sigma_{d+1} \|\boldsymbol{v}\|, \end{aligned} \tag{1.3}$$

and thus $\|\boldsymbol{Y} - \boldsymbol{Z}\| \ge \sigma_{d+1}$ by Proposition 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 1.7. *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 1.8. *The Frobenius norm is indeed a norm.*

Theorem 1.9 (Eckart-Young-Mirsky for Frobenius norm)**.** *For any $\boldsymbol{Y} \in \mathbb{R}^{m\times n}$ and $1 \le d < m$,*

$$\left\{ \begin{array}{ll} \textit{minimize} & \|\boldsymbol{Y} - \boldsymbol{Z}\|_F \\ \textit{subject to} & \boldsymbol{Z} \in \mathbb{R}^{m\times n}, \\ & \mathsf{rank}(\boldsymbol{Z}) \le d \end{array} \right\} = \left(\sum_{i=d+1}^{\min\{m,n\}} \sigma_i(\boldsymbol{Y})^2 \right)^{1/2}, \tag{1.4}$$

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 $\boldsymbol{v}_1, \dots, \boldsymbol{v}_1$ 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.

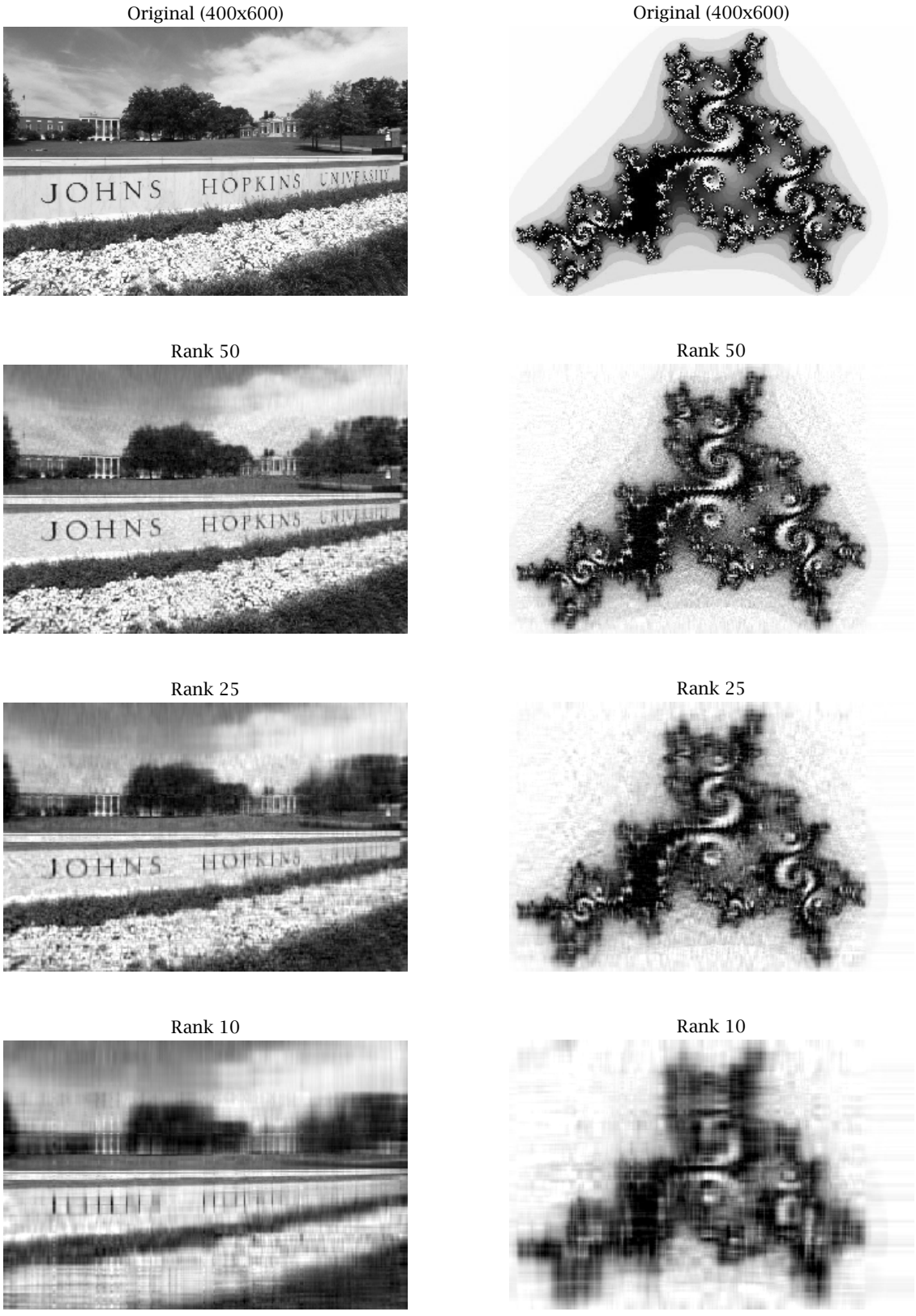

Figure 1.1: Two examples of image compression using truncated singular value decomposition, showing the deterioration of the approximation as the rank decreases.

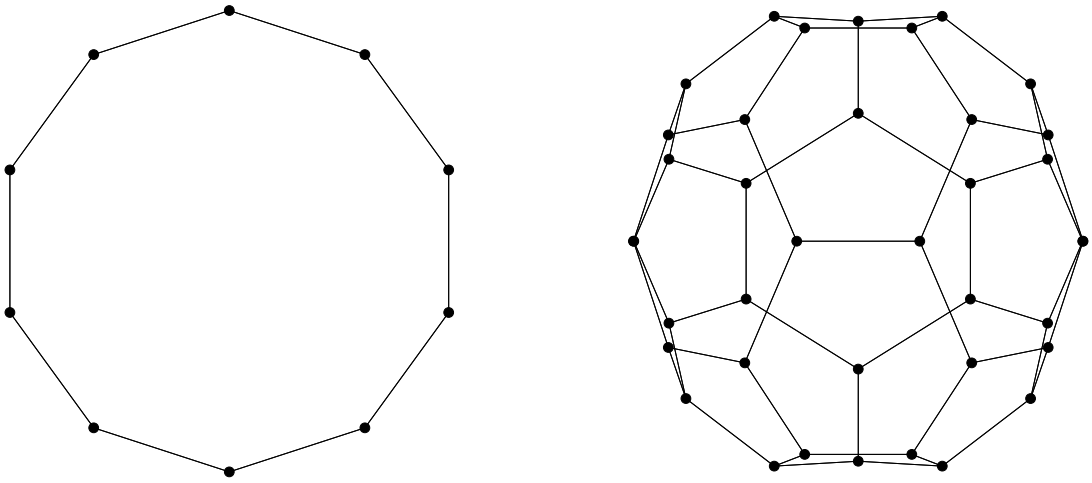

Figure 1.2: Two examples of two-dimensional graph drawings obtained by embedding using eigenvectors. The first is a cycle, while the second is a so-called *fullerene*, a graph describing a three-dimensional structure that can be formed by carbon atoms.

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

$$\boldsymbol{y}_i \approx \sum_{j=1}^{d} x_{ij}\boldsymbol{w}_j = \boldsymbol{W}\boldsymbol{x}_i \text{ for each } i \in [n], \tag{1.5}$$

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

$$\boldsymbol{Y} \approx \boldsymbol{W}\boldsymbol{X}, \tag{1.6}$$

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

$$\|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{X}\|_F^2 = \sum_{i=1}^{n} \left\| \boldsymbol{y}_i - \sum_{j=1}^{d} x_{ij} \boldsymbol{w}_j \right\|^2, \tag{1.7}$$

then we will end up with precisely one of the variational problems that leads to truncating the SVD, per Theorem 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 1.10. *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 1.11. *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*

$$\mathsf{Law}(\boldsymbol{A}\boldsymbol{v}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}^\top \boldsymbol{\Sigma} \boldsymbol{A}). \tag{1.8}$$

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:

$$\mathbb{E}[\boldsymbol{G}\boldsymbol{y}] = \mathbb{E}[\boldsymbol{G}]\boldsymbol{y} = \boldsymbol{0}, \tag{1.9}$$

$$\begin{aligned} \mathsf{Cov}((\boldsymbol{G}\boldsymbol{y})_i, (\boldsymbol{G}\boldsymbol{y})_j) &= \mathbb{E}(\boldsymbol{G}\boldsymbol{y})_i (\boldsymbol{G}\boldsymbol{y})_j \\ &= \mathbb{E}\left(\sum_{a=1}^n G_{ia} y_a\right)\left(\sum_{b=1}^n G_{jb} y_b\right) \\ &= \left\{ \begin{array}{ll} 0 & \text{if } i \neq j, \\ \|\boldsymbol{y}\|^2 & \text{if } i = j \end{array} \right\}. \end{aligned} \tag{1.10}$$

In short, we have

$$\mathsf{Law}(\boldsymbol{G}\boldsymbol{y}) = \mathcal{N}(\boldsymbol{0}, \|\boldsymbol{y}\|^2 \boldsymbol{I}_d). \tag{1.11}$$

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

$$\mathbb{E}\|\boldsymbol{G}\boldsymbol{y}\|^2 = \mathsf{Tr}\,\mathsf{Cov}(\boldsymbol{G}\boldsymbol{y}) = d\|\boldsymbol{y}\|^2. \tag{1.12}$$

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

$$\widehat{\boldsymbol{G}} := \frac{1}{\sqrt{d}} \boldsymbol{G}. \tag{1.13}$$

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

$$\mathbb{E}\|\widehat{\boldsymbol{G}}\boldsymbol{y}\|^2 = \|\boldsymbol{y}\|^2. \tag{1.14}$$

What about angles? By a similar calculation you can find

$$\mathbb{E}\langle \widehat{\boldsymbol{G}}\boldsymbol{y}_1, \widehat{\boldsymbol{G}}\boldsymbol{y}_2 \rangle = \langle \boldsymbol{y}_1, \boldsymbol{y}_2 \rangle, \tag{1.15}$$

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

$$\mathbb{E}(\widehat{\boldsymbol{G}}\boldsymbol{Y})^\top (\widehat{\boldsymbol{G}}\boldsymbol{Y}) = \boldsymbol{Y}^\top \boldsymbol{Y}. \tag{1.16}$$

Remark 1.12. *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 and Nearest Neighbors

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 $\boldsymbol{y}_i$ (see Homework 1), but is an intuitive notion and is relevant to various applications.

Definition 1.13. *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 $i, j \in [n]$,*

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

[4] My own non-standard terminology.

Theorem 1.14 (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*

$$d := C\frac{\log n}{\epsilon^2}. \tag{1.17}$$

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

$$\mathbb{P}\left[\left|\|\boldsymbol{g}\|^2 - d\right| \geq t\right] = \mathbb{P}\left[\left|\sum_{i=1}^d g_i^2 - d\right| \geq t\right] \leq 2\left\{\begin{array}{ll} \exp\left(-\frac{t^2}{8d}\right) & \text{if } t \leq d, \\ \exp\left(-\frac{t}{8}\right) & \text{if } t \geq d \end{array}\right\}. \tag{1.18}$$

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

$$\psi(\lambda) := \log \mathbb{E}\exp(\lambda x_1) = \log \mathbb{E}\exp(\lambda(g_1^2 - 1)), \tag{1.19}$$

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

$$\mathbb{E}\exp(\lambda g_1^2) = \frac{1}{\sqrt{1-2\lambda}}, \tag{1.20}$$

whereby

$$\psi(\lambda) = -\lambda + \frac{1}{2}\log\left(\frac{1}{1-2\lambda}\right). \tag{1.21}$$

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

Thus we have

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

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

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

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

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

Remark 1.16. *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 $\lambda \ge \frac{1}{2}$ 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:

$$\text{“ } \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d) \approx \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d})) \text{ ”} \tag{1.22}$$

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

Theorem 1.17 (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 $k \ge 1$, $(\sqrt{d}\, x_1, \dots, \sqrt{d}\, x_k)$ converges weakly in law to $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$.*

1.3.2 Proof of Theorem 1.14

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

Proof of Theorem 1.14. 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 1.15 we have

$$\leq 2\exp(-c\epsilon^2 d). \tag{1.23}$$

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

$$\mathop{\mathbb{P}}_{\hat{G}\sim\mathcal{N}(0,\frac{1}{d})^{\otimes d\times m}}\left[(1-\epsilon)\|\boldsymbol{y}\| \leq \|\widehat{\boldsymbol{G}}\boldsymbol{y}\| \leq (1+\epsilon)\|\boldsymbol{y}\|\right] \geq 1 - 4\exp(-c\epsilon^2 d). \tag{1.24}$$

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

$$\begin{aligned}
\mathbb{P}[\text{some } E_{ij} \text{ does not occur}] &\leq \binom{n}{2}\cdot 2\exp(-c\epsilon^2 d) \\
&\leq n^2\exp(-c\epsilon^2 d) \\
&= \exp(2\log n - c\epsilon^2 d),
\end{aligned} \tag{1.25}$$

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

Remark 1.18. *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

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

$$P_{ij} \sim \left\{ \begin{array}{ll} \mathcal{N}(0, \frac{1}{qd}) & \text{with probability } q, \\ 0 & \text{with probability } 1-q \end{array} \right\}. \tag{1.26}$$

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

$$\begin{aligned} \mathsf{Var}[\|\boldsymbol{P}\boldsymbol{e}_1\|^2] &= \mathsf{Var}\left[\sum_{i=1}^{d} P_{1d}^2\right] \\ &= d\,\mathsf{Var}[P_{11}^2] && \text{(independence)} \\ &= d(\mathbb{E}P_{11}^4 - (\mathbb{E}P_{11}^2)^2) \\ &= d\left(\frac{3}{q^2d^2} - \frac{1}{d^2}\right) \\ &= \frac{1}{d}\left(\frac{3}{q^2} - 1\right). \end{aligned} \tag{1.27}$$

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

$$q = \Theta\left(\frac{\log^2 n}{m}\right). \tag{1.28}$$

With this, we end up with a projection matrix $\boldsymbol{PHD}$ which achieves the same guarantee as in Theorem 1.14 (with constant ϵ), multiplication by which requires time

$$O\Big(\underbrace{\frac{\log^2 n}{m} \cdot m \log n}_{\text{for } \boldsymbol{P}} + \underbrace{m \log m}_{\text{for } \boldsymbol{H}} + \underbrace{m}_{\text{for } \boldsymbol{D}}\Big) = O\left(\log^3 n + m \log m\right), \tag{1.29}$$

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

$$\max_{i,j \in [n]} \frac{\big| \|f(\boldsymbol{y}_i) - f(\boldsymbol{y}_j)\| - \|\boldsymbol{y}_i - \boldsymbol{y}_j\| \big|}{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|}. \tag{1.30}$$

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:

$$\hat{\boldsymbol{a}} := \arg\min_{\boldsymbol{a}} \sum_{i=1}^{n} (\langle \boldsymbol{a}, \boldsymbol{x}_i \rangle - y_i)^2. \tag{1.31}$$

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

$$\hat{\boldsymbol{a}} = \widehat{\boldsymbol{C}}^{-1} \sum_{i=1}^{n} y_i \boldsymbol{x}_i. \tag{1.32}$$

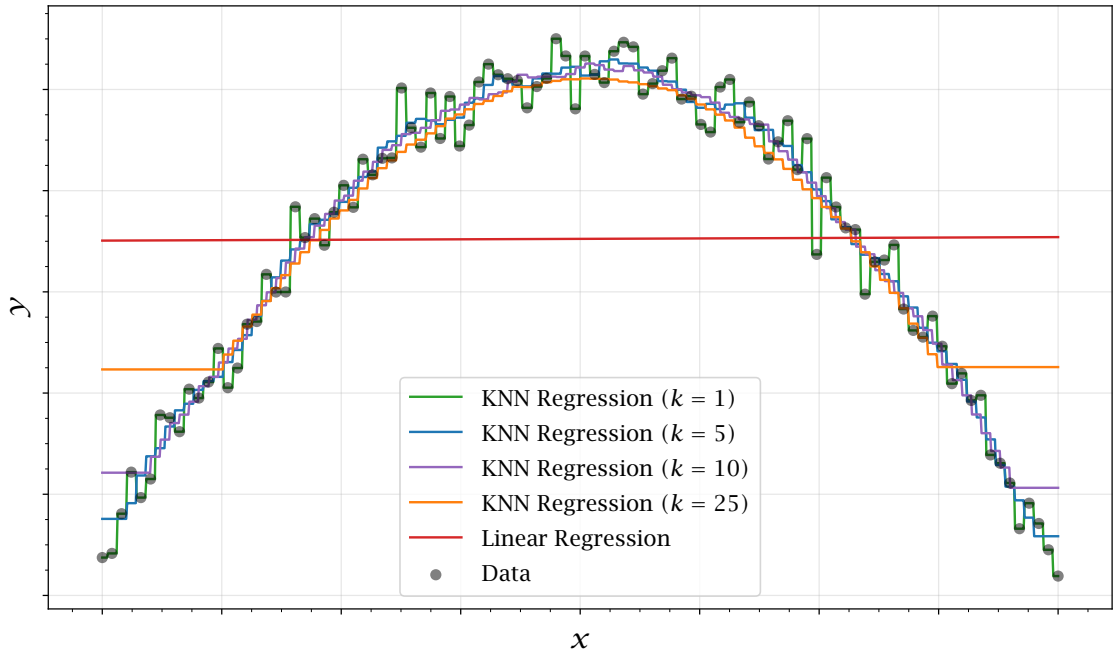

Figure 1.3: An example of linear regression and nearest neighbors regression fits to evaluations of a quadratic function with noise added.

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

$$f(\boldsymbol{x}) = \langle \hat{\boldsymbol{a}}, \boldsymbol{x} \rangle = \sum_{i=1}^{n} \langle \widehat{\boldsymbol{C}}^{-1/2} \boldsymbol{x}_i, \widehat{\boldsymbol{C}}^{-1/2} \boldsymbol{x} \rangle y_i. \tag{1.33}$$

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

$$\widehat{w}(\boldsymbol{x}_i, \boldsymbol{x}) := \langle \widehat{\boldsymbol{C}}^{-1/2} \boldsymbol{x}_i, \widehat{\boldsymbol{C}}^{-1/2} \boldsymbol{x} \rangle. \tag{1.34}$$

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

$$\widehat{w}(\boldsymbol{x}_i, \boldsymbol{x}) := \frac{1}{k} \mathbb{1}\{\boldsymbol{x}_i \text{ one of the } k \text{ closest training points to } \boldsymbol{x}\}. \tag{1.35}$$

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 1.19 (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 1.20. *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 $\boldsymbol{G} \sim \mathcal{N}(0, \frac{1}{d})^{\otimes d \times m}$,*

$$\mathbb{P}\left[\, \left| \|\boldsymbol{G}\boldsymbol{x}_i\|^2 - \|\boldsymbol{x}_i\|^2 \right| \leq \epsilon \|\boldsymbol{x}_i\|^2 \text{ for all } i \in [N] \right] \geq 1 - \frac{1}{N}. \tag{1.36}$$

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

$$\left| \|\boldsymbol{G}\boldsymbol{x}_i\|^2 - \|\boldsymbol{x}_i\|^2 \right| = \left| \boldsymbol{x}_i^\top (\boldsymbol{G}^\top \boldsymbol{G}) \boldsymbol{x}_i - \boldsymbol{x}_i^\top \boldsymbol{I}_m \boldsymbol{x}_i \right|. \tag{1.37}$$

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 1.21. *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,

$$\mathbb{P}[\lambda_d = 0] = \mathbb{P}[\mathsf{det}(\boldsymbol{G}\boldsymbol{G}^\top) = 0] = \mathbb{P}[p(\boldsymbol{G}) = 0] = 0. \tag{1.38}$$

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

$$\mathsf{det}(t\boldsymbol{I}_d - \boldsymbol{G}\boldsymbol{G}^\top) = \sum_{k=0}^{d} (-1)^{d-k} s_{d-k}(\boldsymbol{G}) t^k, \tag{1.39}$$

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:

$$s_k(\boldsymbol{G}) = \sum_{1 \leq i_1 < \cdots < i_k \leq d} \lambda_{i_1} \cdots \lambda_{i_k}. \tag{1.40}$$

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:

$$p(\boldsymbol{G}) := \prod_{1 \leq i < j \leq d} (\lambda_i - \lambda_j)^2. \tag{1.41}$$

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,

$$\mathbb{P}[\lambda_i = \lambda_j \text{ for some } i \neq j] = \mathbb{P}[p(\boldsymbol{G}) = 0] = 0, \tag{1.42}$$

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

$$\boldsymbol{G}^\top \boldsymbol{G} = \sum_{i=1}^{d} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top. \tag{1.43}$$

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

$$\mathsf{Stief}(m, d) := \{\boldsymbol{V} \in \mathbb{R}^{m \times d} : \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_d\}, \tag{1.44}$$

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

$$\mathsf{Stief}(m, m) = \mathcal{O}(m). \tag{1.45}$$

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

$$\mathsf{Law}(\boldsymbol{Qg}) = \mathsf{Law}(\boldsymbol{g}) = \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_m). \tag{1.46}$$

Proof. By Proposition 1.11 or direct calculation of the first two moments. □

Corollary 1.23. *If $\boldsymbol{G} \sim \mathcal{N}(0, \sigma^2)^{\otimes d \times m}$ and $\boldsymbol{Q} \in \mathcal{O}(m)$, then*

$$\mathsf{Law}(\boldsymbol{GQ}) = \mathsf{Law}(\boldsymbol{G}) = \mathcal{N}(0, \sigma^2)^{\otimes d \times m}. \tag{1.47}$$

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 1.24. *$\boldsymbol{L}(\boldsymbol{GQ}) = \boldsymbol{Q}^\top \boldsymbol{L}(\boldsymbol{G})$, 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 1.25. *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 1.26 (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:

$$\begin{aligned}\int_{\mathcal{O}(m)} g(\boldsymbol{Q}) d\mu(\boldsymbol{Q}) &= \int_{\mathcal{O}(m)} f(\boldsymbol{Q}) g(\boldsymbol{Q}) d\left(\frac{1}{2}\mu + \frac{1}{2}\nu\right)(\boldsymbol{Q}) \\ &= \frac{1}{2}\int_{\mathcal{O}(m)} f(\boldsymbol{Q}) g(\boldsymbol{Q}) d\mu(\boldsymbol{Q}) + \frac{1}{2}\int_{\mathcal{O}(m)} f(\boldsymbol{Q}) g(\boldsymbol{Q}) d\nu(\boldsymbol{Q}). \end{aligned} \tag{1.48}$$

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:

$$\boldsymbol{G}\boldsymbol{G}^\top = \sum_{i=1}^m \boldsymbol{g}_i \boldsymbol{g}_i^\top \overset{\text{(LLN)}}{\approx} m \cdot \mathbb{E}\boldsymbol{g}_1 \boldsymbol{g}_1^\top = \frac{m}{d}\boldsymbol{I}_d. \tag{1.49}$$

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:

$$\|\boldsymbol{M}\| = \sup_{\boldsymbol{x} \in \mathbb{S}^{d-1}(1)} |\boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}|. \tag{1.50}$$

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 1.27 (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 1.28. *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,*

$$K \le \|\boldsymbol{M}\| \le \frac{1}{1-2\epsilon} K. \tag{1.51}$$

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 1.29. *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 1.30. *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 1.31. *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 1.32. *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,*

$$P(\epsilon) \le N(\epsilon) \le P(\epsilon/2). \tag{1.52}$$

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 1.33. *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

$$|\mathcal{X}| \leq 1 + \frac{\mu(B(\mathbf{0}, 1 + \frac{\epsilon}{2}))}{\mu(B(\mathbf{0}, \frac{\epsilon}{2}))} = \left(1 + \frac{2}{\epsilon}\right)^d. \tag{1.53}$$

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 1.34. *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 1.35. *For any $\epsilon > 0$, there is an ϵ-net $\mathcal{X}$ of $\mathbb{S}^{d-1}(1) \subset \mathbb{R}^d$ with $|\mathcal{X}| \le (1 + 8/\epsilon^2)^d$.*

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

$$\mathbb{P}\left[|\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}| > t\right] \le 2\exp\left(-\frac{d^2t^2}{8m}\right). \tag{1.54}$$

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 1.15, 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 1.37. *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 $m \ge C_1 d$, then*

$$\mathbb{P}\left[\left\|\boldsymbol{G}\boldsymbol{G}^\top - \frac{m}{d}\boldsymbol{I}_d\right\| \ge C_2\sqrt{\frac{m}{d}}\right] \le \exp(-cd). \tag{1.55}$$

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

$$\mathbb{P}\left[\|\boldsymbol{G}\boldsymbol{G}^\top - m\boldsymbol{I}_d\| \ge C_2\sqrt{md}\right] \le \exp(-cd), \tag{1.56}$$

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

$$\mathbb{P}\left[\sqrt{m} - C_2\sqrt{d} \le \sigma_d(\boldsymbol{G}) \le \cdots \le \sigma_1(\boldsymbol{G}) \le \sqrt{m} + C_2\sqrt{d}\right] \ge 1 - \exp(-cd). \tag{1.57}$$

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 1.28, $\|\boldsymbol{M}\| \le 2K$.

Then, by the union bound and Proposition 1.36,

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

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 1.36, 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 1.38. *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

$$\frac{m}{d}\|\boldsymbol{P}\boldsymbol{x}\|^2 \overset{\text{(law)}}{=} \frac{m}{d}\|\boldsymbol{Q}\boldsymbol{P}^{(0)}\boldsymbol{Q}^\top\boldsymbol{x}\|^2 = \frac{m}{d}\|\boldsymbol{P}^{(0)}\boldsymbol{y}\|^2 = \frac{m}{d}\sum_{i=1}^{d} y_i^2, \tag{1.58}$$

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

$$\|\boldsymbol{x}\|_0 := \#\{i \in [m] : x_i \ne 0\}. \tag{1.59}$$

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 1.39. *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 1.40 (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 1.41. $\boldsymbol{G}$ *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 1.42 (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* $\|\boldsymbol{x}\|_0 \le k$,

$$(1 - \delta)\|\boldsymbol{x}\|^2 \le \|\boldsymbol{G}\boldsymbol{x}\|^2 \le (1 + \delta)\|\boldsymbol{x}\|^2. \tag{1.60}$$

The following implication is immediate.

Proposition 1.43. *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 1.44. *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*

$$\mathbb{P}[\boldsymbol{G} \text{ has the } (k,\delta)\text{-RIP}] \geq 1 - \frac{2}{m^k}. \tag{1.61}$$

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

$$\|\boldsymbol{x}\|^2 = \|\boldsymbol{x}^{(S)}\|^2, \tag{1.62}$$

$$\boldsymbol{G}\boldsymbol{x} = \boldsymbol{G}^{(S)}\boldsymbol{x}^{(S)}. \tag{1.63}$$

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

$$(1-\delta)\|\boldsymbol{x}^{(S)}\|^2 \leq \|\boldsymbol{G}^{(S)}\boldsymbol{x}^{(S)}\|^2 \leq (1+\delta)\|\boldsymbol{x}^{(S)}\|^2. \tag{1.64}$$

But this just amounts to asking that

$$1 - \delta \leq \lambda_k(\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)}) \leq \lambda_1(\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)}) \leq 1 + \delta, \tag{1.65}$$

or again equivalently that

$$\|\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)} - \boldsymbol{I}_k\| \leq \delta. \tag{1.66}$$

We will then proceed by union bounding,

$$\mathbb{P}[\boldsymbol{G} \text{ does not have the } (k,\delta)\text{-RIP}] \leq \binom{m}{k} \mathbb{P}[\|\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)} - \boldsymbol{I}_k\| > \delta], \tag{1.67}$$

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 1.37. 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 1.37 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 1.37 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 1.15 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

$$\leq 2\exp\left(-2k\log m\right). \tag{1.68}$$

Finally, we have

$$\begin{aligned}
\mathbb{P}[\boldsymbol{G} \text{ does not have the } (k,\delta)\text{-RIP}] &\leq m^k \mathbb{P}[\|\boldsymbol{G}^{(S)^\top}\boldsymbol{G}^{(S)} - \boldsymbol{I}_k\| > \delta] \\
&\leq 2\exp(k\log m - 2k\log m) \\
&\leq \frac{2}{m^k},
\end{aligned} \tag{1.69}$$

completing the proof. □

Bibliography

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

[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.

[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.

[Bou85] Jean Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel Journal of Mathematics*, 52:46–52, 1985.

[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.

[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.

[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.

[Sha13] Cosma Shalizi. Advanced data analysis from an elementary point of view, 2013.