Bandeira-Singer-Strohmer

Mathematics of Data Science

Draft: version 0.1

June 11, 2020

Draft

Contents

0

Notes on this Draft and Current Status

This is a draft of a book in preparation by the authors.

While the contents are already fairly self-contained there are still chapters we plan to add. In particular, chapters on Low Rank Modelling, Randomized Linear Algebra, Statistics, Optimization, and Deep Learning are in the works.

The Introduction (Chapter 1) is also not complete at this point.

While Chapters 2 through 10 are not at their final state, we anticipate their focus and content not to change drastically and they can already by used for a graduate course in Mathematics of Data Science; they have been used as such by the authors at their home institutions.

We welcome suggestions and comments, and would like to learn about any possible errors and typos. Please contact the authors at bandeira@math.ethz.ch, strohmer@math.ucdavis.edu, or amits@math.princeton.edu.

Thank you,
Afonso, Thomas, and Amit.

2

Curses, Blessings, and Surprises in High Dimensions

This chapter discusses the curse of dimensionality, as well as many of its blessings. The first is caused by the exponential increase in volume associated with adding extra dimensions to Euclidean space. The latter is a manifestation of an intriguing phenomenon called the concentration of measure. This concentration phenomenon will give rise to many surprising facts about high dimensional geometry that we will discuss. Since several of the results discussed in this chapter require basic tools from probability, we will also review some fundamental probabilistic concepts.

2.1 The Curse of Dimensionality

The *curse of dimensionality* refers to the fact that many algorithmic approaches to problems in $\mathbb{R}^d$ become *exponentially* more difficult as the dimension d grows. The expression "curse of dimensionality" was coined by Richard Bellman to describe the problem caused by the exponential increase in volume associated with adding extra dimensions to Euclidean space [27].

For instance, if we want to sample the unit interval such that the distance between adjacent points is at most 0.01, 100 evenly-spaced sample points suffice; an equivalent sampling of a five-dimensional unit hypercube with a grid with a spacing of 0.01 between adjacent points would require 10^{10} sample points. Thus, a modest increase in dimensions results in a dramatic increase in required data points to cover the space at the same density.

Intimately connected to the curse of dimensionality is the problem of *overfitting* and *underfitting*. Here, overfitting refers to the issue that an algorithm may show good performance on the training data, but poor generliazation to other data. Underfitting in turn, corresponds to poor performance on the training data (and poor generalization to other data). This problem manifests itself in many machine learning algorithms.

We will discuss a toy example from image classification in more detail to illustrate the underlying issues. Assume we want to classify images into two

groups, cars and bicycles, say. From the vast number of images depicting cars or bicycles, we are only able to obtain a small number of training images, say five images of cars and five images of bicycles. We want to train a simple linear classifier based on these ten labeled training images to correctly classify the remaining unlabeled car/bicycle images. We start with a simple feature, e.g. the amount or red pixels' in each image. However, this is unlikely to give a linear separation of the training data. We add more features and eventually the training images become linearly separable. This might suggest that increasing the number of features until perfect classification of the training data is achieved, is a sound strategy. However, as we *linearly increase* the dimension of the feature space, the density of our training data *decreases exponentially* with the feature dimension.

In other words, to maintain a comparable density of our training data, we would need to increase the size of the datset exponentially – the curse of dimensionality. Thus, we risk producing a model that could be very good at predicting the target class on the training set, but it may fail miserably when faced with new data. This means that our model does not *generalize* from the training data to the test data.

2.2 Surprises in High Dimensions

When we peel an orange, then after having removed the rind we are still left with the majority of the orange. Suppose now we peel a d-dimensional orange for large d, then after removing the orange peel we would be left with essentially nothing. The reason for this – from a healthy nutrition viewpoint discouraging – fact is that for a d-dimensional unit ball almost all of its volume is concentrated near the boundary sphere. This is just one of many surprising phenomena in high dimensions. Many of these surprises are actually a manifestation of some form of concentration of measure that we will analyze in more detail in the next section (and then these surprises are not so surprising anymore ...).

When introducing data analysis concepts, we typically use few dimensions in order to facilitate visualization. However, our intuition about space, which is naturally based on two and three dimensions, can often be misleading in high dimensions. Many properties of even very basic objects become counterintuitive in higher dimensions. Understanding these paradoxical properties is essential in data analysis as it allows us to avoid pitfalls in the design of algorithms and statistical methods for high-dimensional data. It is therefore instructive to analyze the shape and properties of some basic geometric forms that we understand very well in dimensions two and three, in high dimensions.

To that end, we will look at some of the properties of the sphere and the cube as the dimension increases. The d-dimensional hyperball of radius R is defined by

$$B^d(R) = \{x \in \mathbb{R}^d : x_1^2 + \cdots + x_d^2 \leq R^2\},$$

the d-dimensional hypersphere (or d-sphere) of radius R is given by

$$S^{d-1}(R) = \{x \in \mathbb{R}^d : x_1^2 + \cdots + x_d^2 = R^2\},$$

and the d-dimensional hypercube with side length $2R$ is the subset of $\mathbb{R}^d$ defined as the d-fold product of intervals $[-R, R]$:

$$C^d(R) = \underbrace{[-R, R] \times \cdots \times [-R, R]}_{d \text{ times}}.$$

If there is no danger of confusion, we may write B^d for $B^d(1)$, S^{d-1} for $S^{d-1}(1)$, and C^d for $C^d(\frac{1}{2})$.

2.2.1 Geometry of spheres and balls in high dimension

Volume of the hyperball

Theorem 2.1. *The volume of $B^d(R)$ is given by*

$$\mathrm{Vol}(B^d(R)) = \frac{\pi^{\frac{d}{2}} R^d}{\frac{d}{2}\, \Gamma(\frac{d}{2})}. \tag{2.1}$$

Proof. The volume of $B^d(R)$ is given by

$$\mathrm{Vol}(B^d(R)) = \int_0^R s_d r^{d-1} dr = \frac{s_d R^d}{d}, \tag{2.2}$$

where s_d denotes the (hyper-)surface area of a unit d-sphere. A unit d-sphere must satisfy

$$s_d \int_0^\infty e^{-r^2} r^{d-1} dr = \underbrace{\int_{-\infty}^\infty \cdots \int_{-\infty}^\infty}_{d \text{ times}} e^{-(x_1^2 + \cdots + x_d^2)} dx_1 \ldots dx_d = \Big(\int_{-\infty}^\infty e^{-x^2} dx \Big)^d.$$

Recall that the Gamma function is given by

$$\Gamma(n) = \int_0^\infty r^{n-1} e^{-r} dr = 2 \int_0^\infty e^{-r^2} r^{2n-1} dr,$$

hence

$$\frac{1}{2} s_d \Gamma(\frac{d}{2}) = \Big[\Gamma(\frac{1}{2})\Big]^d = \big(\pi^{\frac{1}{2}}\big)^d,$$

and thus

$$s_d = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}.$$

Plugging this expression into (2.2) gives

$$\mathrm{Vol}(B^d(R)) = \frac{\pi^{\frac{d}{2}} R^d}{\frac{d}{2}\, \Gamma(\frac{d}{2})}. \tag{2.3}$$

□

For positive integers n there holds $\Gamma(n) = (n-1)!$. Using Stirling's Formula,

$$\Gamma(n) \sim \sqrt{\frac{2\pi}{n}} \left(\frac{n}{e}\right)^n$$

we obtain as approximation for the volume of the unit d-ball for large d

$$\mathrm{Vol}(B^d) \approx \frac{1}{\sqrt{d\pi}} \left(\frac{2\pi e}{d}\right)^{\frac{d}{2}}. \tag{2.4}$$

Since the denominator in the parenthesis of equation (2.4) goes to infinity much faster than the numerator, the volume of the unit d-sphere goes rapidly to 0 as the dimension d increases to infinity, see also Figure 2.1.

Thus, unit spheres in high dimensions have almost no volume—compare this to the unit cube, which has volume 1 in any dimension. For $B^d(R)$ to have volume equal to 1, its radius R must be approximately (asymptotically) equal to $\sqrt{\frac{d}{2\pi e}}$.

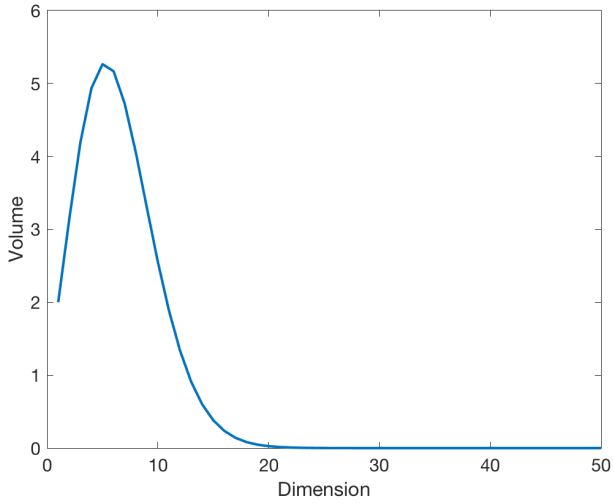


Fig. 2.1: The volume of the unit d-ball using the exact formula in equation (2.3). The volume reaches its maximum for $d = 5$ and decreases rapidly to zero with increasing dimension d.

Concentration of the volume of a ball near its equator

If we take an orange and cut it into slices, then the slices near the center are larger since the sphere is wider there. This effect increases dramatically

(exponentially with the dimension) with increasing dimension. Assume we want to cut off a slab around the "equator[1]" of the d-unit ball such that 99% of its volume is contained inside the slab. In two dimensions the width of the slab has to be almost 2, so that 99% of the volume are captured by the slab. But as the dimension increases the width of the slab gets rapidly smaller. Indeed, in high dimensions only a very thin slab is required, since nearly all the volume of the unit ball lies a very small distance away from the equator. The following theorem makes the considerations above precise.

Theorem 2.2. *Almost all the volume of $B^d(R)$ lies near its equator.*

Proof. It suffices to prove the result for the unit d-ball. Without loss of generality we pick as "north" the direction x_1. The intersection of the sphere with the plane $x_1 = 0$ forms our equator, which is formally given by the $d-1$-dimensional region $\{x : \|x\| \leq 1, x_1 = 0\}$. This intersection is a sphere of dimension $d-1$ with volume $\mathrm{Vol}(B^{d-1})$ given by the $(d-1)$-analog of formula (2.3) with $R = 1$.

We now compute the volume of B^d that lies between $x_1 = 0$ and $x_1 = p_0$. Let $P_0 = \{x : \|x\| \leq 1, x_1 \geq p_0\}$ be the "polar cap", i.e., part of the sphere above the slab of width $2p_0$ around the equator. To compute the volume of the cap P we will integrate over all slices of the cap from 0 to p_0. Each such slice will be a sphere of dimension $d-1$ and radius $\sqrt{1-p^2}$, hence its volume is $(1-p^2)^{\frac{d-1}{2}} \mathrm{Vol}(B^{d-1})$. Therefore

$$\mathrm{Vol}(P) = \int_{p_0}^{1} (1-p^2)^{\frac{d-1}{2}} \mathrm{Vol}(B^{d-1})\, dp = \mathrm{Vol}(B^{d-1}) \int_{p_0}^{1} (1-p^2)^{\frac{d-1}{2}}\, dp.$$

Using $e^x \geq 1 + x$ for all x we can upper bound this integral by

$$\mathrm{Vol}(P) \leq \mathrm{Vol}(B^{d-1}) \int_{p_0}^{\infty} e^{-\frac{d-1}{2}p^2}\, dp \leq \frac{\mathrm{Vol}(B^{d-1})}{d-1} e^{-\frac{(d-1)p_0^2}{2}},$$

where we have bounded the integral via the complementary error function $\mathrm{erfc}(x)$ and used the fact that $\mathrm{erfc}(x) \leq e^{-x^2}$.

Recall, from (2.3) that $\mathrm{Vol}(B^d) = \frac{\pi^{\frac{d}{2}}}{\frac{d}{2}\,\Gamma(\frac{d}{2})}$, so, for d large enough (since $\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-1}{2})} \approx \sqrt{\frac{d}{2}}$),

$$\mathrm{Vol}(B^{d-1}) = \frac{\pi^{-1/2}}{\frac{d-1}{d}} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-1}{2})} \mathrm{Vol}(B^d) \leq \frac{d-1}{2} \mathrm{Vol}(B^d).$$

Finally, a simple calculation shows that the ratio between the volume of the polar caps and the entire hypersphere is bounded by

[1] To define the "equator" of a the d-dimensional ball, we need to pick a "north pole" as reference. Without loss of generality we could pick the unit vector in the x_1-direction as defining "north".

$$\frac{2\,\mathrm{Vol}(P)}{\mathrm{Vol}(B^d)} \leq \frac{2\,\mathrm{Vol}(P)}{\mathrm{Vol}(B^{d-1})} \leq e^{-\frac{d-1}{2}p_0^2}.$$

The expression above shows that this ratio decreases exponentially as both d and p increase, proving our claim that the volume of the sphere concentrates strongly around its equator. □

Concentration of the volume of a ball on shells

We consider two concentric balls $B^d(1)$ and $B^d(1-\varepsilon)$. Using equation (2.3), the ratio of their volumes is

$$\frac{\mathrm{Vol}(B^d(1-\varepsilon))}{\mathrm{Vol}(B^d(1))} = (1-\varepsilon)^d.$$

Clearly, for every ε this ratio tends to zero as $d \to \infty$. This implies that the spherical shell given by the region between $B^d(1)$ and $B^d(1-\varepsilon)$ will contain most of the volume of $B^d(1)$ for large enough d even if ε is very small. How quickly does the volume concentrate at the surface of $B^d(1)$? We choose ε as a function of d, e.g. $\varepsilon = \frac{t}{d}$, then

$$\frac{\mathrm{Vol}(B^d(1-\varepsilon))}{\mathrm{Vol}(B^d(1))} = (1-\frac{t}{d})^d \to e^{-t}.$$

Thus, almost all the volume of $B^d(R)$ is contained in an annulus of width R/d.

Therefore, if we peel a d-dimensional orange and even if we peel it very carefully so that we remove only a very thin layer of its peel, we will have removed most of the orange and are left with almost nothing.

2.2.2 Geometry of the Hypercube

We have seen that most of the volume of the hypersphere is concentrated near its surface. A similar result also holds for the hypercube, and in general for high-dimensional geometric objects. Yet, the hypercube exhibits an even more interesting volume concentration behavior, which we will establish below.

We start with a basic observation.

Proposition 2.3. *The hypercube C^d has volume 1 and diameter $\sqrt{d}$.*

The above proposition, although mathematically trivial, hints already at a somewhat counterintuitive behavior of the cube in high dimensions. Its corners seem to get "stretched out" more and more, while the rest of the cube must "shrink" to keep the volume constant. This property becomes even more striking when we compare the cube with the sphere as the dimension increases.

In two dimensions (Figure 2.2), the unit square is completely contained in the unit sphere. The distance from the center to a vertex (radius of the circumscribed sphere) is $\frac{\sqrt{2}}{2}$ and the apothem (radius of the inscribed sphere) is $\frac{1}{2}$. In

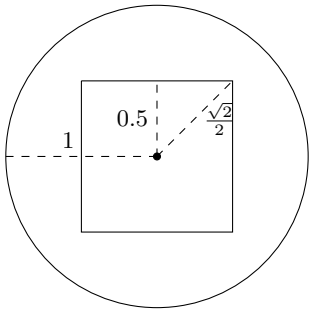


Fig. 2.2: 2-dimensional unit sphere and unit cube, centered at the origin.

four dimensions (Figure 2.3), the distance from the center to a vertex is 1, so the vertices of the cube touch the surface of the sphere. However, the apothem is still $\frac{1}{2}$. The result, when projected in two dimensions no longer appears convex, however all hypercubes are convex. This is part of the strangeness of higher dimensions - hypercubes are both convex and "pointy." In dimensions greater than 4 the distance from the center to a vertex is $\frac{\sqrt{d}}{2} > 1$, and thus the vertices of the hypercube extend far outside the sphere, cf. Figure 2.4.

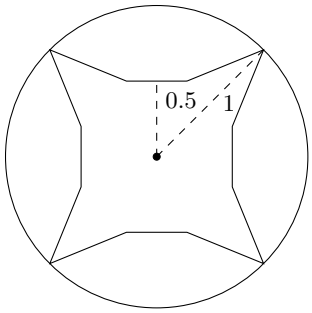


Fig. 2.3: Projections of the 4-dimensional unit sphere and unit cube, centered at the origin (4 of the 16 vertices of the hypercube are shown).

The considerations above suggest the following observation:

"Most of the volume of the high-dimensional cube is located in its corners."

We will prove this observation in Section 2.4.2 using probabilistic techniques which we will introduce in the next sections.

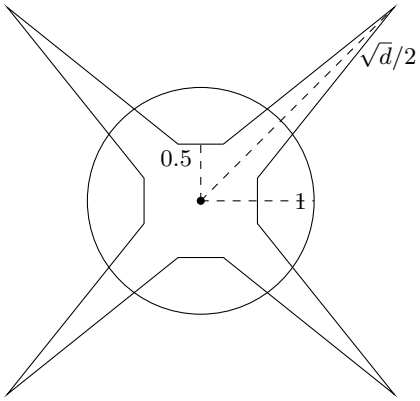


Fig. 2.4: Projections of the d-dimensional unit sphere and unit cube, centered at the origin (4 of the 2^d vertices of the hypercube are shown).

2.3 Basic Concepts from Probability

We briefly review some fundamental concepts from probability theory, which are helpful or necessary to understand the blessings of dimensionality and some of the surprises encountered in high dimensions. More advanced probabilistic concepts will be presented in Chapter 6. We assume that the reader is familiar with elementary probability as is covered in introductory probability courses (see, for example [54, 113]).

The two most basic concepts in probability associated with a random variable X are *expectation* (or *mean*) and *variance*, denoted by

$$\mathbb{E}[X] \qquad \text{and} \qquad \mathrm{Var}(X) := \mathbb{E}[X - \mathbb{E}[X]]^2,$$

respectively. An important tool to describe probability distributions is the *moment generating function* of X, defined by

$$M_X(t) = \mathbb{E}[e^{tX}], \quad t \in \mathbb{R},$$

the choice of nomenclature can be easily justified by expanding $M_X(t)$ in a series. The p-th moment of X is defined by $\mathbb{E}[X^p]$ for $p > 0$ and the p-th absolute moment is $\mathbb{E}[|X|^p]$.

We can introduce L^p-norms of random variables by taking the p-th root of moments, i.e.,

$$\|X\|_{L^p} := \left(\mathbb{E}[|X|^p]\right)^{\frac{1}{p}}, \qquad p \in [0, \infty],$$

with the usual extension to $p = \infty$ by setting

$$\|X\|_\infty := \operatorname{ess\,sup} |X|.$$

Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, where Σ denotes a σ-algebra on the sample space Ω and $\mathbb{P}$ is a probability measure on (Ω, Σ). For fixed p the vector space $L^p(\Omega, \Sigma, \mathbb{P})$ consists of all random variables X on Ω with finite L^p-norm, i.e.,

$$L^p(\Omega, \Sigma, \mathbb{P}) = \{X : \|X\|_{L^p} < \infty\}.$$

We will usually not mention the underlying probability space. For example, we will often simply write L^p for $L^p(\Omega, \Sigma, \mathbb{P})$.

The case $p = 2$ deserves special attention since L^2 is a Hilbert space with inner product and norm

$$\langle X, Y \rangle_{L^2} = \mathbb{E}[XY], \qquad \|X\|_{L^2} = \big(\mathbb{E}[X^2]\big)^{\frac{1}{2}},$$

respectively. Note that the *standard deviation* $\sigma(X) := \sqrt{\mathrm{Var}(X)}$ of X can be written as

$$\sigma(X) = \|X - \mathbb{E}[X]\|_{L^2}.$$

The *covariance* of the random variables X and Y is

$$\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \langle X - \mathbb{E}[X], Y - \mathbb{E}[Y] \rangle_{L^2}. \quad (2.5)$$

We recall a few classical inequalities for random variables. *Hölder's inequality* states that for random variables X and Y on a common probability space and $p, q \geq 1$ with $1/p + 1/q = 1$, there holds

$$|\mathbb{E}[XY]| \leq \|X\|_{L^p} \|Y\|_{L^q}. \quad (2.6)$$

The special case $p = q = 2$ is the *Cauchy-Schwarz inequality*

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[|X|^2]\mathbb{E}[|Y|^2]}. \quad (2.7)$$

Jenssen's inequality states that for any random variable X and a convex function $\varphi : \mathbb{R} \to \mathbb{R}$, we have

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]. \quad (2.8)$$

Since $\varphi(x) = x^{q/p}$ is a convex function for $q \geq p \geq 0$, it follows immediately from Jenssen's inequality that

$$\|X\|_{L^p} \leq \|X\|_{L^q} \qquad \text{for } 0 \leq p \leq q < \infty.$$

Minkovskii's inequality states that for any $p \in [0, \infty]$ and any random variables X, Y, we have

$$\|X + Y\|_{L^p} \leq \|X\|_{L^p} + \|Y\|_{L^p}, \quad (2.9)$$

which can be viewed as the *triangle inequality*.

The *cumulative distribution function* of X is defined by

$$F_X(t) = \mathbb{P}(X \leq t), \quad t \in \mathbb{R}.$$

We have $\mathbb{P}\{X > t\} = 1 - F_X(t)$, where the function $t \mapsto \mathbb{P}\{|X| \geq t\}$ is called the *tail* of X. The following lemma establishes a close connection between expectation and tails.

Proposition 2.4 (Integral identity). *Let X be a non-negative random variable. Then*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}\{X > t\}\, dt.$$

The two sides of this identity are either finite or infinite simultaneously.

Given an event E with non-zero probability,$\mathbb{P}(\cdot|E)$ denotes conditional probability, furthermore for a random variable X we use $\mathbb{E}[X|E]$ to denote the conditional expectation.

Markov's inequality is a fundamental tool to bound the tail of a random variable in terms of its expectation.

Proposition 2.5. *For any non-negative random variable $X : S \to \mathbb{R}$ we have*

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}[X]}{t} \qquad \text{for all } t > 0. \tag{2.10}$$

We provide two versions of the same proof, one using the language of conditional expectations.

Proof. Let $\mathcal{I}$ denote the event $\{X \geq t\}$. Then

$$\mathbb{E}[X] = \sum_{s \in S} p(s)X(s) = \sum_{s \in \mathcal{I}} p(s)X(s) + \sum_{s \in \mathcal{I}^c} p(s)X(s),$$

where $p(s)$ denotes the probability of s; in case of continuous variables this should be replaced with the density function and $\sum$ with an integral.

Since X is non-negative, it holds $\sum_{s \in \mathcal{I}^c} p(s)X(s) \geq 0$ and

$$\mathbb{E}[X] \geq \sum_{s \in \mathcal{I}} p(s)X(s) \geq t \sum_{s \in \mathcal{I}} p(s) = t\mathbb{P}\{\mathcal{I}\}.$$

Proof (Using the language of conditional expectation).

$$\mathbb{E}[X] = \mathbb{P}(X < t)\mathbb{E}[X|X < t] + \mathbb{P}(X > t)\mathbb{E}[X|X \geq t],$$

where we take the product to be zero if the probability is zero.

Since X is non-negative, it holds $\mathbb{P}(X < t)\mathbb{E}[X|X < t] \geq 0$. Also, $\mathbb{E}[X|X \geq t] > t$. Hence,

$$\mathbb{E}[X] \geq \mathbb{P}(X > t)\mathbb{E}[X|X > t] \geq t\mathbb{P}(X \geq t).$$

An important consequence of Markov's inequality is *Chebyshev's inequality.*

Corollary 2.6. *Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$*

$$\mathbb{P}\{|X - \mu| \geq t\} \leq \frac{\sigma^2}{t^2}. \tag{2.11}$$

Chebyshev's inequality, which follows by applying Markov's inequality to the non-negative random variable $Y = (X - \mathbb{E}[X])^2$, is a form of concentration inequality, as it guarantees that X must be close to its mean μ whenever the variance of X is small. Both, Markov's and Chebyshev's inequality are sharp, i.e., in general they cannot be improved.

Markov's inequality only requires the existence of the first moment. We can say a bit more if in addition the random variable X has a moment generating function in a neighborhood around zero, that is, there is a constant $b > 0$ such that $\mathbb{E}[e^{\lambda(X-\mu)}]$ exists for all $\lambda \in [0, b]$. In this case we can apply Markov's inequality to the random variable $Y = e^{\lambda(X-\mu)}$ and obtain the generic *Chernoff bound*

$$\mathbb{P}\{X - \mu \geq t\} = \mathbb{P}\{e^{\lambda(X-\mu)} \geq e^{\lambda t}\} \leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}}. \tag{2.12}$$

In particular, optimizing over λ in order to obtain the tightest bound in (2.12) gives

$$\log \mathbb{P}\{X - \mu \geq t\} \leq - \sup_{\lambda \in [0,b]} \{\lambda t - \log \mathbb{E}[e^{\lambda(X-\mu)}]\}.$$

Gaussian random variables are among the most important random variables. A Gaussian random variable X with mean μ and standard deviation σ has a probability density function given by

$$\psi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right). \tag{2.13}$$

We write $X \sim \mathcal{N}(\mu, \sigma^2)$. We call a Gaussian random variable X with $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$ a *standard Gaussian* or *standard normal* (random variable). In this case we have the following tail bound.

Proposition 2.7 (Gaussian tail bounds). *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then for all $t > 0$*

$$\mathbb{P}(X \geq \mu + t) \leq e^{-t^2/2\sigma^2}. \tag{2.14}$$

Proof. We use the moment-generating function $\lambda \mapsto \mathbb{E}[e^{\lambda X}]$. A simple calculation gives

$$\mathbb{E}[e^{\lambda X}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda x - x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} e^{\lambda^2/2} \int_{-\infty}^{\infty} e^{-(x-\lambda)^2/2}\, dx = e^{\lambda^2/2},$$

where we have used the fact that $\int_{-\infty}^{\infty} e^{-(x-\lambda)^2/2}\, dx$ is just the entire Gaussian integral shifted and therefore its value is $\sqrt{2\pi}$. We now apply Chernoff's bound (2.12) and obtain $\mathbb{P}(X > t) \leq \mathbb{E}[e^{\lambda X}] e^{-\lambda t}$. Minimizing this expression over λ gives $\lambda = t$ and thus $\mathbb{P}(X > t) \leq e^{-t^2/2}$.

Definition 2.8. *A random variable X with mean $\mu = \mathbb{E}[X]$ is called* sub-Gaussian *if there is a positive number σ such that*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\sigma^2\lambda^2/2}, \qquad \text{for all } \lambda \in \mathbb{R}.$$

If X satisfies the above definition, we also say that X is sub-Gaussian with parameter σ, or X is (μ, σ) sub-Gaussian in case we want to emphasize μ as well. Clearly, owing to the symmetry in the definition, $-X$ is sub-Gaussian if and only if X is sub-Gaussian. Obviously, any Gaussian random variable with variance σ^2 is sub-Gaussian with parameter σ. We refer to [138] for other, equivalent, definitions of sub-Gaussian random variables.

Combining the moment condition in Definition 2.8 with calculations similar to those that lead us to the Gaussian tail bounds in 2.7, yields the following concentration inequality for sub-Gaussian random variables.

Proposition 2.9 (Sub-Gaussian tail bounds). *Assume X is sub-Gaussian with parameter σ. Then for all $t > 0$*

$$\mathbb{P}(|X - \mu| \geq t) \leq e^{-t^2/2\sigma^2} \qquad \text{for all } t \in \mathbb{R}. \tag{2.15}$$

An important example of non-Gaussian, but sub-Gaussian random variables are *Rademacher random variables*. A Rademacher random variable ε takes on the values ± 1 with equal probability and is sub-Gaussian with parameter σ. Indeed, any bounded random variable is sub-Gaussian.

While many important random variables have a sub-Gaussian distribution, this class does not include several frequently occurring distributions with heavier tails. A classical example is the *chi-squared distribution*, which we will discuss at the end of this section.

Relaxing slightly the condition on the moment-generating function in Definition 2.8 leads to the class of *sub-exponential* random variables.

Definition 2.10. *A random variable X with mean $\mu = \mathbb{E}[X]$ is called* sub-exponential *if there are parameters ν, b such that*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\nu^2\lambda^2/2}, \qquad \text{for all } \lambda \leq \frac{1}{b}.$$

Clearly, a sub-Gaussian random variable is sub-exponential (set $\nu = \sigma$ and $b = 0$, where $1/b$ is interpreted as $+\infty$). However, the converse is not true. Take for example $X \sim \mathcal{N}(0, 1)$ and consider the random variable $Z = X^2$. For $\lambda < \frac{1}{2}$ it holds that

$$\mathbb{E}[e^{\lambda(Z-1)}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda(x^2-1)} e^{-x^2/2} dx = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}. \tag{2.16}$$

However, for $\lambda \geq \frac{1}{2}$ the moment-generating function does not exist, which implies that X^2 is not sub-Gaussian. But X^2 is sub-exponential. Indeed, a brief computation shows that

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} = e^{4\lambda^2/2}, \qquad \text{for all } |\lambda| \leq 1/4,$$

which in turn implies that X^2 is sub-exponential with parameters $(\nu, b) = (2, 4)$.

Following a similar procedure that yielded sub-Gaussian tail bounds produces concentration inequalities for sub-exponential random variables. However, in this case we see two different types of concentration emerging, depending on the value of t.

Proposition 2.11 (Sub-exponential tail bounds). *Assume X is sub-exponential with parameters (ν, b). Then*

$$\mathbb{P}(X \geq \mu + t) \leq \begin{cases} e^{-t^2/2\nu^2} & \text{if } 0 \leq t \leq \frac{\nu^2}{b}, \\ e^{-t/2b} & \text{if } t > \frac{\nu^2}{b}. \end{cases} \tag{2.17}$$

Both the sub-Gaussian property and the sub-exponential property is preserved under summation for independent random variables, and the associated parameters transform in a simple manner.

A collection $X_1, \ldots, X_n$ of mutually independent random variables that all have the same distribution is called independent identically distributed (i.i.d.). A random variable X' is called an independent copy of X if X and X' are independent and have the same distribution.

Since we are not able to improve Markov's inequality and Chebyshev's inequality in general, the question arises whether we can give a stronger statement for a more restricted class of random variables. Of central importance in this context is the case of a random variable that is the *sum of a number of independent random variables*. This leads to the rich topic of *concentration inequalities* which is discussed in the next sections in this chapter and in Chapter 6.

Before we dive right into a range of concentration inequalities in the next section, we want to investigate one particular example. If $X_1, \ldots, X_n$ are independent, standard normal random variables, then the sum of their squares, $Z = \sum_{k=1}^n X_k^2$ is distributed according to the *chi-squared distribution* with n degrees of freedom. We denote this by $Z \sim \chi^2(n)$. Its probability density function is

$$\varphi(t) = \begin{cases} \frac{t^{\frac{n}{2}-1} e^{-\frac{n}{2}}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})}, & t > 0. \\ 0, & \text{else.} \end{cases}$$

Since the $X_k^2, k = 1, \ldots, n$ are subexponential with parameters $(2, 4)$ and independent, $Z = \sum_{k=1}^n X_k^2$ is subexponential with parameters $(2\sqrt{n}, 4)$. Therefore, using (2.17), we obtain the χ^2 tail bound

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{k=1}^n X_k^2 - 1\right| \geq t\right) \leq \begin{cases} 2e^{-nt^2/8} & \text{for } t \in (0, 1). \\ 2e^{-nt/8} & \text{if } t \geq 1. \end{cases} \tag{2.18}$$

A variation of this bound is stated in Theorem 6.23.

2.4 Blessings of Dimensionality

Suppose we wish to predict the outcome of an event of interest. One natural approach would be to compute the expected value of the object. However, how can we tell how good the expected value is to the actual outcome of the event? Without further information of how well the actual outcome concentrates around its expectation, the expected value is of little use. We would like to have an estimate for the probability that the actual outcome deviates from its expectation by a certain amount. This is exactly the role that *concentration inequalities* play in probability and statistics.

The concentration of measure phenomenon was put forward by Vitali Milman in the asymptotic geometry of Banach spaces regarding probabilities on product spaces in high dimensions [93, 83].

The celebrated law of large numbers of classical probability theory is the most well known form of *concentration of measure*; it states that sums of independent random variables are, under very mild conditions, close to their expectation with a large probability. We will see various quantitative versions of such concentration inequalities throughout this course. Some deal with sums of scalar random variables, others with sums of random vectors or sums of random matrices. Such concentration inequalities are instances of what is sometimes called *Blessings of dimensionality* (cf. [50]). This expression refers to the fact that certain random fluctuations can be well controlled in high dimensions, while it would be very complicated to make such predictive statements in moderate dimensions.

2.4.1 Large Deviation Inequalities

Concentration and large deviations inequalities are among the most useful tools when understanding the performance of some algorithms. We start with two of the most fundamental results in probability. We refer to Sections 1.7 and 2.4 in [54] for the proofs and variations.

Theorem 2.12 (Strong law of large numbers). *Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with mean μ. Denote*

$$S_n := X_1 + \cdots + X_n.$$

Then, as $n \to \infty$

$$\frac{S_n}{n} \to \mu \qquad \textit{almost surely.} \tag{2.19}$$

The celebrated *central limit theorem* tells us that the limiting distribution of a sum of i.i.d. random variables is always Gaussian. The best known version is probably due to Lindeberg-Lévy.

Theorem 2.13 (Lindeberg-Lévy Central limit theorem). *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Denote*

$$S_n := X_1 + \cdots + X_n,$$

and consider the normalized random variable Z_n with mean zero and variance one, given by

$$Z_n := \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\operatorname{Var} S_n}} = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu).$$

Then, as $n \to \infty$

$$Z_n \to \mathcal{N}(0,1) \qquad \text{in distribution.} \tag{2.20}$$

The strong law of large numbers and the central limit theorem give us qualitative statements about the behavior of a sum of i.i.d. random variables. In many applications it is desirable to be able to quantify how such a sum deviates around its mean. This is where concentration inequalities come into play.

The intuitive idea is that if we have a sum of independent random variables

$$X = X_1 + \cdots + X_n,$$

where X_i are i.i.d. centered random variables, then while the value of X can be of order $\mathcal{O}(n)$ it will very likely be of order $\mathcal{O}(\sqrt{n})$ (note that this is the order of its standard deviation). The inequalities that follow are ways of very precisely controlling the probability of X being larger (or smaller) than $\mathcal{O}(\sqrt{n})$. While we could use, for example, Chebyshev's inequality for this, in the inequalities that follow the probabilities will be exponentially small, rather than just quadratically small, which will be crucial in many applications to come. Moreover, unlike the classical central limit theorem, the concentration inequalities below are *non-asymptotic* in the sense that they hold for all fixed n and not just for $n \to \infty$ (but the larger the n, the stronger the inequalities become).

Theorem 2.14 (Hoeffding's Inequality). *Let $X_1, X_2, \ldots, X_n$ be independent bounded random variables, i.e., $|X_i| \leq a_i$ and $\mathbb{E}[X_i] = 0$. Then,*

$$\mathbb{P}\left\{ \left| \sum_{i=1}^{n} X_i \right| > t \right\} \leq 2 \exp\left(- \frac{t^2}{2 \sum_{i=1}^{n} a_i^2} \right).$$

The inequality implies that fluctuations larger than $\mathcal{O}\left(\sqrt{n}\right)$ have small probability. For example, if $a_i = a$ for all i, setting $t = a\sqrt{2n \log n}$ yields that the probability is at most $\frac{2}{n}$.

Proof. We prove the result for the case $|X_i| \leq a$, the extension to the case $|X_i| \leq a_i$ is straightforward. We first get a probability bound for the event $\sum_{i=1}^{n} X_i > t$. The proof, again, will follow from Markov. Since we want an

exponentially small probability, we use a classical trick that involves exponentiating with any $\lambda > 0$ and then choosing the optimal λ.

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^{n} X_i > t\right\} &= \mathbb{P}\left\{\sum_{i=1}^{n} X_i > t\right\} \qquad (2.21)\\ &= \mathbb{P}\left\{e^{\lambda \sum_{i=1}^{n} X_i} > e^{\lambda t}\right\}\\ &\leq \frac{\mathbb{E}[e^{\lambda \sum_{i=1}^{n} X_i}]}{e^{t\lambda}}\\ &= e^{-t\lambda}\prod_{i=1}^{n}\mathbb{E}[e^{\lambda X_i}], \qquad (2.22)\end{aligned}$$

where the penultimate step follows from Markov's inequality and the last equality follows from independence of the X_i's.

We now use the fact that $|X_i| \leq a$ to bound $\mathbb{E}[e^{\lambda X_i}]$. Because the function $f(x) = e^{\lambda x}$ is convex,

$$e^{\lambda x} \leq \frac{a+x}{2a}e^{\lambda a} + \frac{a-x}{2a}e^{-\lambda a},$$

for all $x \in [-a, a]$.

Since, for all i, $\mathbb{E}[X_i] = 0$ we get

$$\mathbb{E}[e^{\lambda X_i}] \leq \mathbb{E}\left[\frac{a+X_i}{2a}e^{\lambda a} + \frac{a-X_i}{2a}e^{-\lambda a}\right] \leq \frac{1}{2}\left(e^{\lambda a} + e^{-\lambda a}\right) = \cosh(\lambda a)$$

Note that[2]

$$\cosh(x) \leq e^{x^2/2}, \quad \text{for all } x \in \mathbb{R}$$

Hence,

$$\mathbb{E}[e^{\lambda X_i}] \leq e^{(\lambda a)^2/2}.$$

Together with (2.21), this gives

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^{n} X_i > t\right\} &\leq e^{-t\lambda}\prod_{i=1}^{n} e^{(\lambda a)^2/2}\\ &= e^{-t\lambda}e^{n(\lambda a)^2/2}\end{aligned}$$

This inequality holds for any choice of $\lambda \geq 0$, so we choose the value of λ that minimizes

$$\min_{\lambda}\left\{n\frac{(\lambda a)^2}{2} - t\lambda\right\}$$

[2]This follows immediately from the Taylor expansions: $\cosh(x) = \sum_{n=0}^{\infty}\frac{x^{2n}}{(2n)!}$, $e^{x^2/2} = \sum_{n=0}^{\infty}\frac{x^{2n}}{2^n n!}$, and $(2n)! \geq 2^n n!$.

Differentiating readily shows that the minimizer is given by

$$\lambda = \frac{t}{na^2},$$

which satisfies $\lambda > 0$. For this choice of λ,

$$n(\lambda a)^2/2 - t\lambda = \frac{1}{n}\left(\frac{t^2}{2a^2} - \frac{t^2}{a^2}\right) = -\frac{t^2}{2na^2}$$

Thus,

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i > t\right\} \leq e^{-\frac{t^2}{2na^2}}$$

By using the same argument on $\sum_{i=1}^{n}(-X_i)$, and union bounding over the two events we get,

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} X_i\right| > t\right\} \leq 2e^{-\frac{t^2}{2na^2}}$$

□

Remark 2.15. Hoeffding's inequality is suboptimal in a sense we now describe. Let's say that we have random variables $r_1, \ldots, r_n$ i.i.d. distributed as

$$r_i = \begin{cases} -1 & \text{with probability } p/2 \\ 0 & \text{with probability } 1-p \\ 1 & \text{with probability } p/2. \end{cases}$$

Then, $\mathbb{E}(r_i) = 0$ and $|r_i| \leq 1$ so Hoeffding's inequality gives:

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} r_i\right| > t\right\} \leq 2\exp\left(-\frac{t^2}{2n}\right).$$

Intuitively, the smaller p is, the more concentrated $|\sum_{i=1}^{n} r_i|$ should be, however Hoeffding's inequality does not capture this behaviour.

A natural way to attempt to capture this behaviour is by noting that the variance of $\sum_{i=1}^{n} r_i$ depends on p as $\mathrm{Var}(r_i) = p$. The inequality that follows, Bernstein's inequality, uses the variance of the summands to improve over Hoeffding's inequality.

The way this is going to be achieved is by strengthening the proof above, more specifically in step (2.22) we will use the bound on the variance to get a better estimate on $\mathbb{E}[e^{\lambda X_i}]$ essentially by realizing that if X_i is centered, $\mathbb{E}X_i^2 = \sigma^2$, and $|X_i| \leq a$ then, for $k \geq 2$, $\mathbb{E}X_i^k \leq \mathbb{E}|X_i|^k \leq \sigma^2\mathbb{E}|X_i|^{k-2} \leq \sigma^2 a^{k-2} = \left(\frac{\sigma^2}{a^2}\right)a^k$.

Theorem 2.16 (Bernstein's Inequality).

Let $X_1, X_2, \ldots, X_n$ be independent centered bounded random variables satisfying $|X_i| \leq a$ and $\mathbb{E}[X_i^2] = \sigma^2$. Then,

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} X_i\right| > t\right\} \leq 2\exp\left(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}at}\right).$$

Remark 2.17. Before proving Bernstein's inequality, note that on the example of Remark 2.15 we get

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} r_i\right| > t\right\} \leq 2\exp\left(-\frac{t^2}{2np + \frac{2}{3}t}\right),$$

which exhibits a dependence on p and, for small values of p is considerably smaller than what Hoeffding's inequality gives.

Proof.

As before, we will prove

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i > t\right\} \leq \exp\left(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}at}\right),$$

and then union bound with the same result for $-\sum_{i=1}^{n} X_i$, to prove the Theorem.

For any $\lambda > 0$ we have

$$\begin{aligned}
\mathbb{P}\left\{\sum_{i=1}^{n} X_i > t\right\} &= \mathbb{P}\{e^{\lambda \sum X_i} > e^{\lambda t}\} \\
&\leq \frac{\mathbb{E}[e^{\lambda \sum X_i}]}{e^{\lambda t}} \\
&= e^{-\lambda t} \prod_{i=1}^{n} \mathbb{E}[e^{\lambda X_i}]
\end{aligned}$$

The following calculations reveal the source of the improvement over Hoeffding's inequality.

$$\begin{aligned}
\mathbb{E}[e^{\lambda X_i}] &= \mathbb{E}\left[1 + \lambda X_i + \sum_{m=2}^{\infty} \frac{\lambda^m X_i^m}{m!}\right] \\
&\leq 1 + \sum_{m=2}^{\infty} \frac{\lambda^m a^{m-2}\sigma^2}{m!} \\
&= 1 + \frac{\sigma^2}{a^2} \sum_{m=2}^{\infty} \frac{(\lambda a)^m}{m!} \\
&= 1 + \frac{\sigma^2}{a^2}\left(e^{\lambda a} - 1 - \lambda a\right)
\end{aligned}$$

Therefore,

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i > t\right\} \leq e^{-\lambda t}\left[1 + \frac{\sigma^2}{a^2}\left(e^{\lambda a} - 1 - \lambda a\right)\right]^n$$

We will use a few simple inequalities (that can be easily proved with calculus) such as[3] $1 + x \leq e^x, \text{ for all } x \in \mathbb{R}$.

This means that,

$$1 + \frac{\sigma^2}{a^2}\left(e^{\lambda a} - 1 - \lambda a\right) \leq e^{\frac{\sigma^2}{a^2}(e^{\lambda a} - 1 - \lambda a)},$$

which readily implies

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i > t\right\} \leq e^{-\lambda t} e^{\frac{n\sigma^2}{a^2}(e^{\lambda a} - 1 - \lambda a)}.$$

As before, we try to find the value of $\lambda > 0$ that minimizes

$$\min_{\lambda}\left\{-\lambda t + \frac{n\sigma^2}{a^2}(e^{\lambda a} - 1 - \lambda a)\right\}$$

Differentiation gives

$$-t + \frac{n\sigma^2}{a^2}(ae^{\lambda a} - a) = 0$$

which implies that the optimal choice of λ is given by

$$\lambda^* = \frac{1}{a}\log\left(1 + \frac{at}{n\sigma^2}\right)$$

If we set

$$u = \frac{at}{n\sigma^2}, \tag{2.23}$$

then $\lambda^* = \frac{1}{a}\log(1 + u)$.

Now, the value of the minimum is given by

$$-\lambda^* t + \frac{n\sigma^2}{a^2}(e^{\lambda^* a} - 1 - \lambda^* a) = -\frac{n\sigma^2}{a^2}\left[(1 + u)\log(1 + u) - u\right].$$

This means that

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i > t\right\} \leq \exp\left(-\frac{n\sigma^2}{a^2}\left\{(1 + u)\log(1 + u) - u\right\}\right)$$

The rest of the proof follows by noting that, for every $u > 0$,

[3] In fact $y = 1 + x$ is a tangent line to the graph of $f(x) = e^x$.

$$(1+u)\log(1+u) - u \geq \frac{u}{\frac{2}{u} + \frac{2}{3}}, \tag{2.24}$$

which implies:

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^{n} X_i > t\right\} &\leq \exp\left(-\frac{n\sigma^2}{a^2}\frac{u}{\frac{2}{u}+\frac{2}{3}}\right) \\ &= \exp\left(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}at}\right).\end{aligned}$$

□

We refer to [138] for several useful variations of Bernstein's inequality.

2.4.2 The Geometry of the Hypercube Revisited

Equipped with the probabilistic tools from the previous sections, we are ready to prove the somewhat counterintuitive properties of hypercubes in high dimensions we discussed in Section 2.2.2.

Theorem 2.18. *Almost all the volume of the high-dimensional cube is located in its corners.*

The proof of this statement will be based on a probabilistic argument, thereby illustrating (again) the nice and fruitful connection between geometry and probability in high dimension. Pick a point at random in the box $[-1,1]^d$. We want to calculate the probability that the point is also in the sphere.

Let $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and each $x_i \in [-1,1]$ is chosen uniformly at random. The event that x also lies in the sphere means

$$\|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2} \leq 1.$$

Let $z_i = x_i^2$ and note that

$$\mathbb{E}[z_i] = \frac{1}{2}\int_{-1}^{1} t^2\, dt = \frac{1}{3} \implies \mathbb{E}[\|x\|_2^2] = \frac{d}{3}$$

and

$$\mathrm{Var}(z_i) = \frac{1}{2}\int_{-1}^{1} t^4\, dt - (\frac{1}{3})^2 = \frac{1}{5} - \frac{1}{9} = \frac{4}{45} \leq \frac{1}{10}$$

Using Hoeffding's inequality,

$$\begin{aligned}\mathbb{P}(\|x\|_2^2 \leq 1) &= \mathbb{P}(\sum_{i=1}^{d} x_i^2 \leq 1) \\ &= \mathbb{P}(\sum_{i=1}^{d} (z_i - \mathbb{E}[z_i]) \leq 1 - \frac{d}{3}) \\ &\leq \exp[-\frac{(\frac{d}{3}-1)^2}{2d\left(\frac{2}{3}\right)^2}] \\ &\leq \exp[-\frac{d}{9}],\end{aligned}$$

for sufficiently large d. Since this value converges to 0 as the dimension d goes to infinity, this shows random points in high cubes are most likely outside the sphere. In other words, almost all the volume of a hypercube concentrates in its corners.

Since we now have gained a better understanding of the properties of the cube in high dimensions, we can use this knowledge to our advantage. For instance, this "pointiness" of the hypercube (in form of the ℓ_1-ball) turns out to very useful in the areas of compressive sensing and sparse recovery, see Chapter 10.

2.4.3 How to Generate Random Points on a Sphere

How can we sample a point uniformly at random from S^{d-1}? The first approach that may come to mind is the following method to generate random points on a unit circle. Independently generate each coordinate uniformly at random from the interval $[-1, 1]$. This yields points that are distributed uniformly at random in a square that contains the unit circle. We could now project all points onto the unit circle. However, the resulting distribution will not be uniform since more points fall on a line from the origin to a vertex of the square, than fall on a line from the origin to the midpoint of an edge due to the difference in length of the diagonal of the square to its side length.

To remedy this problem, we could discard all points outside the unit circle and project the remaining points onto the circle. However, if we generalize this technique to higher dimensions, the analysis in the previous section has shown that the ratio of the volume of $S^{d-1}(1)$ to the volume of $C^d(1)$ decreases rapidly. This makes this process not practical, since almost all the generated points will be discarded in this process and we end up with essentially no points inside (and thus, after projection, on) the sphere.

Instead we can proceed as follows. Recall that the multivariate Gaussian distribution is symmetric about the origin. This rotation invariance is exactly what we need. We simply construct a vector in $\mathbb{R}^d$ whose entries are independently drawn from a univariate Gaussian distribution. We then normalize the

resulting vector to lie on the sphere. This gives a distribution of points that is uniform over the sphere.

Picking a point x uniformly at random on the sphere S^{d-1} is not too different from picking a vector at random with entries of the form $(\pm\frac{1}{\sqrt{d}},\dots,\pm\frac{1}{\sqrt{d}})$, since every point on the sphere has to fullfill $x_1^2+\cdots+x_d^2=1$, hence the "average magnitude" of x_i will be $\frac{1}{\sqrt{d}}$.

Having a method of generating points uniformly at random on S^{d-1} at our disposal, we can now give a probabilistic proof that points on S^{d-1} concentrate near its equator. Without loss of generality we pick an arbitrary unit vector x_1 which represents the "north pole", and the intersection of the sphere with the plane $x_1=0$ forms our equator. We extend x_1 to an orthonormal basis $x_1,\dots,x_d$. We create a random vector by sampling $(Z_1,\dots,Z_d)\sim\mathcal{N}(0,I_d)$ and normalize the vector to get $X=(X_1,\dots,X_d)=\frac{1}{\sum_{k=1}^d Z_k^2}(Z_1,\dots,Z_d)$. Because X is on the sphere, it holds that $\sum_{k=1}^d\langle X,x_k\rangle^2=1$. Note that we also have $\mathbb{E}[\sum_{k=1}^d\langle X,x_k\rangle^2]=\mathbb{E}[1]=1$. Thus, by symmetry, $\mathbb{E}[\langle X,x_1\rangle^2]=\frac{1}{d}$. Applying Markov's inequality (2.10) gives

$$\mathbb{P}(|\langle X,x_1\rangle|>\varepsilon)=\mathbb{P}(\langle X,x_1\rangle^2>\varepsilon^2)\le\frac{\mathbb{E}(\langle X,x_1\rangle^2)}{\varepsilon^2}=\frac{1}{d\varepsilon^2}.$$

For fixed ε we can make this probability arbitrarily small by increasing the dimension d. This proves our claim that points on the high-dimensional sphere concentrate near its equator.

2.4.4 Random Vectors in High Dimensions

Two basic geometric questions from a probabilistic point of view are: (i) What length do we expect a random vector $x\in\mathbb{R}^n$ to have? (ii) What angle do we expect two random vectors $x,y\in\mathbb{R}^n$ to have?

Suppose that the coordinates $x_1,\dots,x_n$ of x are independent random variables with zero mean and unit variances (and similarly for y). It holds that

$$\mathbb{E}\|x\|_2^2=\mathbb{E}\left[\sum_{k=1}^n|x_k|^2\right]=\sum_{k=1}^n\mathbb{E}[|x_k|^2]=n.$$

Hence, we expect the typical length $\|x\|_2$ of x to be approximately $\sqrt{n}$. But how well does the length of a random vector concentrate around its "typical length"?

Assume for instance that the entries $x_k\sim\mathcal{N}(0,1)$. In this case we can use the χ^2-concentration bound (2.18), which gives

$$\mathbb{P}\left(\left|\frac{1}{n}\|x\|_2^2-1\right|\ge t\right)\le 2\exp\left(-\frac{n}{8}\min(t,t^2)\right). \tag{2.25}$$

This represents a concentration inequality for $\|x\|_2^2$, but we aim for a concentration inequality for the length $\|x\|$. To do this we follow a simple but effective trick used in the proof of Theorem 3.1.1 in [138]. We use the following elementary observation that holds for all $z \geq 0$:

$$|z-1| \geq \delta \qquad \text{implies} \qquad |z^2-1| \geq \max(\delta, \delta^2).$$

Using this observation we obtain for any $\delta > 0$ that

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\|x\|_2 - 1\right| \geq \delta\right) \leq \mathbb{P}\left(\left|\frac{1}{n}\|x\|_2^2 - 1\right| \geq \max(\delta, \delta^2)\right) \leq 2e^{-nt^2/8}, \quad (2.26)$$

where we have used $t = \max(\delta, \delta^2)$ in (2.25).

With some minor modifications of these steps (and a slightly different constant) one can extend this result to random vectors with sub-Gaussian coordinates, see e.g. Theorem 3.1.1 in [138].

We now turn our attention to the expected angle between two random vectors. We will show that two randomly drawn vectors in high dimensions are almost perpendicular. The following theorem quantifies this statements. We denote the angle θ_d between two vectors x, y by $\theta_{x,y}$ and recall that $\cos\theta_{x,y} = \frac{\langle x,y\rangle}{\|x\|_2\|y\|_2}$.

Theorem 2.19. *Let $x, y \in \mathbb{R}^d$ be two random vectors with i.i.d. Rademacher variables, i.e. the entries x_i, y_i take values ± 1 with equal probability. Then*

$$\mathbb{P}\left(|\cos\theta_{x,y}| \geq \sqrt{\frac{2\log d}{d}}\right) \leq \frac{2}{d}. \quad (2.27)$$

Proof. Note that $\langle x, y\rangle = \sum_i x_i y_i$ is the sum of i.i.d. Rademacher variables. Hence, $\mathbb{E}[\langle x, y\rangle] = \sum_i \mathbb{E}[x_i y_i] = 0$. Therefore, we can apply Hoeffding's inequality. For any given $t > 0$

$$\mathbb{P}(|\langle x, y\rangle| \geq t) = \mathbb{P}\left(\frac{|\langle x, y\rangle|}{\|x\|_2\|y\|_2} \geq \frac{t}{d}\right) \leq 2\exp\left(\frac{-t^2}{2d}\right).$$

To establish the bound (2.27), we set $t = \sqrt{2d\log d}$ and obtain

$$\mathbb{P}\left(|\cos\theta_{x,y}| > \sqrt{\frac{2\log d}{d}}\right) = \mathbb{P}\left(\frac{|\langle x, y\rangle|}{d} \geq \sqrt{\frac{2\log d}{d}}\right) \leq 2\exp(-\log d) = \frac{2}{d}.$$

It is not surprising that a similar result holds for Gaussian random vectors in $\mathbb{R}^d$ or random vectors chosen from the sphere S^{d-1}. Indeed, even more is true. While we can have only d vectors that are *exactly* orthogonal in $\mathbb{R}^d$, for large d we can have exponentially many vectors that are almost orthogonal in $\mathbb{R}^d$. To see this we return to the setting of Theorem 2.19, choosing m random vectors $x_1, \ldots, x_m$ with i.i.d. Rademacher variables as their entries.

We proceed as in the proof of Theorem 2.19 but let $t = \sqrt{2d \log c}$ where $c > 0$ is a constant. This yields

$$\mathbb{P}\left(|\cos \theta_{x_i,x_j}| \geq \sqrt{\frac{2 \log c}{d}} \right) \leq \frac{2}{c}.$$

Note that we need to consider θ_{x_i,x_j} for $(m^2 - m)/2$ such pairs (x_i, x_j). To make things concrete, we can set for instance $m = \sqrt{c}/4$. Using the union bound we obtain that with probability at least $\frac{7}{8}$ it holds that

$$\max_{i,j,i \neq j} |\cos \theta_{x_i,x_j}| \leq \sqrt{\frac{2 \log c}{d}}.$$

We can now choose e.g. $c = e^{\frac{d}{200}}$ and obtain that we have exponentially many (with respect to d) vectors in $\mathbb{R}^d$ that are almost orthogonal in the sense that the cosine of their pairwise angle is at most $\frac{1}{100}$.

3

Singular Value Decomposition and Principal Component Analysis

Data is most often represented as a matrix, even network data and graphs are often naturally represented by they adjacency matrix. For this reason Linear Algebra is one of the key tools in data analysis. Perhaps more surprising is the fact that spectral properties of matrices representing data play a crucial role in data analysis. After a brief review of Linear Algebra we will illustrate this importance with a discussion of Principal Component Analysis and tools from random matrix theory to better understand its performance in the high dimensional regime.

3.1 Brief review of linear algebra tools

We recommend the reader [68] and [61] as base references in the linear algebra.

Singular Value Decomposition

Singular Value Decomposition (SVD) is one of the most useful tools for analyzing data. Given a matrix $M \in \mathbb{R}^{m\times n}$, the SVD of M is given by

$$M = U\Sigma V^T, \tag{3.1}$$

where $U \in O(m)$, $V \in O(n)$ are orthogonal matrices (meaning that $U^TU = UU^T = \mathrm{I}_{m\times m}$ and $V^TV = VV^T = \mathrm{I}_{n\times n}$) and $\Sigma \in \mathbb{R}^{m\times n}$ is a matrix with non-negative entries on its diagonal and otherwise zero entries.

The columns of U and V are referred to, respectively, as left and right singular vectors of M and the diagonal elements of Σ as singular values of M. Through the SVD, any matrix can be written as a sum of rank-1 matrices

$$M = \sum_{k=1}^{r} \sigma_k u_k v_k^T, \tag{3.2}$$

where $\sigma_1 \geq \sigma_2 \geq \sigma_r > 0$ are the non-zero singular values of M, and u_k and v_k are the corresponding left and right singular vectors. In particular, $\text{rank}(M) = r$, that is, the number of non-zero singular values r is the rank of M.

Remark 3.1. Say $m \leq n$, it is easy to see that we can also think of the SVD as having $U \in \mathbb{R}^{m\times n}$ where $UU^T = \text{I}$, $\Sigma \in \mathbb{R}^{n\times n}$ a diagonal matrix with non-negative entries and $V \in O(n)$.

Matrix norms and low rank matrix approximation

A very powerful modelling tool in data science is low rank matrices. In fact, we will devote whole of Chapter **??** to this topic. As already suggested in the expansion (3.2) the SVD will play an important role in this, being used to provide low rank approximation of data matrices.

In order to be able to talk about low rank approximations of matrices, we need a notion of distance between matrices. Just like with vectors, the distance between matrices can be measured using a suitable norm of the difference. One popular norm is the Frobenius norm, or the Hilbert-Schmidt norm, defined as

$$\|M\|_F = \sqrt{\sum_{i,j} M_{ij}^2}, \tag{3.3}$$

which is simply the Euclidean norm of a vector of length mn of the matrix elements. The Frobenius norm can also be expressed in terms of the singular values. To see this, first express the Frobenius norm in terms of the trace of M^TM as

$$\|M\|_F^2 = \sum_{i,j} M_{ij}^2 = \text{Tr}(M^TM), \tag{3.4}$$

where we recall that the trace of a square matrix A is defined as

$$\text{Tr}(A) = \sum_i A_{ii}. \tag{3.5}$$

A particularly important property of the trace is that for any A of size $m \times n$ and B of size $n \times m$

$$\text{Tr}(AB) = \text{Tr}(BA). \tag{3.6}$$

Note that this implies that, e.g., $\text{Tr}(ABC) = \text{Tr}(CAB)$, but it does not imply that, e.g., $\text{Tr}(ABC) = \text{Tr}(ACB)$ which is not true in general. Now, plugging the SVD (3.1) into (3.4) gives

$$\|M\|_F^2 = \text{Tr}(M^TM) = \text{Tr}(V\Sigma^TU^TU\Sigma V^T) = \text{Tr}(\Sigma^T\Sigma) = \sum_{k=1}^{r} \sigma_k^2, \tag{3.7}$$

where we used the orthogonality of U and V and the trace property (3.6). We conclude that the Frobenius norm equals the Euclidean norm of the vector of singular values.

A different way to define the size of a matrix is by viewing it as an operator and measuring by how much it can dilate vectors. For example, the operator 2-norm is defined as

$$\|M\|_2 = \sup_{\|x\|=1} \|Mx\|. \tag{3.8}$$

Again, this operator norm can be succinctly expressed in terms of the singular values. Indeed, for any $x \in \mathbb{R}^n$

$$Mx = \sum_{k=1}^{r} \sigma_k u_k (v_k^T x). \tag{3.9}$$

Using the orthogonality of the left singular vectors u_k we get

$$\|Mx\|^2 = \sum_{k=1}^{r} \sigma_k^2 \langle v_k, x\rangle^2 \leq \sigma_1^2 \sum_{k=1}^{r} \langle v_k, x\rangle^2 \leq \sigma_1^2 \sum_{k=1}^{n} \langle v_k, x\rangle^2 = \sigma_1^2 \|x\|^2, \tag{3.10}$$

where the last equality is due to the orthogonality of the right singular vectors v_k. Moreover, we get equality by choosing $x = v_1$. We conclude that the 2-norm is simply the largest singular value

$$\|M\|_2 = \sigma_1. \tag{3.11}$$

A very important property of the SVD is that it provides the best low rank approximation of a matrix, when the approximation error is measured in terms of the Frobenius norm. Specifically, for any $0 \leq s \leq r$ consider the rank-s matrix $M_s = \sum_{k=1}^{s} \sigma_k u_k v_k^T$. Then, among all matrices of rank s, M_s best approximates M in terms of the Frobenius norm error. Moreover, the approximation error is given in terms of the remaining $r - s$ smallest singular values as

$$\|M - M_s\|_F = \inf_{B \in \mathbb{R}^{m \times n}, \operatorname{rank}(B) \leq s} \|M - B\|_F = \sqrt{\sum_{k=s+1}^{r} \sigma_k^2} \tag{3.12}$$

A similar result holds for the best low rank approximation in the 2-norm

$$\|M - M_s\|_2 = \inf_{B \in \mathbb{R}^{m \times n}, \operatorname{rank}(B) \leq s} \|M - B\|_2 = \sigma_{s+1} \tag{3.13}$$

In fact, M_s is the best low rank approximation for any univariate matrix norm satisfying $\|UMV\| = \|M\|$ for any $U \in O(m), V \in O(n)$, that is, norms that are invariant to multiplication by orthogonal matrices.

The low rank approximation property has a wide ranging implication on data compression. The storage size of an $m \times n$ data matrix is mn. If that

matrix is of rank r, then storage size reduces from mn to $(n+m+1)r$ (for storing r left and right singular vectors and values). For $r \ll \min\{n,m\}$ this reduction can be quite dramatic. For example, if $r = 10$ and $n = m = 10^6$, then storage reduces from 10^{12} entries to just $2 \cdot 10^7$. But even if the matrix is not precisely of rank r, but only approximately, in the sense that $\sigma_{r+1} \ll \sigma_1$, then we are guaranteed by the above approximation results to incur only a small approximation due to compression using the top r singular vectors and values. In many cases, the singular values of large data matrices decrease very quickly, motivating this type of low rank approximation which oftentimes is the only way to handle massive data sets that otherwise cannot be stored and/or manipulated efficiently. Remarkably, even treating an image as a matrix os pixel intensity values and compressing it this way yields good image compression and de-noising algorithms (as it keeps mitigates the noise corresponding to singular values that are truncated).

Remark 3.2. The computational complexity of computing the SVD of a matrix of size $m \times n$ with $m \geq n$ is $\mathcal{O}(mn^2)$. This cubic scaling could be prohibitive for massive data matrices, and in Chapter **??** we discuss numerical algorithms that use randomization for efficient computation the low rank approximation of such large matrices.

Spectral Decomposition

If $M \in \mathbb{R}^{n\times n}$ is symmetric then it admits a spectral decomposition

$$M = V\Lambda V^T,$$

where $V \in O(n)$ is a matrix whose columns v_k are the eigenvectors of M and Λ is a diagonal matrix whose diagonal elements λ_k are the eigenvalues of M. Similarly, we can write

$$M = \sum_{k=1}^{n} \lambda_k v_k v_k^T.$$

When all of the eigenvalues of M are non-negative we say that M is positive semidefinite and write $M \succeq 0$. In that case we can write

$$M = \left(V\Lambda^{1/2}\right)\left(V\Lambda^{1/2}\right)^T.$$

A decomposition of M of the form $M = UU^T$ (such as the one above) is called a Cholesky decomposition.

For symmetric matrices, the operator 2-norm is also known as the spectral norm, given by

$$\|M\| = \max_k |\lambda_k(M)|.$$

Quadratic Forms

In both this and following chapters, we will be interested in solving problems of the type

$$\max_{\substack{V\in\mathbb{R}^{n\times d}\\ V^TV=\mathrm{I}_{d\times d}}} \operatorname{Tr}\left(V^TMV\right),$$

where M is a symmetric $n\times n$ matrix.

Note that this is equivalent to

$$\max_{\substack{v_1,\ldots,v_d\in\mathbb{R}^n\\ v_i^Tv_j=\delta_{ij}}} \sum_{k=1}^{d} v_k^TMv_k, \tag{3.14}$$

where δ is the Kronecker delta ($\delta_{ij}=1$ for $i=j$ and $\delta_{ij}=0$ otherwise).

When $d=1$ this reduces to the more familiar

$$\max_{\substack{v\in\mathbb{R}^n\\ \|v\|_2=1}} v^TMv. \tag{3.15}$$

It is easy to see (for example, using the spectral decomposition of M) that (3.15) is maximized by the leading eigenvector of M and

$$\max_{\substack{v\in\mathbb{R}^n\\ \|v\|_2=1}} v^TMv=\lambda_{\max}(M).$$

Furthermore (3.14) is maximized by taking $v_1,\ldots,v_d$ to be the k leading eigenvectors of M and its value is simply the sum of the k largest eigenvalues of M. This follows, for example, from a Theorem of Fan (see page 3 of [96]). A fortunate consequence is that the solution to (3.14) can be computed sequentially: we can first solve for $d=1$, computing v_1, then update the solution for $d=2$ by simply computing v_2.

Remark 3.3. All of the tools and results above have natural analogues when the matrices have complex entries (and are Hermitian instead of symmetric).

3.2 Principal Component Analysis and Dimension Reduction

When faced with a high dimensional dataset, a natural approach is to attempt to reduce its dimension, either by projecting it to a lower dimensional space or by finding a better representation for the data using a small number of meaningful features. Beyond data compression and visualization, dimension reduction facilitates downstream analysis such as clustering and regression that perform significantly better in lower dimensions. We will explore a few different ways of reducing the dimension, both linearly and non-linearly.

We will start with the classical Principal Component Analysis (PCA). PCA continues to be one of the most effective and simplest tools for exploratory data analysis. Remarkably, it dates back to a 1901 paper by Karl Pearson [105].

Suppose we have n data points $x_1, \ldots, x_n$ in $\mathbb{R}^p$, for some p, and we are interested in (linearly) projecting the data to $d < p$ dimensions. This is particularly useful if, say, one wants to visualize the data in two or three dimensions ($d = 2, 3$). There are a couple of seemingly different criteria we can use to choose this projection:

1. Finding the d-dimensional affine subspace for which the projections of $x_1, \ldots, x_n$ on it best approximate the original points $x_1, \ldots, x_n$.
2. Finding the d-dimensional projection of $x_1, \ldots, x_n$ that preserves as much variance of the data as possible.

As we will see below, these two approaches are equivalent and they correspond to Principal Component Analysis.

Before proceeding, we recall a couple of simple statistical quantities associated with $x_1, \ldots, x_n$, that will reappear below.

Given $x_1, \ldots, x_n$ we define its sample mean as

$$\mu_n = \frac{1}{n} \sum_{k=1}^{n} x_k, \tag{3.16}$$

and its sample covariance as

$$\Sigma_n = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \mu_n) (x_k - \mu_n)^T . \tag{3.17}$$

Remark 3.4. If $x_1, \ldots, x_n$ are independently sampled from a distribution, μ_n and Σ_n are unbiased estimators for, respectively, the mean and covariance of the distribution.

PCA as the best d-dimensional affine fit

We start with the first interpretation of PCA and then show that it is equivalent to the second. We are trying to approximate each x_k by

$$x_k \approx \mu + \sum_{i=1}^{d} (\beta_k)_i \, v_i, \tag{3.18}$$

where $v_1, \ldots, v_d$ is an orthonormal basis for the d-dimensional subspace, $\mu \in \mathbb{R}^p$ represents the translation, and $\beta_k \in \mathbb{R}^d$ corresponds to the coefficients of x_k. Without loss of generality we take

$$\sum_{k=1}^{n} \beta_k = 0, \tag{3.19}$$

as any joint translation of β_k can be absorbed into μ.

If we represent the subspace by $V = [v_1 \cdots v_d] \in \mathbb{R}^{p\times d}$ then we can rewrite (3.20) as

$$x_k \approx \mu + V\beta_k, \tag{3.20}$$

where $V^TV = \mathrm{I}_{d\times d}$, because the vectors v_i are orthonormal.

We will measure goodness of fit in terms of least squares and attempt to solve

$$\min_{\substack{\mu,\ V,\ \beta_k \\ V^TV=\mathrm{I}}} \sum_{k=1}^{n} \|x_k - (\mu + V\beta_k)\|_2^2 \tag{3.21}$$

We start by optimizing for μ. It is easy to see that the first order condition for μ corresponds to

$$\nabla_\mu \sum_{k=1}^{n} \|x_k - (\mu + V\beta_k)\|_2^2 = 0 \Leftrightarrow \sum_{k=1}^{n} (x_k - (\mu + V\beta_k)) = 0.$$

Thus, the optimal value μ^* of μ satisfies

$$\left(\sum_{k=1}^{n} x_k\right) - n\mu^* - V\left(\sum_{k=1}^{n} \beta_k\right) = 0.$$

Since we assumed in (3.19) that $\sum_{k=1}^{n} \beta_k = 0$, we have that the optimal μ is given by

$$\mu^* = \frac{1}{n}\sum_{k=1}^{n} x_k = \mu_n,$$

the sample mean.

We can then proceed to finding the solution for (3.21) by solving

$$\min_{\substack{V,\ \beta_k \\ V^TV=\mathrm{I}}} \sum_{k=1}^{n} \|x_k - \mu_n - V\beta_k\|_2^2 . \tag{3.22}$$

Let us proceed by optimizing for β_k. The problem almost fully decouples in each k, the only constraint coupling them being (3.19). We will ignore this constraint, solve the decoupled problems, and verify that it is automatically satisfied. Hence we focus on, for each k,

$$\min_{\beta_k} \|x_k - \mu_n - V\beta_k\|_2^2 = \min_{\beta_k} \left\| x_k - \mu_n - \sum_{i=1}^{d} (\beta_k)_i \, v_i \right\|_2^2 . \tag{3.23}$$

Since $v_1, \ldots, v_d$ are orthonormal, it is easy to see that the solution is given by $(\beta_k^*)_i = v_i^T (x_k - \mu_n)$ which can be succinctly written as $\beta_k = V^T (x_k - \mu_n)$, which satisfied (3.19). Thus, (3.22) is equivalent to

$$\min_{V^TV=\mathrm{I}} \sum_{k=1}^{n} \left\| (x_k - \mu_n) - VV^T (x_k - \mu_n) \right\|_2^2 . \tag{3.24}$$

Note that

$$\begin{aligned} \left\| (x_k - \mu_n) - VV^T (x_k - \mu_n) \right\|_2^2 &= (x_k - \mu_n)^T (x_k - \mu_n) \\ &\quad -2 (x_k - \mu_n)^T VV^T (x_k - \mu_n) \\ &\quad + (x_k - \mu_n)^T V \left(V^T V \right) V^T (x_k - \mu_n) \\ &= (x_k - \mu_n)^T (x_k - \mu_n) \\ &\quad - (x_k - \mu_n)^T VV^T (x_k - \mu_n) . \end{aligned}$$

Since $(x_k - \mu_n)^T (x_k - \mu_n)$ does not depend on V, minimizing (3.24) is equivalent to

$$\max_{V^TV=\mathrm{I}} \sum_{k=1}^{n} (x_k - \mu_n)^T VV^T (x_k - \mu_n) . \tag{3.25}$$

A few algebraic manipulations using properties of the trace yields:

$$\begin{aligned} \sum_{k=1}^{n} (x_k - \mu_n)^T VV^T (x_k - \mu_n) &= \sum_{k=1}^{n} \mathrm{Tr} \left[(x_k - \mu_n)^T VV^T (x_k - \mu_n) \right] \\ &= \sum_{k=1}^{n} \mathrm{Tr} \left[V^T (x_k - \mu_n) (x_k - \mu_n)^T V \right] \\ &= \mathrm{Tr} \left[V^T \sum_{k=1}^{n} (x_k - \mu_n) (x_k - \mu_n)^T V \right] \\ &= (n-1) \, \mathrm{Tr} \left[V^T \Sigma_n V \right] . \end{aligned}$$

This means that the solution to (3.25) is given by

$$\max_{V^TV=\mathrm{I}} \mathrm{Tr} \left[V^T \Sigma_n V \right] . \tag{3.26}$$

As we saw above (recall (3.14)) the solution is given by $V = [v_1, \cdots, v_d]$ where $v_1, \ldots, v_d$ correspond to the d leading eigenvectors of Σ_n.

PCA as the *d*-dimensional projection that preserves the most variance

We now show that the alternative interpretation of PCA, of finding the d-dimensional projection of $x_1, \ldots, x_n$ that preserves the most variance, also arrives to the optimization problem (3.26). We aim to find an orthonormal basis $v_1, \ldots, v_d$ (organized as $V = [v_1, \ldots, v_d]$ with $V^TV = \mathrm{I}_{d \times d}$) of a d-dimensional space such that the projection of $x_1, \ldots, x_n$ onto this subspace has the most variance. Equivalently we can ask for the points

$$\left\{ \begin{bmatrix} v_1^T x_k \\ \vdots \\ v_d^T x_k \end{bmatrix} \right\}_{k=1}^{n},$$

to have as much variance as possible. Hence, we are interested in solving

$$\max_{V^T V = \mathrm{I}} \sum_{k=1}^{n} \left\| V^T x_k - \frac{1}{n} \sum_{r=1}^{n} V^T x_r \right\|^2. \tag{3.27}$$

Note that

$$\sum_{k=1}^{n} \left\| V^T x_k - \frac{1}{n} \sum_{r=1}^{n} V^T x_r \right\|^2 = \sum_{k=1}^{n} \left\| V^T (x_k - \mu_n) \right\|^2 = (n-1) \operatorname{Tr}\left(V^T \Sigma_n V \right),$$

showing that (3.27) is equivalent to (3.26) and that the two interpretations of PCA are indeed equivalent.

Finding the Principal Components

When given a dataset $x_1, \ldots, x_n \in \mathbb{R}^p$, in order to compute the Principal Components one needs to compute the leading eigenvectors of

$$\Sigma_n = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \mu_n)(x_k - \mu_n)^T.$$

A naive way of doing this is to construct Σ_n (which takes $\mathcal{O}(np^2)$ work) and then finding its spectral decomposition (which takes $\mathcal{O}(p^3)$ work). This means that the computational complexity of this procedure is $\mathcal{O}\left(\max\left\{np^2, p^3\right\}\right)$ (see [68] or [61]).

An alternative is to use the Singular Value Decomposition (3.1). Let $X = [x_1 \cdots x_n]$ recall that,

$$\Sigma_n = \frac{1}{n-1} \left(X - \mu_n \mathbf{1}^T \right) \left(X - \mu_n \mathbf{1}^T \right)^T.$$

Let us take the SVD of $X - \mu_n \mathbf{1}^T = U_L D U_R^T$ with $U_L \in O(p)$, D diagonal, and $U_R^T U_R = \mathrm{I}$. Then,

$$\Sigma_n = \frac{1}{n-1} \left(X - \mu_n \mathbf{1}^T \right) \left(X - \mu_n \mathbf{1}^T \right)^T = U_L D U_R^T U_R D U_L^T = U_L D^2 U_L^T,$$

meaning that U_L correspond to the eigenvectors of Σ_n. Computing the SVD of $X - \mu_n \mathbf{1}^T$ takes $\mathcal{O}(\min\{n^2 p, p^2 n\})$ work but if one is interested in simply computing the top d eigenvectors then this computational costs reduces to $\mathcal{O}(dnp)$. This can be further improved with randomized algorithms. There are randomized algorithms that compute an approximate solution in

$\mathcal{O}\left(pn \log d + (p+n)d^2\right)$ time (This will be discussed in Chapter **??**. See also, for example, [65, 112, 99]).

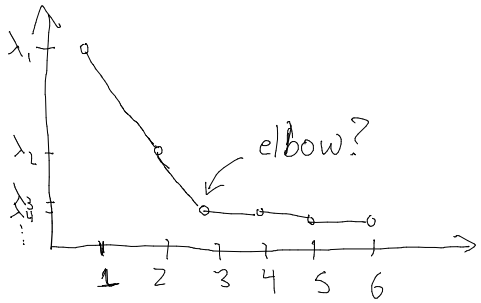
Numerical stability is another important reason why computing the principal components using the SVD is preferable. Since the eigenvalues of Σ_n are proportional to the squares of the singular values of $X - \mu_n \mathbf{1}^T$, problems arise when the ratio of singular values exceeds 10^8, causing the ratio of the corresponding eigenvalues of Σ_n to be larger than 10^{16}. In this case, the smaller eigenvalue would be rounded to zero (due to machine precision), which is certainly not desirable.

Which d should we pick?

Given a dataset, if the objective is to visualize it then picking $d = 2$ or $d = 3$ might make the most sense. However, PCA is useful for many other purposes, for example:

1. Denoising: often times the data belongs to a lower dimensional space but is corrupted by high dimensional noise. When using PCA it is oftentimes possible to reduce the noise while keeping the signal.
2. Downstream analysis: One may be interested in running an algorithm (clustering, regression, etc.) that would be too computationally expensive or too statistically insignificant to run in high dimensions. Dimension reduction using PCA may help there.

In these applications (and many others) it is not clear how to pick d. A fairly popular heuristic is to try to choose the cut-off at a component that has significantly more variance than the one immediately after. Since the total variance is $\mathrm{Tr}(\Sigma_n) = \sum_{k=1}^{p} \lambda_k$, the proportion of variance in the i'th component is nothing but $\frac{\lambda_i}{\mathrm{Tr}(\Sigma_n)}$. A plot of the values of the ordered eigenvalues, also known as a scree plot, helps identify a reasonable choice of d. Here is an example:



It is common to then try to identify an "elbow" on the scree plot to choose the cut-off. In the next Section we will look into Random Matrix Theory to better understand the behavior of the eigenvalues of Σ_n and gain insight into choosing cut-off values.

3.3 PCA in high dimensions and Marčenko-Pastur law

Let us assume that the data points $x_1, \ldots, x_n \in \mathbb{R}^p$ are independent draws of a zero mean Gaussian random variable $g \sim \mathcal{N}(0, \Sigma)$ with some covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. In this case, when we use PCA we are hoping to find a low dimensional structure in the distribution, which should correspond to the large eigenvalues of Σ (and their corresponding eigenvectors). For that reason, and since PCA depends on the spectral properties of Σ_n, we would like to understand whether the spectral properties of the sample covariance matrix Σ_n (eigenvalues and eigenvectors) are close to the ones of Σ, also known as the population covariance.

Since $\mathbb{E}\Sigma_n = \Sigma$, if p is fixed and $n \to \infty$ the law of large numbers guarantees that indeed $\Sigma_n \to \Sigma$. However, in many modern applications it is not uncommon to have p in the order of n (or, sometimes, even larger). For example, if our dataset is composed by images then n is the number of images and p the number of pixels per image; it is conceivable that the number of pixels be on the order of the number of images in a set. Unfortunately, in that case, it is no longer clear that $\Sigma_n \to \Sigma$. Dealing with this type of difficulties is the goal of high dimensional statistics.

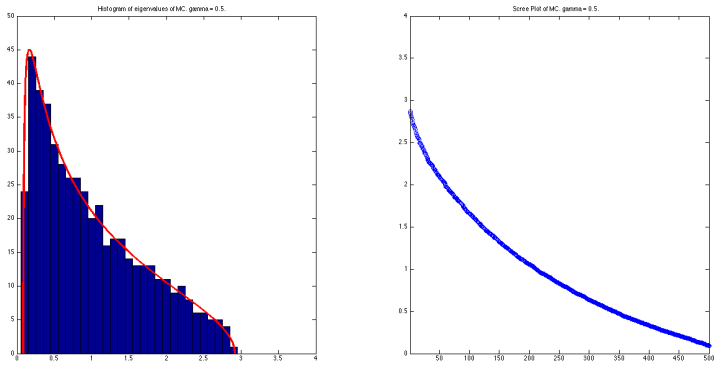
For simplicity we will try to understand the spectral properties of

$$S_n = \frac{1}{n} X X^T,$$

where $x_1, \ldots, x_n$ are the columns of X. Since $x \sim \mathcal{N}(0, \Sigma)$ we know that $\mu_n \to 0$ (and, clearly, $\frac{n}{n-1} \to 1$), hence the spectral properties of S_n will be essentially the same as Σ_n.[1]

Let us start by looking into a simple example, $\Sigma = \mathrm{I}$. In that case, the distribution has no low dimensional structure, as the distribution is rotation invariant. The following is a histogram (left) and a scree plot of the eigenvalues of a sample of S_n (when $\Sigma = \mathrm{I}$) for $p = 500$ and $n = 1000$. The red line is the eigenvalue distribution predicted by the Marčenko-Pastur distribution (3.28), that we will discuss below.

[1] In this case, S_n is actually the maximum likelihood estimator for Σ; we will discuss maximum likelihood estimation later in Chapter **??**.



As one can see in the image, there are many eigenvalues considerably larger than 1, as well as many eigenvalues significantly smaller than 1. Notice that, if given this profile of eigenvalues of Σ_n one could potentially be led to believe that the data has low dimensional structure, when in truth the distribution it was drawn from is isotropic.

Understanding the distribution of eigenvalues of random matrices is in the core of Random Matrix Theory (there are many good books on Random Matrix Theory, e.g. [123] and [11]). This particular limiting distribution was first established in 1967 by Marčenko and Pastur [90] and is now referred to as the Marčenko-Pastur distribution. They showed that, if p and n are both going to ∞ with their ratio fixed $p/n = \gamma \leq 1$, the sample distribution of the eigenvalues of S_n (like the histogram above), in the limit, will be

$$dF_\gamma(\lambda) = \frac{1}{2\pi} \frac{\sqrt{(\gamma_+ - \lambda)(\lambda - \gamma_-)}}{\gamma\lambda} 1_{[\gamma_-,\gamma_+]}(\lambda) d\lambda, \tag{3.28}$$

with support $[\gamma_-, \gamma_+]$, where $\gamma_- = (1-\gamma)^2$, $\gamma_+ = (1+\gamma)^2$, and $\gamma = p/n$. This is plotted as the red line in the figure above.

Remark 3.5. We will not provide the proof of the Marčenko-Pastur law here (you can see, for example, [14] for several different proofs of it), but an approach to a proof is using the so-called moment method. The central idea is to note that one can compute moments of the eigenvalue distribution in two ways and note that (in the limit) for any k,

$$\frac{1}{p}\mathbb{E}\,\mathrm{Tr}\left[\left(\frac{1}{n}XX^T\right)^k\right] = \frac{1}{p}\mathbb{E}\,\mathrm{Tr}\left(S_n^k\right) = \mathbb{E}\frac{1}{p}\sum_{i=1}^{p}\lambda_i^k(S_n) = \int_{\gamma_-}^{\gamma_+}\lambda^k dF_\gamma(\lambda),$$

and that the quantities $\frac{1}{p}\mathbb{E}\,\mathrm{Tr}\left[\left(\frac{1}{n}XX^T\right)^k\right]$ can be estimated (these estimates rely essentially in combinatorics). The distribution $dF_\gamma(\lambda)$ can then be computed from its moments.

3.3.1 Spike Models and BBP phase transition

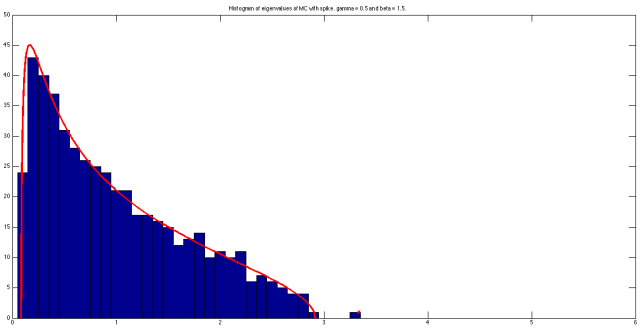
What if there actually is some (linear) low dimensional structure in the data? When can we expect to capture it with PCA? A particularly simple, yet relevant, example to analyze is when the covariance matrix Σ is an identity with a rank 1 perturbation, which we refer to as a spike model $\Sigma = I + \beta uu^T$, for u a unit norm vector and $\beta > 0$.

One way to think about this instance is as each data point x consisting of a signal part $\sqrt{\beta} g_0 u$ where g_0 is a one-dimensional standard Gaussian $\mathcal{N}(0,1)$ (i.e. a normally distributed multiple of a fixed vector $\sqrt{\beta}u$) and a noise part $g \sim \mathcal{N}(0, I)$ (independent of g_0). Then $x = g + \sqrt{\beta} g_0 u$ is a Gaussian random variable

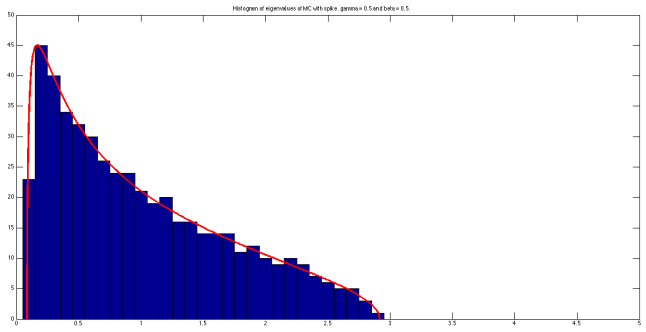
$$x \sim \mathcal{N}(0, I + \beta uu^T).$$

Whereas the signal part $\sqrt{\beta} g_0 u$ resides on a central line in the direction of u, the noise part is high dimensional and isotropic. We therefore refer to β as the signal-to-noise ratio (SNR). Indeed, β is the ratio of the signal variance (in the u-direction) to the noise variance (in each direction).

A natural question is whether this rank-1 perturbation can be seen in S_n. Or equivalently, can one detect the direction of the line u from corrupted measurements in high dimension? Let us build some intuition with an example. The following is the histogram of the eigenvalues of a sample of S_n for $p = 500$, $n = 1000$, u is the first element of the canonical basis $u = e_1$, and $\beta = 1.5$:



The histogram suggests that there is an eigenvalue of S_n that "pops out" of the support of the Marčenko-Pastur distribution (below we will estimate the location of this eigenvalue, and that estimate corresponds to the red "x"). It is worth noting that the largest eigenvalues of Σ is simply $1 + \beta = 2.5$ while the largest eigenvalue of S_n appears considerably larger than that. Let us try now the same experiment for $\beta = 0.5$:



It appears that, for $\beta = 0.5$, the histogram of the eigenvalues is indistinguishable from when $\Sigma = \mathrm{I}$. In particular, no eigenvalue is separated from the Marčenko-Pastur distribution.

This motivates the following question:

Question 3.6. For which values of γ and β do we expect to see an eigenvalue of S_n popping out of the support of the Marčenko-Pastur distribution, and what is the limit value that we expect it to take?

As we will see below, there is a critical value of β, denoted β_c, below which we do not expect to see a change in the distribution of eigenvalues and above which we expect one of the eigenvalues to pop outside of the support. This phenomenon is known as the BBP phase transition (after Baik, Ben Arous, and Péché [15]). There are many very nice papers about this and similar phenomena, including [103, 72, 15, 104, 16, 73, 28, 29]. [2]

In what follows we will find the critical value β_c and estimate the location of the largest eigenvalue of S_n for any β. While the argument we will use can be made precise (and is borrowed from [103]) we will be ignoring a few details for the sake of exposition. In other words, the argument below can be transformed into a rigorous proof, but it is not one at the present form.

We want to understand the behavior of the leading eigenvalue of the sample covariance matrix

$$S_n = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T.$$

Since $x \sim \mathcal{N}(0, I + \beta uu^T)$ we can write $x = (I + \beta uu^T)^{1/2} z$ where $z \sim \mathcal{N}(0, I)$ is an isotropic Gaussian. Then,

$$S_n = \frac{1}{n}\sum_{i=1}^{n} (I + \beta uu^T)^{1/2} z_i z_i^T (I + \beta uu^T)^{1/2} = (I + \beta uu^T)^{1/2} Z_n (I + \beta uu^T)^{1/2},$$

[2] Notice that the Marčenko-Pastur theorem does not imply that all eigenvalues are actually in the support of the Marčenko-Pastur distribution, it just rules out that a non-vanishing proportion are. However, it is possible to show that indeed, in the limit, all eigenvalues will be in the support (see, for example, [103]).

where $Z_n = \frac{1}{n}\sum_{i=1}^n z_i z_i^T$ is the sample covariance matrix of independent isotropic Gaussians. The matrices $S_n = (I+\beta uu^T)^{1/2} Z_n (I+\beta uu^T)^{1/2}$ and $Z_n(I+\beta uu^T)$ are related by a similarity transformation, and therefore have exactly the same eigenvalues. Hence, it suffices to find the leading eigenvalue of the matrix $Z_n(I+\beta uu^T)$, which is a rank-1 perturbation of Z_n (indeed, $Z_n(I+\beta uu^T) = Z_n + \beta Z_n uu^T$). We already know that the eigenvalues of Z_n follow the Marčenko-Pastur distribution, so we are left to understand the effect of a rank-1 perturbation on its eigenvalues.

To find the leading eigenvalue λ of $Z_n(I+\beta uu^T)$, let v be the corresponding eigenvector, that is,

$$Z_n(I+\beta uu^T)v = \lambda v.$$

Subtract $Z_n v$ from both sides to get

$$\beta Z_n uu^T v = (\lambda I - Z_n)v.$$

Assuming λ is not an eigenvalue of Z_n, we can multiply by $(\lambda I - Z_n)^{-1}$ to get[3]

$$\beta(\lambda I - Z_n)^{-1} Z_n uu^T v = v.$$

Our assumption also implies that $u^T v \neq 0$, for otherwise $v = 0$. Multiplying by u^T gives

$$\beta u^T(\lambda I - Z_n)^{-1} Z_n u(u^T v) = u^T v.$$

Dividing by $\beta u^T v$ (which is not 0 as explained above) yields

$$u^T(\lambda I - Z_n)^{-1} Z_n u = \frac{1}{\beta}. \tag{3.29}$$

Suppose $w_1, \ldots, w_p$ are orthonormal eigenvectors of Z_n (with corresponding eigenvalues $\lambda_1, \ldots, \lambda_p$), and expand u in that basis:

$$u = \sum_{i=1}^p \alpha_i w_i.$$

Plugging this expansion in (3.29) gives

$$\sum_{i=1}^p \frac{\lambda_i}{\lambda - \lambda_i}\alpha_i^2 = \frac{1}{\beta} \tag{3.30}$$

For large p, each α_i^2 concentrates around its mean value $\mathbb{E}[\alpha_i^2] = \frac{1}{p}$ (again, this statement can be made rigorous), and (3.30) becomes

$$\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^p \frac{\lambda_i}{\lambda - \lambda_i} = \frac{1}{\beta} \tag{3.31}$$

[3]Intuitively, λ is larger than all the eigenvalues of Z_n, because it corresponds to a perturbation of Z_n by a positive definite matrix βuu^T; yet, a formal justification is beyond the present discussion.

Since the eigenvalues λ_1, λ_p follow the Marčenko-Pastur distribution, the limit on the left hand side can be replaced by the integral

$$\int_{\gamma_-}^{\gamma_+} \frac{t}{\lambda - t}\, dF_\gamma(t) = \frac{1}{\beta} \tag{3.32}$$

Using an integral table (or an integral software), we find that

$$\frac{1}{\beta} = \int_{\gamma_-}^{\gamma_+} \frac{t}{\lambda - t}\, dF_\gamma(t) = \frac{1}{4\gamma}\left[2\lambda - (\gamma_- + \gamma_+) - 2\sqrt{(\lambda - \gamma_-)(\lambda - \gamma_+)}\right]. \tag{3.33}$$

For $\lambda = \gamma_+$, that is, when the top eigenvalue touches the right edge of the Marčenko-Pastur distribution, (3.33) becomes $\frac{1}{4\gamma}(\gamma_+ - \gamma_-)$. This is the critical point that one gets the pop out of the top eigenvalue from the bulk of the Marčenko-Pastur distribution. To calculate the critical value β_c, we recall that $\gamma_- = (1 - \sqrt{\gamma})^2$ and $\gamma_+ = (1 + \sqrt{\gamma})^2$, hence

$$\frac{1}{\beta_c} = \frac{1}{4\gamma}\left((1 + \sqrt{\gamma})^2 - (1 - \sqrt{\gamma})^2\right). \tag{3.34}$$

Therefore, the critical SNR is

$$\beta_c = \sqrt{\gamma} = \sqrt{\frac{p}{n}}. \tag{3.35}$$

When $\beta > \sqrt{\frac{p}{n}}$ one can observe the pop out of the top eigenvalue from the bulk.

Eq. (3.35) illustrates the interplay of the SNR β, the number of samples n, and the dimension p. Low SNR, small sample size, and high dimensionality are all obstacles for detecting linear structure in noisy high dimensional data.

More generally, inverting the relationship between β and λ given by (3.33) (which simply amounts to solving a quadratic), we find that the largest eigenvalue λ of the sample covariance matrix S_n has the limiting value

$$\lambda \to \begin{cases} (\beta + 1)\left(1 + \frac{\gamma}{\beta}\right) & \text{for} \quad \beta \geq \sqrt{\gamma}, \\ \\ (1 + \sqrt{\gamma})^2 & \text{for} \quad \beta < \sqrt{\gamma}. \end{cases} \tag{3.36}$$

In the finite sample case λ will be fluctuating around that value.

Notice that the critical SNR value, $\beta_c = \sqrt{\gamma}$ is buried deep inside the support of the Marčenko-Pastur distribution, because $\sqrt{\gamma} < \gamma_+ = (1 + \sqrt{\gamma})^2$. In other words, the SNR does not have to be greater than the operator norm of the noise matrix in order for it to pop out. We see that the noise effectively pushes the eigenvalue to the right (indeed, $\lambda > \beta$).

The asymptotic squared correlation $|\langle u, v\rangle|^2$ between the top eigenvector v of the sample covariance matrix and true signal vector u can be calculated in a similar fashion. The limiting correlation value turns out to be

$$|\langle v, u\rangle|^2 \to \begin{cases} \frac{1-\frac{\gamma}{\beta^2}}{1+\frac{\gamma}{\beta^2}} & \text{for} \quad \beta \geq \sqrt{\gamma} \\ \\ 0 & \text{for} \quad \beta < \sqrt{\gamma} \end{cases} \tag{3.37}$$

Notice that the correlation value tends to 1 as $\beta \to \infty$, but is strictly less than 1 for any finite SNR.

Wigner matrices

Another very important random matrix model is the Wigner matrix (and it will make appearances in Chapters 6 and 8). Given an integer n, a standard Gaussian Wigner matrix $W \in \mathbb{R}^{n\times n}$ is a symmetric matrix with independent $\mathcal{N}(0,1)$ off-diagonal entries (except for the fact that $W_{ij} = W_{ji}$) and jointly independent $\mathcal{N}(0,2)$ diagonal entries. In the limit, the eigenvalues of $\frac{1}{\sqrt{n}}W$ are distributed according to the so-called semi-circular law

$$dSC(x) = \frac{1}{2\pi}\sqrt{4-x^2}1_{[-2,2]}(x)dx,$$

and there is also a BBP like transition for this matrix ensemble [56]. More precisely, if v is a unit-norm vector in $\mathbb{R}^n$ and $\xi \geq 0$ then the largest eigenvalue of $\frac{1}{\sqrt{n}}W + \xi vv^T$ satisfies

- If $\xi \leq 1$ then
$$\lambda_{\max}\left(\frac{1}{\sqrt{n}}W + \xi vv^T\right) \to 2,$$
- and if $\xi > 1$ then
$$\lambda_{\max}\left(\frac{1}{\sqrt{n}}W + \xi vv^T\right) \to \xi + \frac{1}{\xi}. \tag{3.38}$$

The typical correlation, with v, of the leading eigenvector $v_{\max}$ of $\frac{1}{\sqrt{n}}W + \xi vv^T$ is also known:

- If $\xi \leq 1$ then
$$|\langle v_{\max}, v\rangle|^2 \to 0,$$
- and if $\xi > 1$ then
$$|\langle v_{\max}, v\rangle|^2 \to 1 - \frac{1}{\xi^2}.$$

Form a statistical viewpoint, a central question is to understand for difference distributions of matrices, when is it that it is possible to detect and estimate a spike in a random matrix [106]. When the underlying random matrix corresponds to a random graph and the spike to a bias on distribution of the graph, corresponding to structural properties of the graph the estimates above are able to predict important phase transitions in community detection in networks, as we will see in Chapter 8.

3.3.2 Rank and covariance estimation

The spike model and random matrix theory thus offers a principled way for determining the number of principal components, or equivalently of the rank of the hidden linear structure: simply count the number of eigenvalues to the right of the Marčenko-Pastur distribution. In practice, this approach for rank estimation is often too simplistic for several reasons. First, for actual datasets, n and p are finite, and one needs to take into account non-asymptotic corrections and finite sample fluctuations [77, 78]. Second, the noise may be heteroskedastic (that is, noise variance is different in different directions). Moreover, the noise statistics could also be unknown and it can be non-Gaussian [87]. In some situations it might be possible to estimate the noise statistics directly from the data and to homogenize the noise (a procedure sometimes known as "whitening") [86]. These situations call for careful analysis, and many open problems remain in the field.

Another popular method for rank estimation is using permutation methods. In permutation methods, each column of the data matrix is randomly permutated, so that the low-rank linear structure in the data is destroyed through scrambling, while only the noise is preserved. The process can be repeated multiple times, and the statistics of the singular values of the scrambled data matrices are then used to determine the rank. In particular, only singular values of the original (unscrambled) data matrix that are larger than the largest singular value of the scrambled matrices (taking fluctuations into account of course) are considered as corresponding to signal and are counted towards the rank. The mathematical analysis of permutation methods is another active field of research [47, 48].

In some applications, the objective is to estimate the low rank covariance matrix of the clean signal Σ from the noisy measurements. We saw that in the spike model, the eigenvalues of the sample covariance matrix are inflated due to noise. It is therefore required to shrink the computed eigenvalues of S_n in order to obtain a better estimate of the eigenvalues of Σ. That is, if

$$S_n = \sum_{i=1}^{p} \lambda_i v_i v_i^T$$

is the spectral decomposition of S_n, then we seek an estimator of Σ, denoted $\hat{\Sigma}$ of the form

$$\hat{\Sigma} = \sum_{i=1}^{p} \eta(\lambda_i) v_i v_i^T.$$

The scalar nonlinearity $\eta : \mathbb{R}^+ \to \mathbb{R}^+$ is known as the shrinkage function. An obvious shrinkage procedure is to estimate $\beta = \eta(\lambda)$ from the computed λ by inverting (3.36) (and setting $\beta = 0$ for $\lambda < \gamma_+$). It turns out that this particular shrinker is optimal in terms of the operator norm loss. However, for other loss functions (such as the Frobenius norm loss), the optimal shrinkage function takes a different form [51]. The reason why the shrinker depends

on the loss function is that the eigenvectors of S_n are not perfectly correlated with those of Σ but rather make some non-trivial angle, as in (3.37). In other words, the eigenvectors are noisy, and it may require more aggressive shrinkage to account for that error in the eigenvector. It can be shown that the eigenvector v of the sample covariance is uniformly distributed in a cone around u whose opening angle is given by (3.37). While we can improve the estimation of the eigenvalue via shrinkage, it is however unclear how to improve the estimation of the eigenvector (without any a priori knowledge about it). Finally, we remark that eigenvalue shrinkage also plays an important role in denoising, as will be discussed in Chapter **??**.

4

Graphs, Networks, and Clustering

A crucial part of data science consists of the studying of networks. Network science, or graph theory, unifies the study of diverse types of networks, such as social networks, protein-protein interaction networks, gene-regulation networks, and the internet. In this chapter we introduce graph theory and treat the problem of clustering, to identify similar data points, or vertices, in (network) data.

4.1 PageRank

Before we introduce the formalism of graph theory, we describe the celebrated PageRank algorithm. This algorithm is a principal component[1] behind the web search algorithms, in particular in Google. The goal of PageRank is to quantitatively rate the importance of each page on the web, allowing the search algorithm to rank the pages and thereby present to the user the more important pages first. Search engines such as Google have to carry out three basic steps:[2]

- Crawl the web and locate all, or as many as possible, accessible webpages.
- Index the data of the webpages from step 1, so that they can be searched efficiently for relevant key words or phrases.
- Rate the importance of each page in the database, so that when a user does a search and the subset of pages in the database with the desired information has been found, the more important pages can be presented first.

Here, we will focus on the third step. We follow mainly the derivation in [**?**]. We aim to develop a score of importance for each webpage. A score will be a

[1] It is difficult to resist using this pun.

[2] Another important component of modern search engines is personalization, which we do not discuss here.

non-negative number. A key idea in assigning a score to any given webpage is that the page's score is derived from the links made to that page from other webpages — "A person is important not if it knows a lot of people, but if a lot of people know that person".

Suppose the web of interest contains n pages, each page indexed by an integer k, $1 \leq k \leq n$. A typical example is illustrated in Figure 4.1, in which an arrow from page k to page j indicates a link from page k to page j. Such a web is an example of a directed graph. The links to a given page are called the backlinks for that page. We will use x_k to denote the importance score of page k in the web. x_k is nonnegative and $x_j > x_k$ indicates that page j is more important than page k.

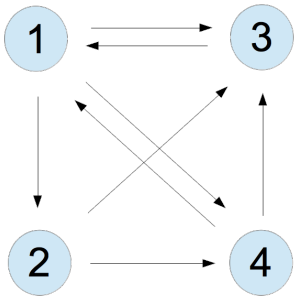


Fig. 4.1: A toy example of the Internet

A very simple approach is to take x_k as the number of backlinks for page k. In the example in Figure 4.1, we have $x_1 = 2, x_2 = 1, x_3 = 3$, and $x_4 = 2$, so that page 3 is the most important, pages 1 and 4 tie for second, and page 2 is least important. A link to page k becomes a vote for page k's importance. This approach ignores an important feature one would expect a ranking algorithm to have, namely, that a link to page k from an important page should boost page k's importance score more than a link from an unimportant page. In the web of Figure 4.1, pages 1 and 4 both have two backlinks: each links to the other, but the second backlink from page 1 is from the seemingly important page 3, while the second backlink for page 4 is from the relatively unimportant page 2. As such, perhaps the algorithm should rate the importance of page 1 higher than that of page 4.

As a first attempt at incorporating this idea, let us compute the score of page j as the sum of the scores of all pages linking to page j. For example, consider the web in our toy example. The score of page 1 would be determined by the relation $x_1 = x_3 + x_4$. However, since x_3 and x_4 will depend on x_1, this seems like a circular definition, since it is self-referential (it is exactly

this self-referential property that will establish a connection to eigenvector problems!).

We also seek a scheme in which a webpage does not gain extra influence simply by linking to lots of other pages. We can do this by reducing the impact of each link, as more and more outgoing links are added to a webpage. If page j contains n_j links, one of which links to page k, then we will boost page k's score by x_j/n_j, rather than by x_j. In this scheme, each webpage gets a total of one vote, weighted by that web page's score, that is evenly divided up among all of its outgoing links. To quantify this for a web of n pages, let $L_k \subset \{1, 2, \ldots, n\}$ denote the set of pages with a link to page k, that is, L_k is the set of page k's backlinks. For each k we require

$$x_k = \sum_{j \in L_k} \frac{x_j}{n_j},$$

where n_j is the number of outgoing links from page j.

If we apply these scheme to the toy example in Figure 4.1, then for page 1 we have $x_1 = x_3/1 + x_4/2$, since pages 3 and 4 are backlinks for page 1 and page 3 contains only one link, while page 4 contains two links (splitting its vote in half). Similarly, $x_2 = x_1/3, x_3 = x_1/3 + x_2/2 + x_4/2$, and $x_4 = x_1/3 + x_2/2$. These conditions can be expressed as linear system of equations $Ax = x$, where $x = [x_1, x_2, x_3, x_4]^T$ and

$$A = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

Thus, we end up with an eigenvalue/eigenvector problem: Find the eigenvector x of the matrix A, associated with the eigenvalue 1. We note that A is a column-stochastic matrix, since it is a square matrix for which all of its entries are nonnegative and the entries in each column sum to 1. Stochastic matrices arise in the study of Markov chains and in a variety of modelling problems in economics and operations research. See e.g. [69] for more details on stochastic matrices. The fact that 1 is an eigenvalue of A is not just coincidence in this example, but holds true in general for stochastic matrices.

Theorem 4.1. *A column-stochastic matrix A has an eigenvalue equal to 1 and 1 is also its largest eigenvalue.*

Proof. Let A be an $n \times n$ column-stochastic matrix. We first note that A and A^T have the same eigenvalues (their eigenvector will usually be different though). Let $\mathbf{1} = [1, 1, \ldots, 1]^T$ be the vector of length n which has all ones as entries. Since A is column-stochastic, we have $A^T\mathbf{1} = \mathbf{1}$ (since all columns of A sum up to 1). Hence $\mathbf{1}$ is an eigenvector of A^T (but not of A) with eigenvalue 1. Thus 1 is also an eigenvalue of A.

To show that 1 is the largest eigenvalue of A we apply the Gershgorin Circle Theorem [69] to A^T. Consider row k of A^T. Let us call the diagonal element $a_{k,k}$ and the radius will be $\sum_{i \neq k} |a_{k,i}| = \sum_{i \neq k} a_{k,i}$ since $a_{k,i} \geq 0$. This is a circle with its center at $a_{k,k} \in [0, 1]$ and with radius $\sum_{i \neq k} a_{k,i} = 1 - a_{k,k}$. Hence, this circle has 1 on its perimeter. This holds for all Gershgorin circles for this matrix. Thus, since all eigenvalues lie in the union of the Gershgorin circles, all eigenvalues λ_i satsify $|\lambda_i| \leq 1$.

In our example, we obtain as eigenvector x of A associated with eigenvalue 1 the vector $x = [x_1, x_2, x_3, x_4]^T$ with entries $x_1 = \frac{12}{31}, x_2 = \frac{4}{31}, x_3 = \frac{9}{31}$, and $x_4 = \frac{6}{31}$. Hence, perhaps somewhat surprisingly, page 3 is no longer the most important one, but page 1. This can be explained by the fact, that the in principle quite important page 3 (which has three webpages linking to it) has only one outgoing link, which gets all its "voting power", and that link points to page 1.

In reality, A can easily be of size a billion times a billion. Fortunately, we do not need compute all eigenvectors of A, only the eigenvector associated with the eigenvalue 1, which, as we know, is also the largest eigenvalue of A. This in turn means we can resort to standard *power iteration* to compute x fairly efficiently (and we can also make use of the fact that A will be a sparse matrix, i.e., many of its entries will be zero). The actual PageRank algorithms adds some minor modifications, but the essential idea is as described above.

4.2 Graph Theory

We now introduce the formalism for undirected[3] graphs, one of the main objects of study in what follows. A graph $G = (V, E)$ contains a set of nodes $V = \{v_1, \ldots, v_n\}$ and edges $E \subseteq \binom{V}{2}$. An edge $(i, j) \in E$ if v_i and v_j are connected. Here is one of the graph theorists favorite examples, the Petersen graph[4]:

Let us recall some concepts about graphs that we will need.

- A graph is connected if, for all pairs of vertices, there is a path between these vertices on the graph. The number of connected components is simply the size of the smallest partition of the nodes into connected subgraphs. The Petersen graph is connected (and thus it has only 1 connected component).
- A clique of a graph G is a subset S of its nodes such that the subgraph corresponding to it is complete. In other words S is a clique if all pairs of vertices in S share an edge. The clique number $c(G)$ of G is the size of the largest clique of G. The Petersen graph has a clique number of 2.

[3] The previous Section featured directed graphs, in which edges (links) have a meaningful direction. In what follows we will focus in undirected graphs in which an edge represents a connection, without meaningful direction.

[4] The Peterson graph is often used as a counter-example in graph theory.

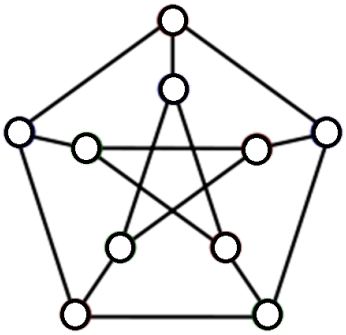


Fig. 4.2: The Petersen graph

- An independence set of a graph G is a subset S of its nodes such that no two nodes in S share an edge. Equivalently it is a clique of the complement graph $G^c := (V, E^c)$. The independence number of G is simply the clique number of S^c. The Petersen graph has an independence number of 4.

A particularly useful way to represent a graph is through its adjacency matrix. Given a graph $G = (V, E)$ on n nodes ($|V| = n$), we define its adjacency matrix $A \in \mathbb{R}^{n \times n}$ as the symmetric matrix with entries

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Sometime, we will consider weighted graphs $G = (V, E, W)$, where edges may have weights w_{ij}, we think of the weights as non-negative $w_{ij} \geq 0$ and symmetric $w_{ij} = w_{ji}$.

Much of the sequel will deal with graphs. Chapter **??** will treat (network) data visualization, dimension reduction, and embeddings of graphs on Euclidean space. Chapter 8 will introduce and study important random graph models. The rest of this Chapter will be devoted to clustering.

4.3 Clustering

Clustering is one of the central tasks in machine learning. Given a set of data points, or nodes of a graph, the purpose of clustering is to partition the data into a set of clusters where data points assigned to the same cluster correspond to similar data points (depending on the context, it could be for example having small distance to each other if the points are in Euclidean space, or being connected if on a graph). We will start with an example of clustering points in Euclidean space, and later move back to graphs.

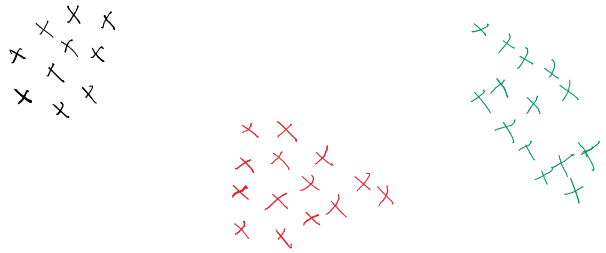


Fig. 4.3: Examples of points separated in clusters.

4.3.1 *k*-means Clustering

One the most popular methods used for clustering is k-means clustering. Given $x_1, \ldots, x_n \in \mathbb{R}^p$ the k-means clustering partitions the data points in clusters $S_1 \cup \cdots \cup S_k$ with centers $\mu_1, \ldots, \mu_k \in \mathbb{R}^p$ as the solution to:

$$\min_{\substack{\text{partition } S_1,\ldots,S_k \\ \mu_1,\ldots,\mu_k}} \sum_{l=1}^{k} \sum_{i \in S_i} \|x_i - \mu_l\|^2 . \tag{4.1}$$

Note that, given the partition, the optimal centers are given by

$$\mu_l = \frac{1}{|S_l|} \sum_{i \in S_l} x_i.$$

Lloyd's algorithm [88] (also sometimes known as the k-means algorithm), is an iterative algorithm that alternates between

- Given centers $\mu_1, \ldots, \mu_k$, assign each point x_i to the cluster

$$l = \text{argmin}_{l=1,\ldots,k} \|x_i - \mu_l\| .$$

- Update the centers $\mu_l = \frac{1}{|S_l|} \sum_{i \in S_l} x_i$.

Unfortunately, Lloyd's algorithm is not guaranteed to converge to the solution of (4.1). Indeed, Lloyd's algorithm oftentimes gets stuck in local optima of (4.1). In the sequel we will discuss convex relaxations for clustering, which can be used as an alternative algorithmic approach to Lloyd's algorithm, but since optimizing (4.1) is NP-hard there is no polynomial time algorithm that works in the worst-case (assuming the widely believed conjecture $P \neq NP$, see also Chapter 7)

While popular, k-means clustering has some potential issues:

- One needs to set the number of clusters a priori (a typical way to overcome this issue is by trying the algorithm for different number of clusters).
- The way (4.1) is defined it needs the points to be defined in an Euclidean space, oftentimes we are interested in clustering data for which we only have some measure of affinity between different data points, but not necessarily an embedding in $\mathbb{R}^p$ (this issue can be overcome by reformulating (4.1) in terms of distances only).
- The formulation is computationally hard, so algorithms may produce sub-optimal instances.
- The solutions of k-means are always convex clusters. This means that k-means may have difficulty in finding cluster such as in Figure 4.4.

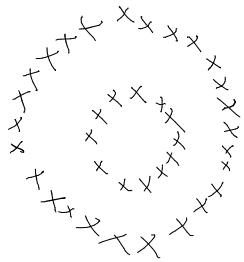


Fig. 4.4: Because the solutions of k-means are always convex clusters, it is not able to handle some cluster structures.

4.3.2 Spectral Clustering

A natural way to try to overcome the issues of k-means depicted in Figure 4.4 is by using transforming the data into a graph and cluster the graph: Given the data points we can construct a weighted graph $G = (V, E, W)$ using a similarity kernel K_ε, such as $K_\varepsilon(u) = \exp\left(\frac{1}{2\varepsilon}u^2\right)$, by associating each point to a vertex and, for which pair of nodes, set the edge weight as

$$w_{ij} = K_\varepsilon\left(\|x_i - x_j\|\right).$$

Other popular procedures to transform data into a graph is by constructing the graph where data points are connected if they correspond to the nearest neighbours. We note that this procedures only needs a measure of distance, or similarity, of data points and not necessarily that they lie in Euclidean Space. Given this motivation, and the prevalence of network data, we will now address the problem of clustering the nodes of a graph.

Normalized Cut

Given a graph $G = (V, E, W)$, the goal is to partition the graph in clusters in a way that keeps as many of the edges, or connections, within the clusters and has as few edges as possible across clusters. We will focus on the case of two clusters, and briefly address extensions in the end of this chapter. A natural way to measure a vertex partition (S, S^c) is

$$\text{cut}(S) = \sum_{i \in S} \sum_{j \in S^c} w_{ij}.$$

If we represent the partition by a vector $y \in \{\pm 1\}^n$ where $y_i = 1$ is $i \in S$, and $y_i = -1$ otherwise, then the cut is a quadratic form on the Graph Laplacian.

Definition 4.2 (Graph Laplacian and Degree Matrix). *Let $G = (V, E, W)$ be a graph and W the matrix of weights (or adjacency matrix if the graph is unweighted). The degree matrix D is a diagonal matrix with diagonal entries*

$$D_{ii} = \deg(i).$$

The graph Laplacian of G is given by

$$L_G = D - W.$$

Equivalently

$$L_G := \sum_{i<j} w_{ij} \left(e_i - e_j\right) \left(e_i - e_j\right)^T.$$

Note that the entries of L_G are given by

$$(L_G)_{ij} = \begin{cases} -w_{ij} & \text{if } i \neq j \\ \deg(i) & \text{if } i = j. \end{cases}$$

If $S \subset V$ and $y \in \{\pm 1\}^n$ such that $y_i = 1$ is $i \in S$, and $y_i = -1$ otherwise, then it is easy to see that

$$\text{cut}(S) = \frac{1}{4} \sum_{i<j} w_{ij} (y_i - y_j)^2.$$

The following proposition establishes

$$\text{cut}(S) = \frac{1}{4} y^T L_G y, \tag{4.2}$$

for $y \in \{\pm 1\}^n$ such that $y_i = 1$ if and only if $i \in S$.

Proposition 4.3. *Let $G = (V, E, W)$ be a graph and L_G its graph Laplacian, let $x \in \mathbb{R}^n$*

$$x^T L_G x = \sum_{i<j} w_{ij} (x_i - x_j)^2$$

Proof.

$$\begin{aligned}\sum_{i<j} w_{ij}\left(x_i - x_j\right)^2 &= \sum_{i<j} w_{ij}\left[\left(e_i - e_j\right)x\right]\left[\left(e_i - e_j\right)x\right]^T \\ &= \sum_{i<j} w_{ij}\left[\left(e_i - e_j\right)^T x\right]^T\left[\left(e_i - e_j\right)^T x\right] \\ &= \sum_{i<j} w_{ij} x^T\left(e_i - e_j\right)\left(e_i - e_j\right)^T x \\ &= x^T\left[\sum_{i<j} w_{ij}\left(e_i - e_j\right)\left(e_i - e_j\right)^T\right] x\end{aligned}$$

□

While $\mathrm{cut}(S)$ is a good way of measuring the fit of a partition, it suffers from an issue: the minimum cut is achieved for $S = \emptyset$ (since $\mathrm{cut}(\emptyset) = 0$) which is a rather meaningless choice of partition. Simply constraining the partition to be non-trivial would still have soft versions of this issue, it would favour very unbalanced partitions. Below we discuss how to promote (almost) balanced partitions.

Remark 4.4. One simple way to address this is to simply ask for an exactly balanced partition, $|S| = |S^c|$ (let us assume the number of vertices $n = |V|$ is even). We can then identify a partition with a label vector $y \in \{\pm 1\}^n$ where $y_i = 1$ is $i \in S$, and $y_i = -1$ otherwise. Also, the balanced condition can be written as $\sum_{i=1}^n y_i = 0$. This means that we can write the minimum balanced cut as

$$\min_{\substack{S\subset V \\ |S|=|S^c|}} \mathrm{cut}(S) = \frac{1}{4}\min_{\substack{y\in\{-1,1\}^n \\ \mathbf{1}^T y = 0}} y^T L_G y,$$

which is suggestive of the connection between clustering and spectral properties of L_G. This connection will be made precise below.

Asking for the partition to be exactly balanced is too restrictive in many cases. There are several ways to evaluate a partition that are variations of $\mathrm{cut}(S)$ that take into account the intuition that one wants both S and S^c to not be too small (although not necessarily equal to $|V|/2$). A prime example is Cheeger's cut.

Definition 4.5 (Cheeger's cut). *Given a graph and a vertex partition (S, S^c), the cheeger cut (also known as conductance, or expansion) of S is given by*

$$h(S) = \frac{\mathrm{cut}(S)}{\min\{\mathrm{vol}(S), \mathrm{vol}(S^c)\}},$$

where $\mathrm{vol}(S) = \sum_{i\in S} \deg(i)$.

Also, the Cheeger's constant of G is given by

$$h_G = \min_{S \subset V} h(S).$$

A similar object is the Normalized Cut, Ncut, which is given by

$$\mathrm{Ncut}(S) = \frac{\mathrm{cut}(S)}{\mathrm{vol}(S)} + \frac{\mathrm{cut}(S^c)}{\mathrm{vol}(S^c)}.$$

Note that $\mathrm{Ncut}(S)$ and $h(S)$ are tightly related, in fact it is easy to see that:

$$h(S) \leq \mathrm{Ncut}(S) \leq 2h(S).$$

Normalized Cut as a spectral relaxation

Below we will show that Ncut can be written in terms of a minimization of a quadratic form involving the graph Laplacian L_G, analogously to the balanced partition as described on Remark 4.4.

Recall that balanced partition can be written as

$$\frac{1}{4} \min_{\substack{y \in \{-1,1\}^n \\ \mathbf{1}^T y = 0}} y^T L_G y.$$

An intuitive way to relax the balanced condition is to allow the labels y to take values in two different real values a and b (e.g. $y_i = a$ if $i \in S$ and $y_j = b$ if $i \notin S$) but not necessarily ± 1. We can then use the notion of volume of a set to ensure a less restrictive notion of balanced by asking that

$$a\,\mathrm{vol}\,(S) + b\,\mathrm{vol}\,(S^c) = 0, \tag{4.3}$$

where

$$\mathrm{vol}(S) = \sum_{i \in S} \deg(i). \tag{4.4}$$

Thus (4.3) corresponds to $\mathbf{1}^T D y = 0$.

We also need to fix a scale for a and b:

$$a^2\,\mathrm{vol}\,(S) + b^2\,\mathrm{vol}\,(S^c) = 1,$$

which corresponds to $y^T D y = 1$.

This suggests considering

$$\min_{\substack{y \in \{a,b\}^n \\ \mathbf{1}^T D y = 0,\, y^T D y = 1}} y^T L_G y.$$

As we will see below, this corresponds precisely to Ncut.

Proposition 4.6. *For a and b to satisfy $a\,\mathrm{vol}\,(S) + b\,\mathrm{vol}\,(S^c) = 0$ and $a^2\,\mathrm{vol}\,(S) + b^2\,\mathrm{vol}\,(S^c) = 1$ it must be that*

$$a = \left(\frac{\mathrm{vol}(S^c)}{\mathrm{vol}(S)\,\mathrm{vol}(G)}\right)^{\frac{1}{2}} \quad \text{and} \quad b = -\left(\frac{\mathrm{vol}(S)}{\mathrm{vol}(S^c)\,\mathrm{vol}(G)}\right)^{\frac{1}{2}},$$

corresponding to

$$y_i = \begin{cases} \left(\frac{\mathrm{vol}(S^c)}{\mathrm{vol}(S)\,\mathrm{vol}(G)}\right)^{\frac{1}{2}} & \text{if } i \in S \\ -\left(\frac{\mathrm{vol}(S)}{\mathrm{vol}(S^c)\,\mathrm{vol}(G)}\right)^{\frac{1}{2}} & \text{if } i \in S^c. \end{cases}$$

Note that vol *is defined as* (4.4).

Proof. The proof involves only doing simple algebraic manipulations together with noticing that $\mathrm{vol}(S) + \mathrm{vol}(S^c) = \mathrm{vol}(G)$. □

Proposition 4.7.

$$\mathrm{Ncut}(S) = y^T L_G y,$$

where y is given by

$$y_i = \begin{cases} \left(\frac{\mathrm{vol}(S^c)}{\mathrm{vol}(S)\,\mathrm{vol}(G)}\right)^{\frac{1}{2}} & \text{if } i \in S \\ -\left(\frac{\mathrm{vol}(S)}{\mathrm{vol}(S^c)\,\mathrm{vol}(G)}\right)^{\frac{1}{2}} & \text{if } i \in S^c. \end{cases}$$

Proof.

$$\begin{aligned} y^T L_G y &= \frac{1}{2}\sum_{i,j} w_{ij}(y_i - y_j)^2 \\ &= \sum_{i\in S}\sum_{j\in S^c} w_{ij}(y_i - y_j)^2 \\ &= \sum_{i\in S}\sum_{j\in S^c} w_{ij}\left[\left(\frac{\mathrm{vol}(S^c)}{\mathrm{vol}(S)\,\mathrm{vol}(G)}\right)^{\frac{1}{2}} + \left(\frac{\mathrm{vol}(S)}{\mathrm{vol}(S^c)\,\mathrm{vol}(G)}\right)^{\frac{1}{2}}\right]^2 \\ &= \sum_{i\in S}\sum_{j\in S^c} w_{ij}\frac{1}{\mathrm{vol}(G)}\left[\frac{\mathrm{vol}(S^c)}{\mathrm{vol}(S)} + \frac{\mathrm{vol}(S)}{\mathrm{vol}(S^c)} + 2\right] \\ &= \sum_{i\in S}\sum_{j\in S^c} w_{ij}\frac{1}{\mathrm{vol}(G)}\left[\frac{\mathrm{vol}(S^c)}{\mathrm{vol}(S)} + \frac{\mathrm{vol}(S)}{\mathrm{vol}(S^c)} + \frac{\mathrm{vol}(S)}{\mathrm{vol}(S)} + \frac{\mathrm{vol}(S^c)}{\mathrm{vol}(S^c)}\right] \\ &= \sum_{i\in S}\sum_{j\in S^c} w_{ij}\left[\frac{1}{\mathrm{vol}(S)} + \frac{1}{\mathrm{vol}(S^c)}\right] \\ &= \mathrm{cut}(S)\left[\frac{1}{\mathrm{vol}(S)} + \frac{1}{\mathrm{vol}(S^c)}\right] \\ &= \mathrm{Ncut}(S). \end{aligned}$$

□

This means that finding the minimum Ncut corresponds to solving

$$\begin{aligned} &\min\ y^T L_G y \\ &\text{s. t. } y \in \{a,b\}^n \text{ for some } a \text{ and } b \\ &\quad y^T D y = 1 \\ &\quad y^T D\mathbf{1} = 0. \end{aligned} \tag{4.5}$$

Since solving (4.5) is, in general, NP-hard, we consider a similar problem where the constraint that y can only take two values is removed:

$$\begin{aligned} &\min\ y^T L_G y \\ &\text{s. t. } y \in \mathbb{R}^n \\ &\quad y^T D y = 1 \\ &\quad y^T D\mathbf{1} = 0. \end{aligned} \tag{4.6}$$

Given a solution of (4.6) we can *round* it to a partition by setting a threshold τ and taking $S = \{i \in V : y_i \leq \tau\}$. We will see below that (4.6) is an eigenvector problem (for this reason we call (4.6) a spectral relaxation).

In order to better see that (4.6) is an eigenvector problem (and thus computationally tractable), set $z = D^{\frac{1}{2}} y$ and

$$\mathcal{L}_G = D^{-\frac{1}{2}} L_G D^{-\frac{1}{2}}, \tag{4.7}$$

then (4.6) is equivalent

$$\begin{aligned} &\min\ z^T \mathcal{L}_G z \\ &\text{s. t. } z \in \mathbb{R}^n \\ &\quad \|z\|^2 = 1 \\ &\quad \left(D^{\frac{1}{2}}\mathbf{1}\right)^T z = 0. \end{aligned} \tag{4.8}$$

Note that $\mathcal{L}_G = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. We order its eigenvalues in increasing order $0 = \lambda_1(\mathcal{L}_G) \leq \lambda_2(\mathcal{L}_G) \leq \cdots \leq \lambda_n(\mathcal{L}_G)$. The eigenvector associated to the smallest eigenvector is given by $D^{\frac{1}{2}}\mathbf{1}$ this means that (by the variational interpretation of the eigenvalues) that the minimum of (4.8) is $\lambda_2(\mathcal{L}_G)$ and the minimizer is given by the second smallest eigenvector of $\mathcal{L}_G = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, which we call v_2. Note that this corresponds also to the second largest eigenvector of $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. This means that the optimal y in (4.6) is given by $\varphi_2 = D^{-\frac{1}{2}} v_2$. This motivates Algorithm 4.1, which is often referred to as Spectral Clustering:

Because the relaxation (4.6) is obtained from (4.5) by removing a constraint we immediately have that

$$\lambda_2(\mathcal{L}_G) \leq \min_{S \subset V} \mathrm{Ncut}(S).$$

This means that

Algorithm 4.1 Spectral Clustering

Given a graph $G = (V, E, W)$, let v_2 be the eigenvector corresponding to the second smallest eigenvalue of the normalized Laplacian $\mathcal{L}_G$, as defined in (4.7). Let $\varphi_2 = D^{-\frac{1}{2}} v_2$. Given a threshold τ (one can try all different possibilities, or run k-means for $k = 2$), set

$$S = \{i \in V : \varphi_2(i) \leq \tau\}.$$

$$\frac{1}{2}\lambda_2 (\mathcal{L}_G) \leq h_G.$$

In what follows we will show a guarantee for Algorithm 4.1.

Lemma 4.8. *There is a threshold τ producing a partition S such that*

$$h(S) \leq \sqrt{2\lambda_2 (\mathcal{L}_G)}.$$

This implies in particular that

$$h(S) \leq \sqrt{4h_G},$$

meaning that Algorithm 4.1 is suboptimal at most by a square-root factor.

Note that this also directly implies the famous Cheeger's Inequality

Theorem 4.9 (Cheeger's Inequality). *Recall the definitions above. The following holds:*

$$\frac{1}{2}\lambda_2 (\mathcal{L}_G) \leq h_G \leq \sqrt{2\lambda_2 (\mathcal{L}_G)}.$$

Cheeger's inequality was first established for manifolds by Jeff Cheeger in 1970 [41], the graph version is due to Noga Alon and Vitaly Milman [7, 9] in the mid 80s.

The upper bound in Cheeger's inequality (corresponding to Lemma 4.8) is more interesting but more difficult to prove, it is often referred to as the "the difficult part" of Cheeger's inequality. We will prove this Lemma in what follows. There are several proofs of this inequality (see [42] for four different proofs!). The proof that follows is an adaptation of the proof in this blog post [127] for the case of weighted graphs.

Proof. [of Lemma 4.8]

We will show that given $y \in \mathbb{R}^n$ satisfying

$$\mathcal{R}(y) := \frac{y^T L_G y}{y^T D y} \leq \delta,$$

and $y^T D\mathbf{1} = 0$. there is a "rounding of it", meaning a threshold τ and a corresponding choice of partition

$$S = \{i \in V : y_i \leq \tau\}$$

such that

$$h(S) \leq \sqrt{2\delta},$$

since $y = \varphi_2$ satisfies the conditions and gives $\delta = \lambda_2(\mathcal{L}_G)$ this proves the Lemma.

We will pick this threshold at random and use the probabilistic method to show that at least one of the thresholds works.

First we can, without loss of generality, assume that $y_1 \leq \cdot \leq y_n$ (we can simply relabel the vertices). Also, note that scaling of y does not change the value of $\mathcal{R}(y)$. Also, if $y^T D\mathbf{1} = 0$ adding a multiple of $\mathbf{1}$ to y can only decrease the value of $\mathcal{R}(y)$: the numerator does not change and the denominator $(y + c\mathbf{1})^T D(y + c\mathbf{1}) = y^T Dy + c^2 \mathbf{1}^T D\mathbf{1} \geq y^T Dy$.

This means that we can construct (from y by adding a multiple of $\mathbf{1}$ and scaling) a vector x such that

$$x_1 \leq ... \leq x_n, \ x_m = 0, \ \text{ and } x_1^2 + x_n^2 = 1,$$

and

$$\frac{x^T L_G x}{x^T D x} \leq \delta,$$

where m be the index for which $\text{vol}(\{1, \ldots, m-1\}) \leq \text{vol}(\{m, \ldots, n\})$ but $\text{vol}(\{1, \ldots, m\}) > \text{vol}(\{m, \ldots, n\})$.

We consider a random construction of S with the following distribution. $S = \{i \in V : x_i \leq \tau\}$ where $\tau \in [x_1, x_n]$ is drawn at random with the distribution

$$\mathbb{P}\{\tau \in [a, b]\} = \int_a^b 2|\tau| d\tau,$$

where $x_1 \leq a \leq b \leq x_n$.

It is not difficult to check that

$$\mathbb{P}\{\tau \in [a, b]\} = \begin{cases} \left|b^2 - a^2\right| & \text{if } a \text{ and } b \text{ have the same sign} \\ a^2 + b^2 & \text{if } a \text{ and } b \text{ have different signs} \end{cases}$$

Let us start by estimating $\mathbb{E}\,\text{cut}(S)$.

$$\begin{aligned} \mathbb{E}\,\text{cut}(S) &= \mathbb{E}\frac{1}{2}\sum_{i \in V}\sum_{j \in V} w_{ij} \mathbf{1}_{(S, S^c) \text{ cuts the edge } (i,j)} \\ &= \frac{1}{2}\sum_{i \in V}\sum_{j \in V} w_{ij} \mathbb{P}\{(S, S^c) \text{ cuts the edge } (i, j)\} \end{aligned}$$

Note that $\mathbb{P}\{(S, S^c) \text{ cuts the edge } (i, j)\}$ is $\left|x_i^2 - x_j^2\right|$ is x_i and x_j have the same sign and $x_i^2 + x_j^2$ otherwise. Both cases can be conveniently upper bounded by $|x_i - x_j|(|x_i| + |x_j|)$. This means that

$$\begin{aligned}\mathbb{E}\,\mathrm{cut}(S) &\leq \frac{1}{2}\sum_{i,j} w_{ij}\,|x_i - x_j|\,(|x_i| + |x_j|) \\ &\leq \frac{1}{2}\sqrt{\sum_{ij} w_{ij}(x_i - x_j)^2}\sqrt{\sum_{ij} w_{ij}(|x_i| + |x_j|)^2},\end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality.

From the construction of x we know that

$$\sum_{ij} w_{ij}(x_i - x_j)^2 = 2x^T L_G x \leq 2\delta x^T D x.$$

Also,

$$\sum_{ij} w_{ij}(|x_i|+|x_j|)^2 \leq \sum_{ij} w_{ij} 2x_i^2 + 2x_j^2. = 2\left(\sum_i \deg(i) x_i^2\right) + 2\left(\sum_j \deg(j) x_j^2\right) = 4x^T D x.$$

This means that

$$\mathbb{E}\,\mathrm{cut}(S) \leq \frac{1}{2}\sqrt{2\delta x^T D x}\sqrt{4x^T D x} = \sqrt{2\delta}\, x^T D x.$$

On the other hand,

$$\mathbb{E}\min\{\mathrm{vol}\, S, \mathrm{vol}\, S^c\} = \sum_{i=1}^n \deg(i)\mathbb{P}\{x_i \text{ is in the smallest set (in terms of volume)}\},$$

to break ties, if $\mathrm{vol}(S) = \mathrm{vol}(S^c)$ we take the "smallest" set to be the one with the first indices.

Note that m is always in the largest set. Any vertex $j < m$ is in the smallest set if $x_j \leq \tau \leq x_m = 0$ and any $j > m$ is in the smallest set if $0 = x_m \leq \tau \leq x_j$. This means that,

$$\mathbb{P}\{x_i \text{ is in the smallest set (in terms of volume)} = x_j^2.$$

Which means that

$$\mathbb{E}\min\{\mathrm{vol}\, S, \mathrm{vol}\, S^c\} = \sum_{i=1}^n \deg(i) x_i^2 = x^T D x.$$

Hence,

$$\frac{\mathbb{E}\,\mathrm{cut}(S)}{\mathbb{E}\min\{\mathrm{vol}\, S, \mathrm{vol}\, S^c\}} \leq \sqrt{2\delta}.$$

Note however that because $\frac{\mathbb{E}\,\mathrm{cut}(S)}{\mathbb{E}\min\{\mathrm{vol}\, S, \mathrm{vol}\, S^c\}}$ is not necessarily the same as $\mathbb{E}\frac{\mathrm{cut}(S)}{\min\{\mathrm{vol}\, S, \mathrm{vol}\, S^c\}}$ and so, we do not necessarily have

$$\mathbb{E}\frac{\text{cut}(S)}{\min\{\text{vol}\, S, \text{vol}\, S^c\}} \leq \sqrt{2\delta}.$$

However, since both random variables are positive,

$$\mathbb{E}\,\text{cut}(S) \leq \mathbb{E}\min\{\text{vol}\, S, \text{vol}\, S^c\}\sqrt{2\delta},$$

or equivalently

$$\mathbb{E}\left[\text{cut}(S) - \min\{\text{vol}\, S, \text{vol}\, S^c\}\sqrt{2\delta}\right] \leq 0,$$

which guarantees, by the probabilistic method, the existence of S such that

$$\text{cut}(S) \leq \min\{\text{vol}\, S, \text{vol}\, S^c\}\sqrt{2\delta},$$

which is equivalent to

$$h(S) = \frac{\text{cut}(S)}{\min\{\text{vol}\, S, \text{vol}\, S^c\}} \leq \sqrt{2\delta},$$

which concludes the proof of the Lemma. □

Multiple Clusters

Much of the above can be easily adapted to multiple clusters. Algorithm 4.2 is a natural extension of spectral clustering to multiple clusters.[5]

Algorithm 4.2 Spectral Clustering

Given a graph $G = (V, E, W)$, let $v_2, \ldots, v_k$ be the eigenvectors corresponding to the second through $(k-1)$th eigenvalues of the normalized Laplacian $\mathcal{L}_G$, as defined in (4.7). Let $\varphi_m = D^{-\frac{1}{2}} v_m$. Consider the map $\varphi : V \to \mathbb{R}^{k-1}$ defined as

$$\varphi(v_i) = \begin{bmatrix} \varphi_2(i) \\ \vdots \\ \varphi_k(i) \end{bmatrix}.$$

Cluster the n points in $k-1$ dimensions into k clusters using k-means.

There is also an analogue of Cheeger's inequality. A natural way of evaluating k-way clustering is via the k-way expansion constant (see [85]):

$$\rho_G(k) = \min_{S_1,\ldots,S_k} \max_{l=1,\ldots,k} \left\{ \frac{\text{cut}(S)}{\text{vol}(S)} \right\},$$

[5] We will see in Chapter 5 that the map $\varphi : V \to \mathbb{R}^{k-1}$ defined in Algorithm 4.2 can also be used for data visualization, not just clustering.

where the maximum is over all choice of k disjoin subsets of V (but not necessarily forming a partition).

Another natural definition is

$$\varphi_G(k) = \min_{S:\mathrm{vol}\, S \leq \frac{1}{k}\,\mathrm{vol}(G)} \frac{\mathrm{cut}(S)}{\mathrm{vol}(S)}.$$

It is easy to see that

$$\varphi_G(k) \leq \rho_G(k).$$

The following are analogues of Cheeger's inequality for multiple clusters.

Theorem 4.10 ([85]). *Let $G = (V, E, W)$ be a graph and k a positive integer*

$$\rho_G(k) \leq \mathcal{O}\left(k^2\right) \sqrt{\lambda_k}. \tag{4.9}$$

Also,

$$\rho_G(k) \leq \mathcal{O}\left(\sqrt{\lambda_{2k} \log k}\right).$$

5

Nonlinear Dimension Reduction and Diffusion Maps

In Chapter 3 we discussed dimension reduction via Principal Component Analysis. Many datasets however have low dimensional structure that is not linear. In this chapter we will discuss nonlinear dimension reduction techniques. Just as with Spectral Clustering in Chapter 4 we will focus on graph data while noting that most types of data can be transform in a weighted graph by means of a similarity kernel (Section 5.1.1). The goal of this chapter is to embed the nodes of a graph in Euclidean space in a way that best preserves the intrinsic geometry of the graph (or the data that gave rise to the graph).

5.1 Diffusion Maps

Diffusion Maps will allows us to represent (weighted) graphs $G = (V, E, W)$ in $\mathbb{R}^d$, i.e. associating, to each node, a point in $\mathbb{R}^d$. Before presenting Diffusion Maps, we'll introduce a few important notions. The reader may notice the similarities with the objects described in the context of PageRank in Chapter 4, the main difference is that here the connections between graphs have no direction, meaning that the weight matrix W is symmetric; this will be crucial in the derivations below.

Given $G = (V, E, W)$ we consider a random walk (with independent steps) on the vertices of V with transition probabilities:

$$\mathbb{P}\left\{X(t+1) = j | X(t) = i\right\} = \frac{w_{ij}}{\deg(i)},$$

where $\deg(i) = \sum_j w_{ij}$. Let M be the matrix of these probabilities,

$$M_{ij} = \frac{w_{ij}}{\deg(i)}. \tag{5.1}$$

It is easy to see that $M_{ij} \geq 0$ and $M\mathbf{1} = \mathbf{1}$ (indeed, M is a transition probability matrix). Recalling that D is the diagonal matrix with diagonal entries

$D_{ii} = \deg(i)$ we have

$$M = D^{-1}W.$$

If we start a random walker at node i ($X(0) = 1$) then the probability that, at step t, is at node j is given by

$$\mathbb{P}\left\{X(t) = j | X(0) = i\right\} = \left(M^t\right)_{ij}.$$

In other words, the probability cloud of the random walker at point t, given that it started at node i is given by the row vector

$$\mathbb{P}\left\{X(t) | X(0) = i\right\} = e_i^T M^t = M^t[i,:].$$

Remark 5.1. A natural representation of the graph would be to associate each vertex to the probability cloud above, meaning

$$i \to M^t[i,:].$$

This would place nodes i_1 and i_2 for which the random walkers starting at i_1 and i_2 have, after t steps, very similar distribution of locations. However, this would require $d = n$. In what follows we will construct a similar mapping but for considerably smaller d.

M is not symmetric, but a matrix similar to M, $S = D^{\frac{1}{2}} M D^{-\frac{1}{2}}$ is, indeed $S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. We consider the spectral decomposition of S

$$S = V \Lambda V^T, \tag{5.2}$$

where $V = [v_1, \ldots, v_n]$ satisfies $V^T V = I_{n\times n}$ and Λ is diagonal with diagonal elements $\Lambda_{kk} = \lambda_k$ (and we organize them as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$). Note that $Sv_k = \lambda_k v_k$. Also,

$$M = D^{-\frac{1}{2}} S D^{\frac{1}{2}} = D^{-\frac{1}{2}} V \Lambda V^T D^{\frac{1}{2}} = \left(D^{-\frac{1}{2}} V\right) \Lambda \left(D^{\frac{1}{2}} V\right)^T.$$

We define $\Phi = D^{-\frac{1}{2}} V$ with columns $\Phi = [\varphi_1, \ldots, \varphi_n]$ and $\Psi = D^{\frac{1}{2}} V$ with columns $\Psi = [\psi_1, \ldots, \psi_n]$. Then

$$M = \Phi \Lambda \Psi^T,$$

and Φ, Ψ form a biorthogonal system in the sense that $\Phi^T \Psi = I_{n\times n}$ or, equivalently, $\varphi_j^T \psi_k = \delta_{jk}$. Note that φ_k and ψ_k are, respectively right and left eigenvectors of M, indeed, for all $1 \leq k \leq n$:

$$M\varphi_k = \lambda_k \varphi_k \quad \text{ and } \psi_k^T M = \lambda_k \psi_k^T.$$

Also, we can rewrite this decomposition as

$$M = \sum_{k=1}^{n} \lambda_k \varphi_k \psi_k^T.$$

and it is easy to see that

$$M^t = \sum_{k=1}^{n} \lambda_k^t \varphi_k \psi_k^T. \tag{5.3}$$

Let's revisit the embedding suggested on Remark 5.1. It would correspond to

$$v_i \to M^t[i,:] = \sum_{k=1}^{n} \lambda_k^t \varphi_k(i) \psi_k^T,$$

it is written in terms of the basis ψ_k. The Diffusion Map will essentially consist of the representing a node i by the coefficients of the above map

$$v_i \to M^t[i,:] = \begin{bmatrix} \lambda_1^t \varphi_1(i) \\ \lambda_2^t \varphi_2(i) \\ \vdots \\ \lambda_n^t \varphi_n(i) \end{bmatrix}, \tag{5.4}$$

Note that $M\mathbf{1} = \mathbf{1}$, meaning that one of the right eigenvectors φ_k is simply a multiple of $\mathbf{1}$ and so it does not distinguish the different nodes of the graph. We will show that this indeed corresponds to the the first eigenvalue.

Proposition 5.2. *All eigenvalues λ_k of M satisfy $|\lambda_k| \leq 1$.*

Proof.

Let φ_k be a right eigenvector associated with λ_k whose largest entry in magnitude is positive $\varphi_k(i_{\max})$. Then,

$$\lambda_k \varphi_k(i_{\max}) = M\varphi_k(i_{\max}) = \sum_{j=1}^{n} M_{i_{\max},j} \varphi_k(j).$$

This means, by triangular inequality that, that

$$|\lambda_k| = \sum_{j=1}^{n} |M_{i_{\max},j}| \frac{|\varphi_k(j)|}{|\varphi_k(i_{\max})|} \leq \sum_{j=1}^{n} |M_{i_{\max},j}| = 1.$$

□

Remark 5.3. It is possible that there are other eigenvalues with magnitude 1 but only if G is disconnected or if G is bipartite. Provided that G is disconnected, a natural way to remove potential periodicity issues (like the graph being bipartite) is to make the walk lazy, i.e. to add a certain probability of the walker to stay in the current node. This can be conveniently achieved by taking, e.g.,

$$M' = \frac{1}{2}M + \frac{1}{2}I.$$

By the proposition above we can take $\varphi_1 = \mathbf{1}$, meaning that the first coordinate of (5.4) does not help differentiate points on the graph. This suggests removing that coordinate:

Definition 5.4 (Diffusion Map). *Given a graph $G = (V, E, W)$ construct M and its decomposition $M = \Phi\Lambda\Psi^T$ as described above. The Diffusion Map is a map $\varphi_t : V \to \mathbb{R}^{n-1}$ given by*

$$\varphi_t(v_i) = \begin{bmatrix} \lambda_2^t \varphi_2(i) \\ \lambda_3^t \varphi_3(i) \\ \vdots \\ \lambda_n^t \varphi_n(i) \end{bmatrix}.$$

This map is still a map to $n-1$ dimensions. But note now that each coordinate has a factor of λ_k^t which, if λ_k is small will be rather small for moderate values of t. This motivates truncating the Diffusion Map by taking only the first d coefficients.

Definition 5.5 (Truncated Diffusion Map). *Given a graph $G = (V, E, W)$ and dimension d, construct M and its decomposition $M = \Phi\Lambda\Psi^T$ as described above. The Diffusion Map truncated to d dimensions is a map $\varphi_t : V \to \mathbb{R}^d$ given by*

$$\varphi_t^{(d)}(v_i) = \begin{bmatrix} \lambda_2^t \varphi_2(i) \\ \lambda_3^t \varphi_3(i) \\ \vdots \\ \lambda_{d+1}^t \varphi_{d+1}(i) \end{bmatrix}.$$

In the following theorem we show that the euclidean distance in the diffusion map coordinates (called diffusion distance) meaningfully measures distance between the probability clouds after t iterations.

Theorem 5.6. *For any pair of nodes v_{i_1}, v_{i_2} we have*

$$\|\varphi_t(v_{i_1}) - \varphi_t(v_{i_2})\|^2 = \sum_{j=1}^{n} \frac{1}{\deg(j)} \left[\mathbb{P}\{X(t) = j | X(0) = i_1\} - \mathbb{P}\{X(t) = j | X(0) = i_2\}\right]^2.$$

Proof.

Note that $\sum_{j=1}^{n} \frac{1}{\deg(j)} \left[\mathbb{P}\{X(t) = j | X(0) = i_1\} - \mathbb{P}\{X(t) = j | X(0) = i_2\}\right]^2$ can be rewritten as

$$\sum_{j=1}^{n} \frac{1}{\deg(j)} \left[\sum_{k=1}^{n} \lambda_k^t \varphi_k(i_1)\psi_k(j) - \sum_{k=1}^{n} \lambda_k^t \varphi_k(i_2)\psi_k(j)\right]^2 = \sum_{j=1}^{n} \frac{1}{\deg(j)} \left[\sum_{k=1}^{n} \lambda_k^t \left(\varphi_k(i_1) - \varphi_k(i_2)\right)\psi_k(j)\right]^2$$

and

$$\sum_{j=1}^{n}\frac{1}{\deg(j)}\left[\sum_{k=1}^{n}\lambda_k^t\left(\varphi_k(i_1)-\varphi_k(i_2)\right)\psi_k(j)\right]^2=\sum_{j=1}^{n}\left[\sum_{k=1}^{n}\lambda_k^t\left(\varphi_k(i_1)-\varphi_k(i_2)\right)\frac{\psi_k(j)}{\sqrt{\deg(j)}}\right]^2$$
$$=\left\|\sum_{k=1}^{n}\lambda_k^t\left(\varphi_k(i_1)-\varphi_k(i_2)\right)D^{-\frac{1}{2}}\psi_k\right\|^2.$$

Note that $D^{-\frac{1}{2}}\psi_k=v_k$ which forms an orthonormal basis, meaning that

$$\left\|\sum_{k=1}^{n}\lambda_k^t\left(\varphi_k(i_1)-\varphi_k(i_2)\right)D^{-\frac{1}{2}}\psi_k\right\|^2=\sum_{k=1}^{n}\left(\lambda_k^t\left(\varphi_k(i_1)-\varphi_k(i_2)\right)\right)^2$$
$$=\sum_{k=2}^{n}\left(\lambda_k^t\varphi_k(i_1)-\lambda_k^t\varphi_k(i_2)\right)^2,$$

where the last inequality follows from the fact that $\varphi_1=\mathbf{1}$ and concludes the proof of the theorem.

□

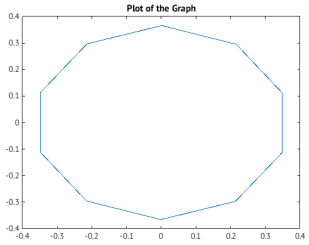


Fig. 5.1: The Diffusion Map of the ring graph gives a very natural way of displaying (indeed, if one is asked to draw the ring graph, this is probably the drawing that most people would do). It is actually not difficult to analytically compute the Diffusion Map of this graph and confirm that it displays the points in a circle.

5.1.1 Diffusion Maps of point clouds

Very often we are interested in embedding in $\mathbb{R}^d$ a point cloud of points $x_1,\ldots,x_n\in\mathbb{R}^p$ and not necessarily a graph. One optionis to use Principal Component Analysis (PCA), but PCA is only designed to find linear structure of the data and the low dimensionality of the dataset may be non-linear. For example, let us say that our dataset is images of the face of someone taken

from different angles and lighting conditions, for example, the dimensionality of this dataset is limited by the amount of muscles in the head and neck and by the degrees of freedom of the lighting conditions (see Figure **??**) but it is not clear that this low dimensional structure is linearly apparent on the pixel values of the images.

Let us consider a point cloud that is sampled from a two dimensional swiss roll embedded in three dimension (see Figure 5.2). In order to learn the two dimensional structure of this object we need to differentiate points that are near each other because they are close by in the manifold and not simply because the manifold is curved and the points appear nearby even when they really are distant in the manifold (see Figure 5.2 for an example). We will achieve this by creating a graph from the data points.

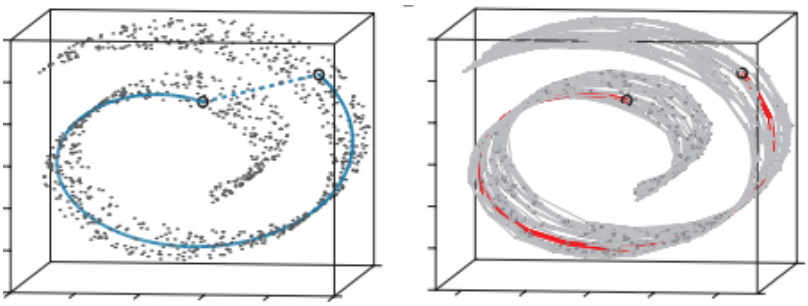


Fig. 5.2: A swiss roll point cloud (see, for example, [124]). The points are sampled from a two dimensional manifold curved in $\mathbb{R}^3$ and then a graph is constructed where nodes correspond to points.

Our goal is for the graph to capture the structure of the manifold. To each data point we will associate a node. For this we should only connect points that are close in the manifold and not points that maybe appear close in Euclidean space simply because of the curvature of the manifold. This is achieved by picking a small scale and linking nodes if they correspond to points whose distance is smaller than that scale. This is usually done smoothly via a kernel K_ε, and to each edge (i, j) associating a weight

$$w_{ij} = K_\varepsilon \left(\|x_i - x_j\|_2 \right),$$

a common example of a Kernel is $K_\varepsilon(u) = \exp\left(-\frac{1}{2\varepsilon}u^2\right)$, that gives essentially zero weight to edges corresponding to pairs of nodes for which $\|x_i - x_j\|_2 \gg \sqrt{\varepsilon}$. We can then take the the Diffusion Maps of the resulting graph.

5.1.2 An illustrative simple example

A simple and illustrative example is to take images of a blob on a background in different positions (image a white square on a black background and each data point corresponds to the same white square in different positions). This dataset is clearly intrinsically two dimensional, as each image can be described by the (two-dimensional) position of the square. However, we don't expect this two-dimensional structure to be directly apparent from the vectors of pixel values of each image; in particular we don't expect these vectors to lie in a two dimensional affine subspace!

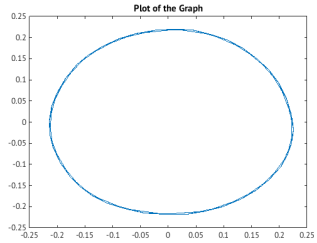


Fig. 5.3: The two-dimensional diffusion map of the dataset of the datase where each data point is an image with the same vertical strip in different positions in the x-axis, the circular structure is apparent.

Let's start by experimenting with the above example for one dimension. In that case the blob is a vertical stripe and simply moves left and right. We think of our space as the in many arcade games, if the square or stripe moves to the right all the way to the end of the screen, it shows up on the left side (and same for up-down in the two-dimensional case). Not only this point cloud should have a one dimensional structure but it should also exhibit a circular structure. Remarkably, this structure is completely apparent when taking the two-dimensional Diffusion Map of this dataset, see Figure 5.3.

For the two dimensional example, we expect the structure of the underlying manifold to be a two-dimensional torus. Indeed, Figure 5.4 shows that the three-dimensional diffusion map captures the toroidal structure of the data.

5.1.3 Similar non-linear dimensional reduction techniques

There are several other similar non-linear dimensional reduction methods. A particularly popular one is ISOMAP [124]. The idea is to find an embedding in $\mathbb{R}_d$ for which euclidean distances in the embedding correspond as much as possible to geodesic distances in the graph. This can be achieved by, between

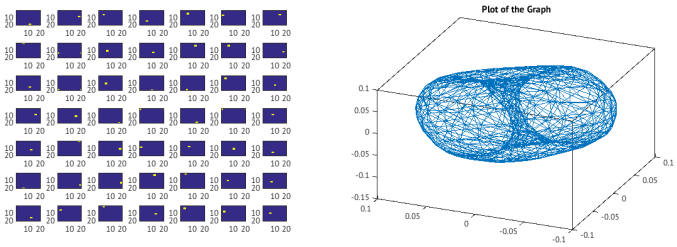


Fig. 5.4: On the left the data set considered and on the right its three dimensional diffusion map, the fact that the manifold is a torus is remarkably captured by the embedding.

pairs of nodes v_i, v_j finding their geodesic distance and then using, for example, Multidimensional Scaling to find points $y_i \in \mathbb{R}^d$ that minimize (for example)

$$\min_{y_1,\ldots,y_n \in \mathbb{R}^d} \sum_{i,j} \left(\|y_i - y_j\|^2 - \delta_{ij}^2 \right)^2,$$

which can be done with spectral methods (it is a good exercise to compute the optimal solution to the above optimization problem).

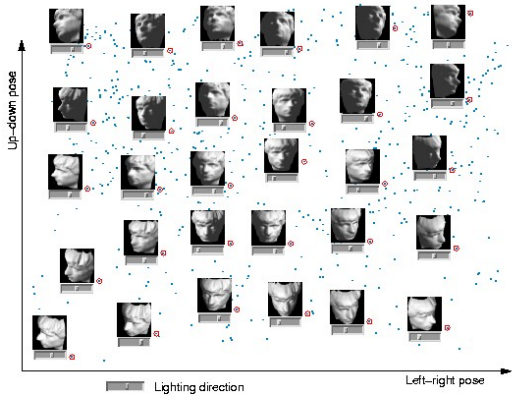


Fig. 5.5: The two dimensional represention of a data set of images of faces as obtained in [124] using ISOMAP. Remarkably, the two dimensionals are interpretable

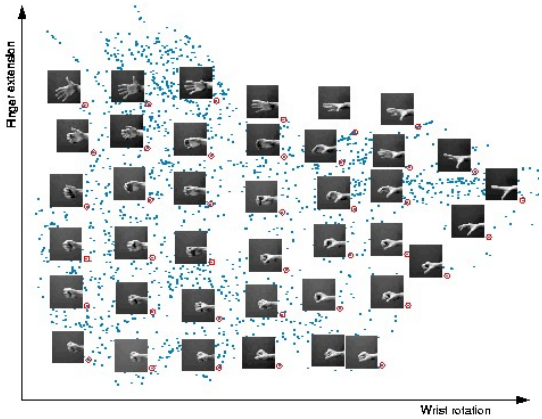


Fig. 5.6: The two dimensional represention of a data set of images of human hand as obtained in [124] using ISOMAP. Remarkably, the two dimensionals are interpretable

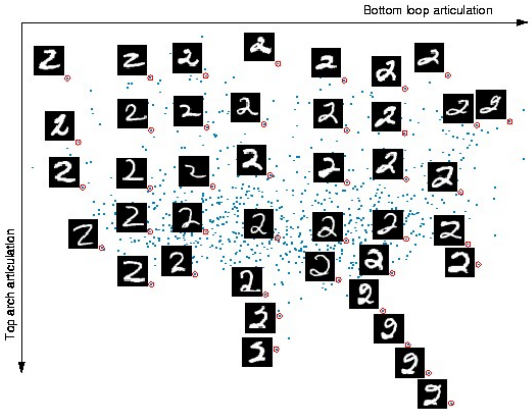


Fig. 5.7: The two dimensional represention of a data set of handwritten digits as obtained in [124] using ISOMAP. Remarkably, the two dimensionals are interpretable

5.2 Connections between Diffusion Maps and Spectral Clustering

Diffusion maps are tightly connected to Spectral Clustering (described in Chapter 4). In fact, Spectral Clustering can be understood as simply performing k-means on the embedding given by Diffusion Maps truncated to $k-1$ dimensions.

A natural way to try to overcome the issues of k-means depicted in Figure 4.4 is by using Diffusion Maps: Given the data points we construct a weighted graph $G = (V, E, W)$ using a kernel K_ε, such as $K_\varepsilon(u) = \exp\left(\frac{1}{2\varepsilon}u^2\right)$, by associating each point to a vertex and, for which pair of nodes, set the edge weight as

$$w_{ij} = K_\varepsilon\left(\|x_i - x_j\|\right).$$

Recall the construction of a matrix $M = D^{-1}W$ as the transition matrix of a random walk

$$\mathbb{P}\left\{X(t+1) = j | X(t) = i\right\} = \frac{w_{ij}}{\deg(i)} = M_{ij},$$

where D is the diagonal with $D_{ii} = \deg(i)$. The d-dimensional Diffusion Maps is given by

$$\varphi_t^{(d)}(i) = \begin{bmatrix} \lambda_2^t \varphi_2(i) \\ \vdots \\ \lambda_{d+1}^t \varphi_{d+1}(i) \end{bmatrix},$$

where $M = \Phi \Lambda \Psi^T$ where Λ is the diagonal matrix with the eigenvalues of M and Φ and Ψ are, respectively, the right and left eigenvectors of M (note that they form a bi-orthogonal system, $\Phi^T\Psi = I$).

If we want to cluster the vertices of the graph in k clusters, then it is natural to truncate the Diffusion Map to have $k-1$ dimensions (since in $k-1$ dimensions we can have k linearly separable sets). If indeed the clusters were linearly separable after embedding then one could attempt to use k-means on the embedding to find the clusters, this is precisely the motivation for Spectral Clustering.

Algorithm 5.1 Spectral Clustering described using Diffusion Maps.

Spectral Clustering: Given a graph $G = (V, E, W)$ and a number of clusters k (and t), Spectral Clustering consists in taking a $(k-1)$ dimensional Diffusion Map

$$\varphi_t^{(k-1)}(i) = \begin{bmatrix} \lambda_2^t \varphi_2(i) \\ \vdots \\ \lambda_k^t \varphi_k(i) \end{bmatrix}$$

and clustering the points $\varphi_t^{(k-1)}(1), \varphi_t^{(k-1)}(2), \ldots, \varphi_t^{(k-1)}(n) \in \mathbb{R}^{k-1}$ using, for example, k-means clustering. Usually, the scaling of λ_m^t is ignored (corresponding to $t = 0$).

In order to show that this indeed coincides with Algorithm 4.2, it is enough to show that $\varphi_m = D^{-\frac{1}{2}}v_m$ where v_m is the eigenvector associated with the m-th smallest eigenvalue of $\mathcal{L}_G$. This follows from the fact that $S = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ as defined in (5.2) is related to $\mathcal{L}_G$ by

$$\mathcal{L}_G = I - S,$$

and $\Phi = D^{-1/2}V$.

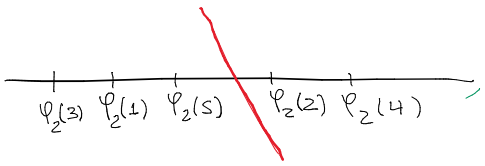


Fig. 5.8: For two clusters, spectral clustering consists in assigning to each vertex i a real number $\varphi_2(i)$, then setting a threshold τ and taking $S = \{i \in V : \varphi_2(i) \leq \tau\}$. This real number can both be interpreted through the spectrum of $\mathcal{L}_G$ as in Algorithm 4.1 or as the Diffusion Maps embedding as in Algorithm 5.1.

Proposition 5.7 below establishes a connection between Ncut (as described in Chapter 4) and the random walks introduced above. Let M as defined in (5.1) denote the matrix of transition probabilities. Recall that $M\mathbf{1} = \mathbf{1}$, corresponding to $M\varphi_1 = \varphi_1$, which means that $\psi_1^T M = \psi_1^T$, where

$$\psi_1 = D^{\frac{1}{2}} v_1 = D\varphi_1 = [\deg(i)]_{1 \leq i \leq n}.$$

This means that $\left[\frac{\deg(i)}{\mathrm{vol}(G)}\right]_{1 \leq i \leq n}$ is the stationary distribution of this random walk. Indeed it is easy to check that, if $X(t)$ has a certain distribution p_t then $X(t+1)$ has a distribution p_{t+1} given by $p_{t+1}^T = p_t^T M$

Proposition 5.7. *Given a graph $G = (V, E, W)$ and a partition (S, S^c) of V, $\mathrm{Ncut}(S)$ corresponds to the probability, in the random walk associated with G, that a random walker in the stationary distribution goes to S^c conditioned on being in S plus the probability of going to S condition on being in S^c, more explicitly:*

$$\mathrm{Ncut}(S) = \mathbb{P}\left\{X(t+1) \in S^c | X(t) \in S\right\} + \mathbb{P}\left\{X(t+1) \in S | X(t) \in S^c\right\},$$

where $\mathbb{P}\{X(t) = i\} = \frac{\deg(i)}{\mathrm{vol}(G)}$.

Proof. Without loss of generality we can take $t = 0$. Also, the second term in the sum corresponds to the first with S replaced by S^c and vice-versa, so we'll focus on the first one. We have:

$$\begin{aligned}
\mathbb{P}\{X(1) \in S^c | X(0) \in S\} &= \frac{\mathbb{P}\{X(1) \in S^c \cap X(0) \in S\}}{\mathbb{P}\{X(0) \in S\}} \\
&= \frac{\sum_{i \in S} \sum_{j \in S^c} \mathbb{P}\{X(1) \in j \cap X(0) \in i\}}{\sum_{i \in S} \mathbb{P}\{X(0) = i\}} \\
&= \frac{\sum_{i \in S} \sum_{j \in S^c} \frac{\deg(i)}{\mathrm{vol}(G)} \frac{w_{ij}}{\deg(i)}}{\sum_{i \in S} \frac{\deg(i)}{\mathrm{vol}(G)}} \\
&= \frac{\sum_{i \in S} \sum_{j \in S^c} w_{ij}}{\sum_{i \in S} \deg(i)} \\
&= \frac{\mathrm{cut}(S)}{\mathrm{vol}(S)}.
\end{aligned}$$

Analogously,

$$\mathbb{P}\{X(t+1) \in S | X(t) \in S^c\} = \frac{\mathrm{cut}(S)}{\mathrm{vol}(S^c)},$$

which concludes the proof. □

5.3 Semi-supervised learning

Classification is a central task in machine learning. In a supervised learning setting we are given many labelled examples and want to use them to infer the label of a new, unlabeled example. For simplicity, let us focus on the case of two labels, $\{-1, +1\}$.

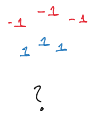


Fig. 5.9: Given a few labeled points, the task is to label an unlabeled point.

Let us consider the task of labelling the point "?" in Figure 5.9 given the labeled points. The natural label to give to the unlabeled point would be 1.

However, if we are given not just one unlabeled point, but many, as in Figure 5.10; then it starts being apparent that -1 is a more reasonable guess.

Intuitively, the unlabeled data points allowed us to better learn the intrinsic geometry of the dataset. That is the idea behind Semi-Supervised Learning (SSL), to make use of the fact that often one has access to many unlabeled data points in order to improve classification.

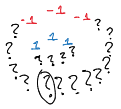


Fig. 5.10: In this example we are given many unlabeled points, the unlabeled points help us learn the geometry of the data.

Just as above, we will use the data points to construct (via a kernel K_ε) a graph $G = (V, E, W)$ where nodes correspond to points. More precisely, let l denote the number of labeled points with labels $f_1, \ldots, f_l$, and u the number of unlabeled points (with $n = l + u$), the first l nodes $v_1, \ldots, v_l$ correspond to labeled points and the rest $v_{l+1}, \ldots, v_n$ are unlabaled. We want to find a function $f : V \to \{-1, 1\}$ that agrees on labeled points: $f(i) = f_i$ for $i = 1, \ldots, l$ and that is "as smooth as possible" the graph. A way to pose this is the following

$$\min_{f:V\to\{-1,1\}:\, f(i)=f_i\ i=1,\ldots,l} \sum_{i<j} w_{ij} \left(f(i) - f(j)\right)^2.$$

Instead of restricting ourselves to giving $\{-1, 1\}$ we allow ourselves to give real valued labels, with the intuition that we can "round" later by, e.g., assigning the sign of $f(i)$ to node i.

We thus are interested in solving

$$\min_{f:V\to\mathbb{R}:\, f(i)=f_i\ i=1,\ldots,l} \sum_{i<j} w_{ij} \left(f(i) - f(j)\right)^2.$$

If we denote by f the vector (in $\mathbb{R}^n$ with the function values) then, recalling Proposition 4.3, we are can rewrite the problem as

$$\sum_{i<j} w_{ij} \left(f(i) - f(j)\right)^2 = f^T L_G f.$$

Remark 5.8. Consider an analogous example on the real line, where one would want to minimize

$$\int f'(x)^2 dx.$$

Integrating by parts

$$\int f'(x)^2 dx = \text{Boundary Terms} - \int f(x) f''(x) dx.$$

Analogously, in $\mathbb{R}^d$:

$$\int \|\nabla f(x)\|^2 dx = \int \sum_{k=1}^{d} \left(\frac{\partial f}{\partial x_k}(x)\right)^2 dx = \text{B. T.} - \int f(x) \sum_{k=1}^{d} \frac{\partial^2 f}{\partial x_k^2}(x) dx = \text{B. T.} - \int f(x) \Delta f(x) dx,$$

which helps motivate the use of the term graph Laplacian for L_G.

Let us consider our problem

$$\min_{f:V\to\mathbb{R}:\, f(i)=f_i\ i=1,\dots,l} f^T L_G f.$$

We can write

$$D=\begin{bmatrix} D_l & 0 \\ 0 & D_u \end{bmatrix},\quad W=\begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix},\quad L_G=\begin{bmatrix} D_l-W_{ll} & -W_{lu} \\ -W_{ul} & D_u-W_{uu} \end{bmatrix},\quad \text{and } f=\begin{bmatrix} f_l \\ f_u \end{bmatrix}.$$

Then we want to find (recall that $W_{ul}=W_{lu}^T$)

$$\min_{f_u\in\mathbb{R}^u} f_l^T\left[D_l-W_{ll}\right]f_l-2f_u^T W_{ul}f_l+f_u^T\left[D_u-W_{uu}\right]f_u.$$

by first-order optimality conditions, it is easy to see that the optimal satisfies

$$\left(D_u-W_{uu}\right)f_u=W_{ul}f_l.$$

If D_u-W_{uu} is invertible[1] then

$$f_u^*=\left(D_u-W_{uu}\right)^{-1}W_{ul}f_l.$$

Remark 5.9. The function f function constructed is called a harmonic extension. Indeed, it shares properties with harmonic functions in euclidean space such as the mean value property and maximum principles; if v_i is an unlabeled point then

$$f(i)=\left[D_u^{-1}\left(W_{ul}f_l+W_{uu}f_u\right)\right]_i=\frac{1}{\deg(i)}\sum_{j=1}^{n}w_{ij}f(j),$$

which immediately implies that the maximum and minimum value of f needs to be attained at a labeled point.

An interesting experience and the Sobolev Embedding Theorem

Let us try a simple experiment. Let's say we have a grid on $[-1,1]^d$ dimensions (with say m^d points for some large m) and we label the center as $+1$ and every node that is at distance larger or equal to 1 to the center, as -1. We are interested in understanding how the above algorithm will label the remaining points, hoping that it will assign small numbers to points far away from the center (and close to the boundary of the labeled points) and large numbers to points close to the center.

See the results for $d=1$ in Figure 5.11, $d=2$ in Figure 5.12, and $d=3$ in Figure 5.13. While for $d\leq 2$ it appears to be smoothly interpolating between

[1] It is not difficult to see that unless the problem is in some form degenerate, such as the unlabeled part of the graph being disconnected from the labeled one, then this matrix will indeed be invertible.

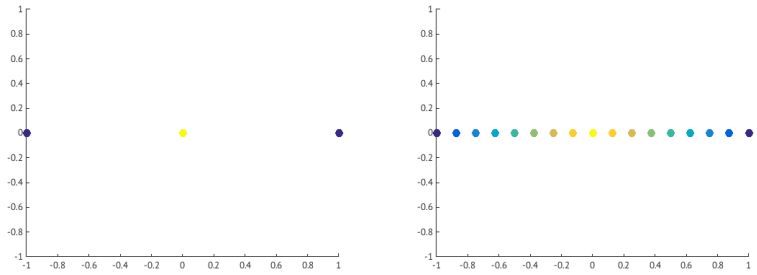


Fig. 5.11: The $d = 1$ example of the use of this method to the example described above, the value of the nodes is given by color coding. For $d = 1$ it appears to smoothly interpolate between the labeled points.

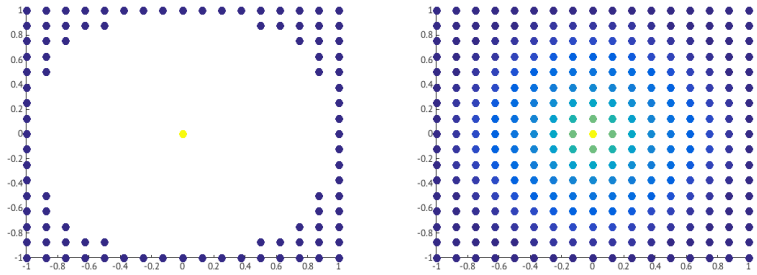


Fig. 5.12: The $d = 2$ example of the use of this method to the example described above, the value of the nodes is given by color coding. For $d = 2$ it appears to smoothly interpolate between the labeled points.

the labels, for $d = 3$ it seems that the method simply learns essentially -1 on all points, thus not being very meaningful. Let us turn to $\mathbb{R}^d$ for intuition:

Let's say that we want to find a function in $\mathbb{R}^d$ that takes the value 1 at zero and -1 at the unit sphere, that minimizes $\int_{B_0(1)} \|\nabla f(x)\|^2 dx$. Let us consider the following function on $B_0(1)$ (the ball centered at 0 with unit radius)

$$f_\varepsilon(x) = \begin{cases} 1 - 2\frac{|x|}{\varepsilon} & \text{if} |x| \leq \varepsilon \\ -1 & \text{otherwise.} \end{cases}$$

A quick calculation suggest that

$$\int_{B_0(1)} \|\nabla f_\varepsilon(x)\|^2 dx = \int_{B_0(\varepsilon)} \frac{1}{\varepsilon^2} dx = \text{vol}(B_0(\varepsilon)) \frac{1}{\varepsilon^2} dx \approx \varepsilon^{d-2},$$

meaning that, if $d > 2$, the performance of this function is improving as $\varepsilon \to 0$, explaining the results in Figure 5.13.

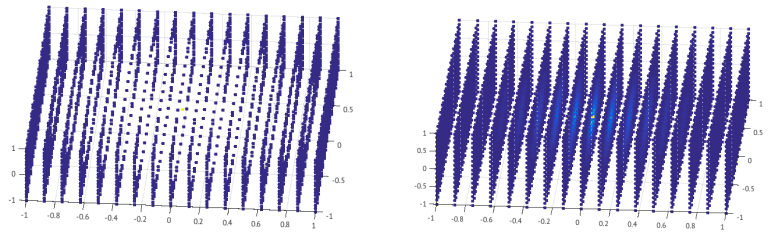


Fig. 5.13: The $d = 3$ example of the use of this method to the example described above, the value of the nodes is given by color coding. For $d = 3$ the solution appears to only learn the label -1.

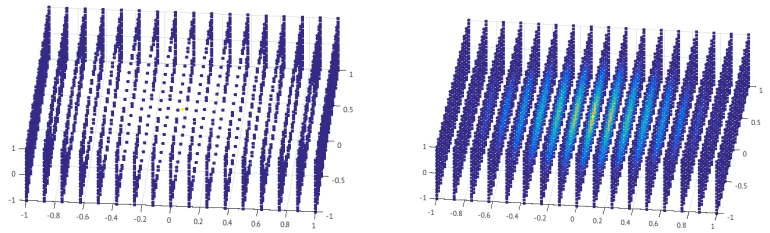


Fig. 5.14: The $d = 3$ example of the use of this method with the extra regularization $f^T L^2 f$ to the example described above, the value of the nodes is given by color coding. The extra regularization seems to fix the issue of discontinuities.

One way of thinking about what is going on is through the Sobolev Embedding Theorem. $H^m\left(\mathbb{R}^d\right)$ is the space of function whose derivatives up to order m are square-integrable in $\mathbb{R}^d$, Sobolev Embedding Theorem says that if $m > \frac{d}{2}$ then, if $f \in H^m\left(\mathbb{R}^d\right)$ then f must be continuous, which would rule out the behavior observed in Figure 5.13. It also suggests that if we are able to control also second derivates of f then this phenomenon should disappear (since $2 > \frac{3}{2}$). While we will not describe it here in detail, there is, in fact, a way of doing this by minimizing not $f^T L f$ but $f^T L^2 f$ instead, Figure 5.14 shows the outcome of the same experiment with the $f^T L f$ replaced by $f^T L^2 f$ and confirms our intuition that the discontinuity issue should disappear (see, e.g., [100] for more on this phenomenon).

6

Concentration of Measure and Matrix Inequalities

In this chapter we significantly expand on the concepts presented in Chapter 2, showcasing several other instances of the *Concentration of Measure* phenomena and focus on matrix versions of these inequalities that will be crucial in the forthcoming chapters.

6.1 Matrix Bernstein Inequality

In many of the chapters that follow we will need to control the largest eigenvalue or spectral norm of random matrices. Depending on the context, these matrices may represent the noise whose effect in a spectral algorithm is controlled by its spectral norm, or the size of a dual variable that needs to be controlled to show the exactness of a convex relaxation. While some of the tools we developed in Chapter 2 could be used to control the size of the entries of random matrices, which could translate to spectral bounds, this would likely introduce many suboptimal dimensional factors. We will start by presenting a general use concentration inequality for sums of independent random matrices, while noting that, as with scalars, many random variables can be written as sums of independent random variables even when it's not trivially apparent.

Let us recall Bernsteins inequality (Theorem 2.16) copied here with slightly different notation, and with only one of the tails: If $X_1, X_2, \ldots, X_n$ are independent centered random variables satisfying $|X_i| \leq r$ and $\mathbb{E}[X_i^2] = \frac{1}{n}\nu^2$. Then,

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i > t\right\} \leq \exp\left(-\frac{t^2}{2\nu^2 + \frac{2}{3}rt}\right). \tag{6.1}$$

A very useful generalization of this inequality exists for bounding the largest eigenvalue of the sum of independent random matrices

Theorem 6.1 (Theorem 1.4 in [130]). *Let $\{X_k\}_{k=1}^n$ be a sequence of independent random symmetric $d \times d$ matrices. Assume that each X_k satisfies:*

$$\mathbb{E}X_k = 0 \text{ and } \lambda_{\max}(X_k) \leq R \text{ almost surely.}$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_{k=1}^{n} X_k\right) \geq t\right\} \leq d \cdot \exp\left(\frac{-t^2}{2\sigma^2 + \frac{2}{3}Rt}\right) \text{ where } \sigma^2 = \left\|\sum_{k=1}^{n} \mathbb{E}\left(X_k^2\right)\right\|.$$

Note that $\|A\|$ denotes the spectral norm of A. Comparing with (6.1) the attentive reader will notice an extra dimensional factor of d; a simple change of variables shows that this corresponds to a poly-logarithmic factor on the random variable, a factor that will be discussed later in this Chapter.

In what follows we will state and prove various matrix concentration results, somewhat similar to Theorem 6.1. We will focus on understanding, and bounding, the typical value of the spectral norm of random matrices by upper bounding $\mathbb{E}\|X\|$, as these tend to be high dimensional objects themselves they often have enough concentration that tail bounds are then easy to obtain. In fact, in the next Section we will illustration exactly this by deriving a tail bound for the spectral norm of a Wigner matrix using Gaussian Concentration. For an approach to matrix concentration that includes a direct proof of Theorem 6.1 we recommend Tropp's excellent monograph [132].

6.2 Gaussian Concentration and the Spectral norm of Wigner Matrices

One of the most important results in concentration of measure is Gaussian concentration. Although being a concentration result specific for normally distributed random variables, it will be very useful throughout this book. Intuitively it says that if $F : \mathbb{R}^n \to \mathbb{R}$ is a function that is stable in terms of its input then $F(g)$ is very well concentrated around its mean, where $g \in \mathcal{N}(0, I)$. More precisely:

Theorem 6.2 (Gaussian Concentration). *Let* $X = [X_1, \ldots, X_n]^T$ *be a vector with i.i.d. standard Gaussian entries and* $F : \mathbb{R}^n \to \mathbb{R}$ *a* σ*-Lipschitz function (i.e.:* $|F(x) - F(y)| \leq \sigma\|x - y\|$*, for all* $x, y \in \mathbb{R}^n$*). Then, for every* $t \geq 0$

$$\mathbb{P}\left\{|F(X) - \mathbb{E}F(X)| \geq t\right\} \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

For the sake of simplicity we will show the proof for a slightly weaker bound: $\mathbb{P}\left\{|F(X) - \mathbb{E}F(X)| \geq t\right\} \leq 2\exp\left(-\frac{2}{\pi^2}\frac{t^2}{\sigma^2}\right)$. This exposition follows closely the proof of Theorem 2.1.12 in [123] and the original argument is due to Maurey and Pisier. For a proof with the optimal constants see, for example, Theorem 3.25 in [134]. We will also assume that the function F is smooth — this is actually not a restriction, as a limiting argument can generalize the result from smooth functions to general Lipschitz functions.

Proof.

If F is smooth, then it is easy to see that the Lipschitz property implies that, for every $x \in \mathbb{R}^n$, $\|\nabla F(x)\|_2 \leq \sigma$. By subtracting a constant from F, we can assume that $\mathbb{E}F(X) = 0$. Also, it is enough to show a one-sided bound

$$\mathbb{P}\left\{F(X) - \mathbb{E}F(X) \geq t\right\} \leq \exp\left(-\frac{2}{\pi^2}\frac{t^2}{\sigma^2}\right),$$

since obtaining the same bound for $-F(X)$ and taking a union bound would gives the result.

We start by using the same idea as in the proof of the large deviation inequalities above. For any $\lambda > 0$, Markov's inequality implies that

$$\begin{aligned}\mathbb{P}\left\{F(X) \geq t\right\} &= \mathbb{P}\left\{\exp\left(\lambda F(X)\right) \geq \exp\left(\lambda t\right)\right\} \\ &\leq \frac{\mathbb{E}\left[\exp\left(\lambda F(X)\right)\right]}{\exp\left(\lambda t\right)}\end{aligned}$$

This means we need to upper bound $\mathbb{E}\left[\exp\left(\lambda F(X)\right)\right]$ using a bound on $\|\nabla F\|$. The idea is to introduce a random independent copy Y of X. Since $\exp\left(\lambda \cdot\right)$ is convex, Jensen's inequality implies that

$$\mathbb{E}\left[\exp\left(-\lambda F(Y)\right)\right] \geq \exp\left(-\mathbb{E}\lambda F(Y)\right) = \exp(0) = 1.$$

Hence, since X and Y are independent,

$$\mathbb{E}\left[\exp\left(\lambda\left[F(X) - F(Y)\right]\right)\right] = \mathbb{E}\left[\exp\left(\lambda F(X)\right)\right]\mathbb{E}\left[\exp\left(-\lambda F(Y)\right)\right] \geq \mathbb{E}\left[\exp\left(\lambda F(X)\right)\right]$$

Now we use the Fundamental Theorem of Calculus in a circular arc from X to Y:

$$F(X) - F(Y) = \int_0^{\frac{\pi}{2}} \frac{\partial}{\partial\theta} F\left(Y\cos\theta + X\sin\theta\right) d\theta.$$

The advantage of using the circular arc is that, for any θ, $X_\theta := Y\cos\theta + X\sin\theta$ is another random variable with the same distribution. And this property holds for its derivative with respect to θ, $X'_\theta = -Y\sin\theta + X\cos\theta$ as well. Moreover, X_θ and X'_θ are independent. In fact, note that

$$\mathbb{E}\left[X_\theta {X'_\theta}^T\right] = \mathbb{E}\left[Y\cos\theta + X\sin\theta\right]\left[-Y\sin\theta + X\cos\theta\right]^T = 0.$$

We use Jensen's inequality again (with respect to the integral now) to get:

$$\begin{aligned}\exp\left(\lambda\left[F(X) - F(Y)\right]\right) &= \exp\left(\lambda\frac{\pi}{2}\frac{1}{\pi/2}\int_0^{\pi/2}\frac{\partial}{\partial\theta}F\left(X_\theta\right)d\theta\right) \\ &\leq \frac{1}{\pi/2}\int_0^{\pi/2}\exp\left(\lambda\frac{\pi}{2}\frac{\partial}{\partial\theta}F\left(X_\theta\right)\right)d\theta\end{aligned}$$

Using the chain rule,

$$\exp\left(\lambda\left[F(X)-F(Y)\right]\right) \leq \frac{2}{\pi}\int_0^{\pi/2} \exp\left(\lambda\frac{\pi}{2}\nabla F\left(X_\theta\right)\cdot X_\theta'\right)d\theta,$$

and taking expectations

$$\mathbb{E}\exp\left(\lambda\left[F(X)-F(Y)\right]\right) \leq \frac{2}{\pi}\int_0^{\pi/2} \mathbb{E}\exp\left(\lambda\frac{\pi}{2}\nabla F\left(X_\theta\right)\cdot X_\theta'\right)d\theta,$$

If we condition on X_θ, since $\left\|\lambda\frac{\pi}{2}\nabla F\left(X_\theta\right)\right\| \leq \lambda\frac{\pi}{2}\sigma$, $\lambda\frac{\pi}{2}\nabla F\left(X_\theta\right)\cdot X_\theta'$ is a gaussian random variable with variance at most $\left(\lambda\frac{\pi}{2}\sigma\right)^2$. This directly implies that, for every value of X_θ

$$\mathbb{E}_{X_\theta'}\exp\left(\lambda\frac{\pi}{2}\nabla F\left(X_\theta\right)\cdot X_\theta'\right) \leq \exp\left[\frac{1}{2}\left(\lambda\frac{\pi}{2}\sigma\right)^2\right]$$

Taking expectation now over X_θ, and putting everything together, gives

$$\mathbb{E}\left[\exp\left(\lambda F(X)\right)\right] \leq \exp\left[\frac{1}{2}\left(\lambda\frac{\pi}{2}\sigma\right)^2\right],$$

which means that

$$\mathbb{P}\left\{F(X)\geq t\right\} \leq \exp\left[\frac{1}{2}\left(\lambda\frac{\pi}{2}\sigma\right)^2-\lambda t\right],$$

Optimizing for λ gives $\lambda^* = \left(\frac{2}{\pi}\right)^2\frac{t}{\sigma^2}$, which in turn gives

$$\mathbb{P}\left\{F(X)\geq t\right\} \leq \exp\left[-\frac{2}{\pi^2}\frac{t^2}{\sigma^2}\right].$$

□

6.2.1 Spectral norm of a Wigner Matrix

We give an illustrative example of the utility of Gaussian concentration. Let $W\in\mathbb{R}^{n\times n}$ be a standard Gaussian Wigner matrix, a symmetric matrix with (otherwise) independent Gaussian entries, the off-diagonal entries have unit variance and the diagonal entries have variance 2. $\|W\|$ depends on $\frac{n(n+1)}{2}$ independent (standard) Gaussian random variables and it is easy to see that it is a $\sqrt{2}$-Lipschitz function of these variables, since

$$\left|\|W^{(1)}\|-\|W^{(2)}\|\right| \leq \left\|W^{(1)}-W^{(2)}\right\| \leq \left\|W^{(1)}-W^{(2)}\right\|_F.$$

The symmetry of the matrix and the variance 2 of the diagonal entries are responsible for an extra factor of $\sqrt{2}$.

Using Gaussian Concentration (Theorem 6.2) we immediately get

$$\mathbb{P}\left\{\|W\| \geq \mathbb{E}\|W\| + t\right\} \leq 2\exp\left(-\frac{t^2}{4}\right).$$

Since[1] $\mathbb{E}\|W\| \leq 2\sqrt{n}$ we get

Proposition 6.3. *Let $W \in \mathbb{R}^{n\times n}$ be a standard Gaussian Wigner matrix, a symmetric matrix with (otherwise) independent Gaussian entries, the off-diagonal entries have unit variance and the diagonal entries have variance 2. Then,*

$$\mathbb{P}\left\{\|W\| \geq 2\sqrt{n} + t\right\} \leq 2\exp\left(-\frac{t^2}{4}\right).$$

Note that this gives an extremely precise control of the fluctuations of $\|W\|$. In fact, for $t = 2\sqrt{\log n}$ this gives

$$\mathbb{P}\left\{\|W\| \geq 2\sqrt{n} + 2\sqrt{\log n}\right\} \leq 2\exp\left(-\frac{4\log n}{4}\right) = \frac{2}{n}.$$

6.2.2 Talagrand's concentration inequality

A remarkable result by Talagrand [121], Talangrad's concentration inequality, provides an analogue of Gaussian concentration for bounded random variables.

Theorem 6.4 (Talangrand concentration inequality, Theorem 2.1.13 [123]). *Let $K > 0$, and let $X_1, \ldots, X_n$ be independent bounded random variables with $|X_i| \leq K$ for all $1 \leq i \leq n$. Let $F : \mathbb{R}^n \to \mathbb{R}$ be a σ-Lipschitz and convex function. Then, for any $t \geq 0$,*

$$\mathbb{P}\left\{|F(X) - \mathbb{E}\left[F(X)\right]| \geq tK\right\} \leq c_1\exp\left(-c_2\frac{t^2}{\sigma^2}\right),$$

for positive constants c_1, and c_2.

Other useful similar inequalities (with explicit constants) are available in [91].

6.3 Non-Commutative Khintchine inequality

We start with a particularly important inequality involving the expected value of a random matrix. It is intimately related to the non-commutative Khintchine inequality [107], and for that reason we will often refer to it as Non-commutative Khintchine (see, for example, (4.9) in [130]).

[1] It is an excellent exercise to prove $\mathbb{E}\|W\| \leq 2\sqrt{n}$ using Slepian's inequality.

Theorem 6.5 (Non-commutative Khintchine (NCK)). *Let $A_1, \dots, A_n \in \mathbb{R}^{d\times d}$ be symmetric matrices and $g_1, \dots, g_n \sim \mathcal{N}(0,1)$ i.i.d., then:*

$$\mathbb{E}\left\|\sum_{k=1}^{n} g_k A_k\right\| \leq \left(2 + 2\log(2d)\right)^{\frac{1}{2}} \sigma,$$

where

$$\sigma^2 = \left\|\sum_{k=1}^{n} A_k^2\right\|. \tag{6.2}$$

Note that, akin to Proposition 6.3, we can also use Gaussian Concentration to get a tail bound on $\|\sum_{k=1}^{n} g_k A_k\|$. We consider the function

$$F : \mathbb{R}^n \to \left\|\sum_{k=1}^{n} g_k A_k\right\|.$$

We now estimate its Lipschitz constant; let $g, h \in \mathbb{R}^n$ then

$$\begin{aligned}
\left|\left\|\sum_{k=1}^{n} g_k A_k\right\| - \left\|\sum_{k=1}^{n} h_k A_k\right\|\right| &\leq \left\|\left(\sum_{k=1}^{n} g_k A_k\right) - \left(\sum_{k=1}^{n} h_k A_k\right)\right\| \\
&= \left\|\sum_{k=1}^{n} (g_k - h_k) A_k\right\| \\
&= \max_{v:\, \|v\|=1} v^T \left(\sum_{k=1}^{n} (g_k - h_k) A_k\right) v \\
&= \max_{v:\, \|v\|=1} \sum_{k=1}^{n} (g_k - h_k)\left(v^T A_k v\right) \\
&\leq \max_{v:\, \|v\|=1} \sqrt{\sum_{k=1}^{n} (g_k - h_k)^2} \sqrt{\sum_{k=1}^{n} \left(v^T A_k v\right)^2} \\
&= \sqrt{\max_{v:\, \|v\|=1} \sum_{k=1}^{n} \left(v^T A_k v\right)^2}\, \|g - h\|_2,
\end{aligned}$$

where in the first inequality we made use of the triangular inequality and in the last one of the Cauchy-Schwarz inequality.

This motivates us to define a new parameter, the weak variance σ_*.

Definition 6.6 (Weak Variance (see, for example, [132])). *Given $A_1, \dots, A_n \in \mathbb{R}^{d\times d}$ symmetric matrices. We define the weak variance parameter as*

$$\sigma_*^2 = \max_{v:\, \|v\|=1} \sum_{k=1}^{n} \left(v^T A_k v\right)^2.$$

This means that, using Gaussian concentration (and setting $t = u\sigma_*$), we have

$$\mathbb{P}\left\{\left\|\sum_{k=1}^n g_k A_k\right\| \geq \left(2+2\log(2d)\right)^{\frac{1}{2}}\sigma + u\sigma_*\right\} \leq \exp\left(-\frac{1}{2}u^2\right). \tag{6.3}$$

Thus, although the expected value of $\|\sum_{k=1}^n g_k A_k\|$ is controlled by the parameter σ, its fluctuations seem to be controlled by σ_*. We compare the two quantities in the following proposition.

Proposition 6.7. *Given $A_1, \ldots, A_n \in \mathbb{R}^{d\times d}$ symmetric matrices, recall that*

$$\sigma = \sqrt{\left\|\sum_{k=1}^n A_k^2\right\|} \text{ and } \sigma_* = \sqrt{\max_{v:\,\|v\|=1}\sum_{k=1}^n (v^T A_k v)^2}.$$

We have

$$\sigma_* \leq \sigma.$$

Proof. Using the Cauchy-Schwarz inequality,

$$\begin{aligned}
\sigma_*^2 &= \max_{v:\,\|v\|=1}\sum_{k=1}^n \left(v^T A_k v\right)^2 \\
&= \max_{v:\,\|v\|=1}\sum_{k=1}^n \left(v^T\left[A_k v\right]\right)^2 \\
&\leq \max_{v:\,\|v\|=1}\sum_{k=1}^n \left(\|v\|\|A_k v\|\right)^2 \\
&= \max_{v:\,\|v\|=1}\sum_{k=1}^n \|A_k v\|^2 \\
&= \max_{v:\,\|v\|=1}\sum_{k=1}^n v^T A_k^2 v \\
&= \left\|\sum_{k=1}^n A_k^2\right\| \\
&= \sigma^2.
\end{aligned}$$

□

6.3.1 Optimality of matrix concentration result for Gaussian series

The following simple calculation is suggestive that the parameter σ in Theorem 6.5 is indeed the correct parameter to understand $\mathbb{E}\|\sum_{k=1}^n g_k A_k\|$.

$$\mathbb{E}\left\|\sum_{k=1}^{n} g_k A_k\right\|^2 = \mathbb{E}\left\|\left(\sum_{k=1}^{n} g_k A_k\right)^2\right\| = \mathbb{E}\max_{v:\ \|v\|=1} v^T\left(\sum_{k=1}^{n} g_k A_k\right)^2 v$$

$$\geq \max_{v:\ \|v\|=1} \mathbb{E} v^T\left(\sum_{k=1}^{n} g_k A_k\right)^2 v = \max_{v:\ \|v\|=1} v^T\left(\sum_{k=1}^{n} A_k^2\right) v = \sigma^2.$$

But a natural question is whether the logarithmic term is needed. Motivated by this question we will explore a couple of examples.

Example 6.8. We can write a $d \times d$ Wigner matrix W as a gaussian series, by taking A_{ij} for $i \leq j$ defined as

$$A_{ij} = e_i e_j^T + e_j e_i^T,$$

if $i \neq j$, and

$$A_{ii} = \sqrt{2} e_i e_i^T.$$

It is not difficult to see that, in this case, $\sum_{i \leq j} A_{ij}^2 = (d+1) I_{d \times d}$, meaning that $\sigma = \sqrt{d+1}$. This implies that Theorem 6.5 gives us

$$\mathbb{E}\|W\| \lesssim \sqrt{d \log d},$$

however, we know that $\mathbb{E}\|W\| \asymp \sqrt{d}$, meaning that the bound given by NCK (Theorem 6.5) is, in this case, suboptimal by a logarithmic factor.[2]

The next example will show that the logarithmic factor is in fact needed in some examples

Example 6.9. Consider $A_k = e_k e_k^T \in \mathbb{R}^{d \times d}$ for $k = 1, \ldots, d$. The matrix $\sum_{k=1}^{n} g_k A_k$ corresponds to a diagonal matrix with independent standard gaussian random variables as diagonal entries, and so its spectral norm is given by $\max_k |g_k|$. It is known that $\max_{1 \leq k \leq d} |g_k| \asymp \sqrt{\log d}$. On the other hand, a direct calculation shows that $\sigma = 1$. This shows that the logarithmic factor cannot, in general, be removed.

This motivates the question of trying to understand when is it that the extra dimensional factor is needed. For both these examples, the resulting matrix $X = \sum_{k=1}^{n} g_k A_k$ has independent entries (except for the fact that it is symmetric). The case of independent entries [111, 116, 81, 24] is now somewhat understood:

Theorem 6.10 ([24]). *If X is a $d \times d$ random symmetric matrix with gaussian independent entries (except for the symmetry constraint) whose entry i, j has variance b_{ij}^2 then*

$$\mathbb{E}\|X\| \lesssim \sqrt{\max_{1 \leq i \leq d} \sum_{j=1}^{d} b_{ij}^2} + \max_{ij} |b_{ij}| \sqrt{\log d}.$$

[2] By $a \asymp b$ we mean $a \lesssim b$ and $a \gtrsim b$.

Remark 6.11. X in the theorem above can be written in terms of a Gaussian series by taking

$$A_{ij} = b_{ij}\left(e_i e_j^T + e_j e_i^T\right),$$

for $i < j$ and $A_{ii} = b_{ii} e_i e_i^T$. One can then compute σ and σ_*:

$$\sigma^2 = \max_{1\le i\le d} \sum_{j=1}^{d} b_{ij}^2 \text{ and } \sigma_*^2 \asymp b_{ij}^2.$$

This means that, when the random matrix in NCK (Theorem 6.5) has independent entries (modulo symmetry) then

$$\mathbb{E}\|X\| \lesssim \sigma + \sqrt{\log d}\sigma_*. \tag{6.4}$$

Theorem 6.10 together with a recent improvement of Theorem 6.5 by Tropp [133][3] motivate the bold possibility of (6.4) holding in more generality.

Conjecture 6.12. Let $A_1, \ldots, A_n \in \mathbb{R}^{d\times d}$ be symmetric matrices and $g_1, \ldots, g_n \sim \mathcal{N}(0,1)$ i.i.d., then:

$$\mathbb{E}\left\|\sum_{k=1}^{n} g_k A_k\right\| \lesssim \sigma + (\log d)^{\frac{1}{2}}\,\sigma_*,$$

While it may very will be that Conjecture 6.12 is false, no counter example is known, up to date.

6.4 Matrix concentration inequalities

In what follows, we closely follow [131] and present an elementary proof of a few useful matrix concentration inequalities. We start with a Master Theorem of sorts for Rademacher series (the Rademacher analogue of Theorem 6.5)

Theorem 6.13. *Let $H_1, \ldots, H_n \in \mathbb{R}^{d\times d}$ be symmetric matrices and $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. Rademacher random variables (meaning $= +1$ with probability $1/2$ and $= -1$ with probability $1/2$), then:*

$$\mathbb{E}\left\|\sum_{k=1}^{n} \varepsilon_k H_k\right\| \le \left(1 + 2\lceil \log(d)\rceil\right)^{\frac{1}{2}}\sigma,$$

where

$$\sigma^2 = \left\|\sum_{k=1}^{n} H_k^2\right\|. \tag{6.5}$$

[3] We briefly discuss this improvement in Remark 6.20

Using Theorem 6.13, we will prove the following theorem.

Theorem 6.14. *Let $T_1, \ldots, T_n \in \mathbb{R}^{d\times d}$ be random independent symmetric positive semidefinite matrices, then*

$$\mathbb{E}\left\|\sum_{i=1}^n T_i\right\| \leq \left[\left\|\sum_{i=1}^n \mathbb{E}T_i\right\|^{\frac{1}{2}} + \sqrt{C(d)}\left(\mathbb{E}\max_i \|T_i\|\right)^{\frac{1}{2}}\right]^2,$$

where

$$C(d) := 4 + 8\lceil \log d\rceil. \tag{6.6}$$

A key step in the proof of Theorem 6.14 is an idea that is extremely useful in Probability, the trick of symmetrization. For this reason we isolate it in a lemma.

Lemma 6.15 (Symmetrization). *Let $T_1, \ldots, T_n$ be independent random matrices (note that they do not necessarily need to be positive semidefinite, for the sake of this lemma) and $\varepsilon_1, \ldots, \varepsilon_n$ random i.i.d. Rademacher random variables (independent also from the matrices). Then*

$$\mathbb{E}\left\|\sum_{i=1}^n T_i\right\| \leq \left\|\sum_{i=1}^n \mathbb{E}T_i\right\| + 2\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i T_i\right\|$$

Proof. The triangular inequality gives

$$\mathbb{E}\left\|\sum_{i=1}^n T_i\right\| \leq \left\|\sum_{i=1}^n \mathbb{E}T_i\right\| + \mathbb{E}\left\|\sum_{i=1}^n (T_i - \mathbb{E}T_i)\right\|.$$

Let us now introduce, for each i, a random matrix T_i' identically distributed to T_i and independent (all $2n$ matrices are independent). Then

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^n (T_i - \mathbb{E}T_i)\right\| &= \mathbb{E}_T\left\|\sum_{i=1}^n \left(T_i - \mathbb{E}T_i - \mathbb{E}_{T_i'}\left[T_i' - \mathbb{E}_{T_i'}T_i'\right]\right)\right\| \\ &= \mathbb{E}_T\left\|\mathbb{E}_{T'}\sum_{i=1}^n (T_i - T_i')\right\| \leq \mathbb{E}\left\|\sum_{i=1}^n (T_i - T_i')\right\|,\end{aligned}$$

where we use the notation $\mathbb{E}_a$ to mean that the expectation is taken with respect to the variable a and the last step follows from Jensen's inequality with respect to $\mathbb{E}_{T'}$.

Since $T_i - T_i'$ is a symmetric random variable,[4] it is identically distributed to $\varepsilon_i (T_i - T_i')$, which gives

[4]Note we use the notation "symmetric random variable to mean $X \sim -X$ and symmetric matrix to mean $X^T = X$

$$\mathbb{E}\left\|\sum_{i=1}^{n}(T_i - T_i')\right\| = \mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i\,(T_i - T_i')\right\| \leq \mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i T_i\right\| + \mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i T_i'\right\| = 2\mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i T_i\right\|,$$

concluding the proof. □

Proof. [of Theorem 6.14]

Using Lemma 6.15 and Theorem 6.13 we get

$$\mathbb{E}\left\|\sum_{i=1}^{n}T_i\right\| \leq \left\|\sum_{i=1}^{n}\mathbb{E}T_i\right\| + \sqrt{C(d)}\mathbb{E}\left\|\sum_{i=1}^{n}T_i^2\right\|^{\frac{1}{2}}$$

The trick now is to make a term like the one in the LHS appear in the RHS. For that we start by noting (you can see Fact 2.3 in [131] for an elementary proof) that, since $T_i \succeq 0$,

$$\left\|\sum_{i=1}^{n}T_i^2\right\| \leq \max_i \|T_i\| \left\|\sum_{i=1}^{n}T_i\right\|.$$

This means that

$$\mathbb{E}\left\|\sum_{i=1}^{n}T_i\right\| \leq \left\|\sum_{i=1}^{n}\mathbb{E}T_i\right\| + \sqrt{C(d)}\mathbb{E}\left[\left(\max_i \|T_i\|\right)^{\frac{1}{2}}\left\|\sum_{i=1}^{n}T_i\right\|^{\frac{1}{2}}\right].$$

Furthermore, applying the Cauchy-Schwarz inequality for $\mathbb{E}$ gives,

$$\mathbb{E}\left\|\sum_{i=1}^{n}T_i\right\| \leq \left\|\sum_{i=1}^{n}\mathbb{E}T_i\right\| + \sqrt{C(d)}\left(\mathbb{E}\max_i \|T_i\|\right)^{\frac{1}{2}}\left(\mathbb{E}\left\|\sum_{i=1}^{n}T_i\right\|\right)^{\frac{1}{2}},$$

Now that the term $\mathbb{E}\left\|\sum_{i=1}^{n}T_i\right\|$ appears in the RHS, the proof can be finished with a simple application of the quadratic formula (see Section 6.1. in [131] for details).

□

We now show an inequality for general symmetric matrices

Theorem 6.16. *Let $Y_1, \ldots, Y_n \in \mathbb{R}^{d\times d}$ be random independent symmetric matrices satisfying $\mathbb{E}Y_i = 0$, then*

$$\mathbb{E}\left\|\sum_{i=1}^{n}Y_i\right\| \leq \sqrt{C(d)}\sigma + C(d)L,$$

where,

$$\sigma^2 = \left\|\sum_{i=1}^{n}\mathbb{E}Y_i^2\right\| \text{ and } L^2 = \mathbb{E}\max_i \|Y_i\|^2 \tag{6.7}$$

and, as in (6.6)*,*

$$C(d) := 4 + 8\lceil \log d \rceil.$$

Proof.

Using Symmetrization (Lemma 6.15) and Theorem 6.13, we get

$$\mathbb{E}\left\|\sum_{i=1}^{n} Y_i\right\| \leq 2\mathbb{E}_Y\left[\mathbb{E}_\varepsilon\left\|\sum_{i=1}^{n}\varepsilon_i Y_i\right\|\right] \leq \sqrt{C(d)}\mathbb{E}\left\|\sum_{i=1}^{n} Y_i^2\right\|^{\frac{1}{2}}.$$

Jensen's inequality gives

$$\mathbb{E}\left\|\sum_{i=1}^{n} Y_i^2\right\|^{\frac{1}{2}} \leq \left(\mathbb{E}\left\|\sum_{i=1}^{n} Y_i^2\right\|\right)^{\frac{1}{2}},$$

and the proof can be concluded by noting that $Y_i^2 \succeq 0$ and using Theorem 6.14. □

Remark 6.17 (The rectangular case). One can extend Theorem 6.16 to general rectangular matrices $S_1, \ldots, S_n \in \mathbb{R}^{d_1 \times d_2}$ by setting

$$Y_i = \begin{bmatrix} 0 & S_i \\ S_i^T & 0 \end{bmatrix},$$

and noting that

$$\left\|Y_i^2\right\| = \left\|\begin{bmatrix} 0 & S_i \\ S_i^T & 0 \end{bmatrix}^2\right\| = \left\|\begin{bmatrix} S_i S_i^T & 0 \\ 0 & S_i^T S_i \end{bmatrix}\right\| = \max\left\{\left\|S_i^T S_i\right\|, \left\|S_i S_i^T\right\|\right\}.$$

For details we refer to [131].

In order to prove Theorem 6.13, we will use an arithmetic mean-geometric mean (AM-GM) like inequality for matrices.

Lemma 6.18. *Given symmetric matrices $H, W, Y \in \mathbb{R}^{d\times d}$ and non-negative integers r, q satisfying $q \leq 2r$,*

$$\operatorname{Tr}\left[HW^q HY^{2r-q}\right] + \operatorname{Tr}\left[HW^{2r-q}HY^q\right] \leq \operatorname{Tr}\left[H^2\left(W^{2r} + Y^{2r}\right)\right],$$

and summing over q gives

$$\sum_{q=0}^{2r}\operatorname{Tr}\left[HW^q HY^{2r-q}\right] \leq \left(\frac{2r+1}{2}\right)\operatorname{Tr}\left[H^2\left(W^{2r} + Y^{2r}\right)\right].$$

We refer to Fact 2.4 in [131] for an elementary proof but note that it is a matrix analogue to the inequality,

$$\mu^\theta \lambda^{1-\theta} + \mu^{1-\theta}\lambda^\theta \leq \lambda + \mu$$

for $\mu, \lambda \geq 0$ and $0 \leq \theta \leq 1$, which can be easily shown by adding two AM-GM inequalities

$$\mu^\theta \lambda^{1-\theta} \leq \theta\mu + (1-\theta)\lambda \text{ and } \mu^{1-\theta}\lambda^\theta \leq (1-\theta)\mu + \theta\lambda.$$

Proof. [of Theorem 6.13]

Let $X = \sum_{k=1}^n \varepsilon_k H_k$, then for any positive integer p,

$$\mathbb{E}\|X\| \leq \left(\mathbb{E}\|X\|^{2p}\right)^{\frac{1}{2p}} = \left(\mathbb{E}\|X^{2p}\|\right)^{\frac{1}{2p}} \leq \left(\mathbb{E}\operatorname{Tr} X^{2p}\right)^{\frac{1}{2p}},$$

where the first inequality follows from Jensen's inequality and the last from $X^{2p} \succeq 0$ and the observation that the trace of a positive semidefinite matrix is at least its spectral norm. In the sequel, we upper bound $\mathbb{E}\operatorname{Tr} X^{2p}$. We introduce X_{+i} and X_{-i} as X conditioned on ε_i being, respectively $+1$ or -1. More precisely

$$X_{+i} = H_i + \sum_{j \neq i} \varepsilon_j H_j \text{ and } X_{-i} = -H_i + \sum_{j \neq i} \varepsilon_j H_j.$$

Then, we have

$$\mathbb{E}\operatorname{Tr} X^{2p} = \mathbb{E}\operatorname{Tr}\left[X X^{2p-1}\right] = \mathbb{E}\sum_{i=1}^n \operatorname{Tr} \varepsilon_i H_i X^{2p-1}.$$

Note that $\mathbb{E}_{\varepsilon_i} \operatorname{Tr}\left[\varepsilon_i H_i X^{2p-1}\right] = \frac{1}{2}\operatorname{Tr}\left[H_i\left(X_{+i}^{2p-1} - X_{-i}^{2p-1}\right)\right]$, this means that

$$\mathbb{E}\operatorname{Tr} X^{2p} = \sum_{i=1}^n \mathbb{E}\frac{1}{2}\operatorname{Tr}\left[H_i\left(X_{+i}^{2p-1} - X_{-i}^{2p-1}\right)\right],$$

where the expectation can be taken over ε_j for $j \neq i$.

Now we rewrite $X_{+i}^{2p-1} - X_{-i}^{2p-1}$ as a telescopic sum:

$$X_{+i}^{2p-1} - X_{-i}^{2p-1} = \sum_{q=0}^{2p-2} X_{+i}^q \left(X_{+i} - X_{-i}\right) X_{-i}^{2p-2-q},$$

which gives

$$\mathbb{E}\operatorname{Tr} X^{2p} = \sum_{i=1}^n \sum_{q=0}^{2p-2} \mathbb{E}\frac{1}{2}\operatorname{Tr}\left[H_i X_{+i}^q \left(X_{+i} - X_{-i}\right) X_{-i}^{2p-2-q}\right].$$

Since $X_{+i} - X_{-i} = 2H_i$ we get

$$\mathbb{E}\operatorname{Tr} X^{2p} = \sum_{i=1}^n \sum_{q=0}^{2p-2} \mathbb{E}\operatorname{Tr}\left[H_i X_{+i}^q H_i X_{-i}^{2p-2-q}\right]. \tag{6.8}$$

We now make use of Lemma 6.18 to get[5] to get

$$\mathbb{E}\,\mathrm{Tr}\,X^{2p} \leq \sum_{i=1}^{n} \frac{2p-1}{2}\mathbb{E}\,\mathrm{Tr}\left[H_i^2\left(X_{+i}^{2p-2}+X_{-i}^{2p-2}\right)\right]. \tag{6.9}$$

Hence,

$$\begin{aligned}
\sum_{i=1}^{n} \frac{2p-1}{2}\mathbb{E}\,\mathrm{Tr}\left[H_i^2\left(X_{+i}^{2p-2}+X_{-i}^{2p-2}\right)\right] &= (2p-1)\sum_{i=1}^{n}\mathbb{E}\,\mathrm{Tr}\left[H_i^2\frac{\left(X_{+i}^{2p-2}+X_{-i}^{2p-2}\right)}{2}\right] \\
&= (2p-1)\sum_{i=1}^{n}\mathbb{E}\,\mathrm{Tr}\left[H_i^2\mathbb{E}_{\varepsilon_i}\left[X^{2p-2}\right]\right] \\
&= (2p-1)\sum_{i=1}^{n}\mathbb{E}\,\mathrm{Tr}\left[H_i^2X^{2p-2}\right] \\
&= (2p-1)\mathbb{E}\,\mathrm{Tr}\left[\left(\sum_{i=1}^{n}H_i^2\right)X^{2p-2}\right]
\end{aligned}$$

Since $X^{2p-2} \succeq 0$ we have

$$\mathrm{Tr}\left[\left(\sum_{i=1}^{n}H_i^2\right)X^{2p-2}\right] \leq \left\|\sum_{i=1}^{n}H_i^2\right\|\mathrm{Tr}\,X^{2p-2} = \sigma^2\,\mathrm{Tr}\,X^{2p-2}, \tag{6.10}$$

which gives

$$\mathbb{E}\,\mathrm{Tr}\,X^{2p} \leq \sigma^2(2p-1)\mathbb{E}\,\mathrm{Tr}\,X^{2p-2}. \tag{6.11}$$

Applying this inequality, recursively, we get

$$\mathbb{E}\,\mathrm{Tr}\,X^{2p} \leq \left[(2p-1)(2p-3)\cdots(3)(1)\right]\sigma^{2p}\mathbb{E}\,\mathrm{Tr}\,X^0 = (2p-1)!!\sigma^{2p}d$$

Hence,

$$\mathbb{E}\|X\| \leq \left(\mathbb{E}\,\mathrm{Tr}\,X^{2p}\right)^{\frac{1}{2p}} \leq \left[(2p-1)!!\right]^{\frac{1}{2p}}\sigma d^{\frac{1}{2p}}.$$

Taking $p = \lceil \log d \rceil$ and using the fact that $(2p-1)!! \leq \left(\frac{2p+1}{e}\right)^p$ (see [131] for an elementary proof consisting essentially of taking logarithms and comparing the sum with an integral) we get

$$\mathbb{E}\|X\| \leq \left(\frac{2\lceil\log d\rceil+1}{e}\right)^{\frac{1}{2}}\sigma d^{\frac{1}{2\lceil\log d\rceil}} \leq (2\lceil\log d\rceil+1)^{\frac{1}{2}}\sigma.$$

.

□

[5] See Remark 6.20 regarding the suboptimality of this step.

Remark 6.19. A similar argument can be used to prove Theorem 6.5 (the Gaussian series case) based on Gaussian integration by parts, see Section 7.2. in [133].

Remark 6.20. Note that, up until the step from (6.8) to (6.9) all steps are equalities suggesting that this step may be the lossy step responsible by the suboptimal dimensional factor in several cases (although (6.10) can also potentially be lossy, it is not uncommon that $\sum H_i^2$ is a multiple of the identity matrix, which would render this step also an equality).

In fact, Joel Tropp [133] recently proved an improvement over the NCK inequality that, essentially, consists in replacing inequality (6.9) with a tighter argument. In a nutshell, the idea is that, if the H_i's are non-commutative, most summands in (6.8) are actually expected to be smaller than the ones corresponding to $q = 0$ and $q = 2p - 2$, which are the ones that appear in (6.9).

6.5 Other useful large deviation inequalities

This section contains several other useful scalar large deviation inequalities. We defer the proofs to references.

6.5.1 Additive Chernoff Bound

The additive Chernoff bound, also known as Chernoff-Hoeffding theorem concerns Bernoulli random variables.

Theorem 6.21. *Given $0 < p < 1$ and $X_1, \ldots, X_n$ i.i.d. random variables distributed as* Bernoulli(p) *random variable (meaning that it is 1 with probability p and 0 with probability $1 - p$), then, for any $\varepsilon > 0$:*

- $$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq p + \varepsilon\right\} \leq \left[\left(\frac{p}{p+\varepsilon}\right)^{p+\varepsilon}\left(\frac{1-p}{1-p-\varepsilon}\right)^{1-p-\varepsilon}\right]^n$$
- $$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \leq p - \varepsilon\right\} \leq \left[\left(\frac{p}{p-\varepsilon}\right)^{p-\varepsilon}\left(\frac{1-p}{1-p+\varepsilon}\right)^{1-p+\varepsilon}\right]^n$$

6.5.2 Multiplicative Chernoff Bound

There is also a multiplicative version (see, for example Lemma 2.3.3. in [53]), which is particularly useful.

Theorem 6.22. *Let $X_1, \ldots, X_n$ be independent random variables taking values is $\{0, 1\}$ (meaning they are Bernoulli distributed but not necessarily identically distributed). Let $\mu = \mathbb{E}\sum_{i=1}^{n} X_i$, then, for any $\delta > 0$:*

- $\mathbb{P}\{X > (1+\delta)\mu\} < \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{\mu}$
- $\mathbb{P}\{X < (1-\delta)\mu\} < \left[\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right]^{\mu}$

6.5.3 Deviation bounds for χ_2 variables

Another particularly useful deviation inequality is Lemma 1 in Laurent and Massart [82]:

Theorem 6.23 (Lemma 1 in Laurent and Massart [82]). *Let $X_1, \ldots, X_n$ be i.i.d. standard Gaussian random variables ($\mathcal{N}(0,1)$), and $a_1, \ldots, a_n$ non-negative numbers. Let*

$$Z = \sum_{k=1}^{n} a_k \left(X_k^2 - 1\right).$$

The following inequalities hold for any $t > 0$:

- $\mathbb{P}\{Z \geq 2\|a\|_2\sqrt{x} + 2\|a\|_\infty x\} \leq \exp(-x)$,
- $\mathbb{P}\{Z \leq -2\|a\|_2\sqrt{x}\} \leq \exp(-x)$,

where $\|a\|_2^2 = \sum_{k=1}^{n} a_k^2$ and $\|a\|_\infty = \max_{1\leq k\leq n} |a_k|$.

Note that if $a_k = 1$, for all k, then Z is a χ_2 random variable with n degrees of freedom, so this theorem immediately gives a deviation inequality for χ_2 random variables, see also the tail bound (2.18).

7

Max Cut, Lifting, and Approximation Algorithms

Many data analysis tasks include in them a step consisting of solving a computational problem, oftentimes in the form of finding a hidden parameter that best explains the data, or model specifications that provide best-fits. Many such problems, including examples in previous chapters, are computationally intractable. In complexity theory this is often captured by NP-hardness. Unless the widely believed $P \neq NP$ conjecture is false, there is no polynomial algorithm that can solve all instances of an NP-hard problem. Thus, when faced with an NP-hard problem (such as the *Max-Cut* problem discussed below) one has three natural options: to use an exponential type algorithm that solves exactly the problem in all instances, to design polynomial time algorithms that only work for some of the instances (hopefully relevant ones), or to design polynomial algorithms that, in all instances, produce guaranteed approximate solutions. This section is about the third option, another example of this approach is the earlier discussion on Spectral Clustering and Cheeger's inequality. The second option, of designing algorithms that work in many, rather than all, instances is discussed in later chapters, notably these goals are often achieved by the same algorithms.

The *Max-Cut* problem is defined as follows: Given a graph $G = (V, E)$ with non-negative weights w_{ij} on the edges, find a set $S \subset V$ for which $\text{cut}(S)$ is maximal. Goemans and Williamson [60] introduced an approximation algorithm that runs in polynomial time, has a randomized component in it, and is able to obtain a cut whose expected value is guaranteed to be no smaller than a particular constant α_{GW} times the optimum cut. The constant α_{GW} is referred to as the approximation ratio.

Let $V = \{1, \ldots, n\}$. One can restate `Max-Cut` as

$$\begin{array}{ll} \max & \frac{1}{2}\sum_{i<j} w_{ij}(1 - y_i y_j) \\ s.t. & |y_i| = 1 \end{array} \tag{7.1}$$

The y_i's are binary variables that indicate set membership, i.e., $y_i = 1$ if $i \in S$ and $y_i = -1$ otherwise.

Consider the following relaxation of this problem:

$$\begin{array}{ll} \max & \frac{1}{2}\sum_{i<j} w_{ij}(1-u_i^T u_j) \\ s.t. & u_i \in \mathbb{R}^n, \|u_i\| = 1. \end{array} \tag{7.2}$$

This is a relaxation because if we restrict u_i to be a multiple of e_1, the first element of the canonical basis, then (7.2) is equivalent to (7.1). For this to be a useful approach, the following two properties should hold:

(a) Problem (7.2) is easy to solve.
(b) The solution of (7.2) is, in some way, related to the solution of (7.1).

Definition 7.1. *Given a graph G, we define* $\mathrm{MaxCut}(G)$ *as the optimal value of* (7.1) *and* $\mathcal{R}\mathrm{MaxCut}(G)$ *as the optimal value of* (7.2).

We start with property (a). Set X to be the Gram matrix of $u_1, \ldots, u_n$, that is, $X = U^T U$ where the i'th column of U is u_i. We can rewrite the objective function of the relaxed problem as

$$\frac{1}{2}\sum_{i<j} w_{ij}(1 - X_{ij})$$

One can exploit the fact that X having a decomposition of the form $X = Y^T Y$ is equivalent to being positive semidefinite, denoted $X \succeq 0$. The set of PSD matrices is a convex set. Also, the constraint $\|u_i\| = 1$ can be expressed as $X_{ii} = 1$. This means that the relaxed problem is equivalent to the following semidefinite program (SDP):

$$\begin{array}{ll} \max & \frac{1}{2}\sum_{i<j} w_{ij}(1 - X_{ij}) \\ s.t. & X \succeq 0 \text{ and } X_{ii} = 1,\ i = 1, \ldots, n. \end{array} \tag{7.3}$$

This SDP can be solved (up to ε accuracy) in time polynomial on the input size and $\log(\varepsilon^{-1})$[135].

There is an alternative way of viewing (7.3) as a relaxation of (7.1). By taking $X = yy^T$ one can formulate a problem equivalent to (7.1)

$$\begin{array}{ll} \max & \frac{1}{2}\sum_{i<j} w_{ij}(1 - X_{ij}) \\ s.t. & X \succeq 0\ ,\ X_{ii} = 1,\ i = 1, \ldots, n,\ \text{and}\ \mathrm{rank}(X) = 1. \end{array} \tag{7.4}$$

The SDP (7.3) can be regarded as a relaxation of (7.4) obtained by removing the non-convex rank constraint. In fact, this is how we will later formulate a similar relaxation for the minimum bisection problem, in Chapter 8.

We now turn to property (b), and consider the problem of forming a solution to (7.1) from a solution to (7.3). From the solution $\{u_i\}_{i=1,\ldots,n}$ of the relaxed problem (7.3), we produce a cut subset S' by randomly picking a vector $r \in \mathbb{R}^n$ from the uniform distribution on the unit sphere and setting

$$S' = \{i | r^T u_i \geq 0\}$$

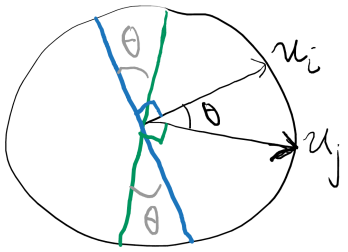


Fig. 7.1: Illustration of the relationship between the angle between vectors and their inner product, $\theta = \arccos(u_i^T u_j)$

In other words, we separate the vectors $u_1, \ldots, u_n$ by a random hyperplane (perpendicular to r). We will show that the cut given by the set S' is comparable to the optimal one.

Let W be the value of the cut produced by the procedure described above. Note that W is a random variable, whose expectation is easily seen (see Figure 7.1) to be given by

$$\begin{aligned}\mathbb{E}[W] &= \sum_{i<j} w_{ij} \Pr\left\{\text{sign}(r^T u_i) \neq \text{sign}(r^T u_j)\right\} \\ &= \sum_{i<j} w_{ij} \frac{1}{\pi} \arccos(u_i^T u_j).\end{aligned}$$

If we define α_{GW} as

$$\alpha_{GW} = \min_{-1 \leq x \leq 1} \frac{\frac{1}{\pi}\arccos(x)}{\frac{1}{2}(1-x)},$$

it can be shown that $\alpha_{GW} > 0.87$ (see, for example [60]).

By linearity of expectation

$$\mathbb{E}[W] = \sum_{i<j} w_{ij} \frac{1}{\pi} \arccos(u_i^T u_j) \geq \alpha_{GW} \frac{1}{2} \sum_{i<j} w_{ij}(1 - u_i^T u_j). \tag{7.5}$$

Let $\text{MaxCut}(G)$ be the maximum cut of G, meaning the maximum of the original problem (7.1). Naturally, the optimal value of (7.2) is larger or equal than $\text{MaxCut}(G)$. Hence, an algorithm that solves (7.2) and uses the random rounding procedure described above produces a cut W satisfying

$$\text{MaxCut}(G) \geq \mathbb{E}[W] \geq \alpha_{GW} \frac{1}{2} \sum_{i<j} w_{ij}(1 - u_i^T u_j) \geq \alpha_{GW} \text{MaxCut}(G), \tag{7.6}$$

thus having an approximation ratio (in expectation) of α_{GW}. Note that one can run the randomized rounding procedure several times and select the best cut.[1] We thus have

$$\text{MaxCut}(G) \geq \mathbb{E}[W] \geq \alpha_{GW}\mathcal{R}\text{MaxCut}(G) \geq \alpha_{GW}\text{MaxCut}(G)$$

Can α_{GW} be improved?

A natural question is to ask whether there exists a polynomial time algorithm that has an approximation ratio better than α_{GW}.

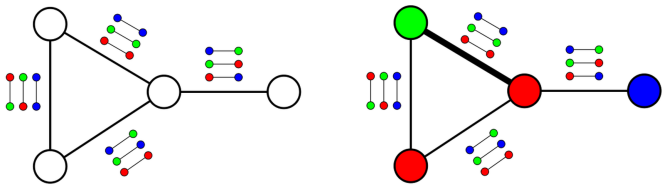


Fig. 7.2: The Unique Games Problem

The unique games problem (as depicted in Figure 7.2) is the following: Given a graph and a set of k colors, and, for each edge, a matching between the colors, the goal in the unique games problem is to color the vertices as to agree with as high of a fraction of the edge matchings as possible. For example, in Figure 7.2 the coloring agrees with $\frac{3}{4}$ of the edge constraints, and it is easy to see that one cannot do better.

The Unique Games Conjecture of Khot [74], has played a major role in hardness of approximation results.

Conjecture 7.2. For any $\varepsilon > 0$, the problem of distinguishing whether an instance of the Unique Games Problem is such that it is possible to agree with a $\geq 1 - \varepsilon$ fraction of the constraints or it is not possible to even agree with a ε fraction of them, is NP-hard.

There is a sub-exponential time algorithm capable of distinguishing such instances of the unique games problem [13], however no polynomial time algorithm has been found so far. At the moment one of the strongest candidates to break the Unique Games Conjecture is a relaxation based on the Sum-of-squares hierarchy that we will discuss below.

Remarkably, approximating `Max-Cut` with an approximation ratio better than α_{GW} is as hard as refuting the Unique Games Conjecture (UG-hard) [75].

[1] It is worth noting that one is only guaranteed to solve 7.2 up to an approximation of ε from its optimum value. However, since this ε can be made arbitrarily small, one can get the approximation ratio arbitrarily close to α_{GW}.

More generality, if the Unique Games Conjecture is true, the semidefinite programming approach described above produces optimal approximation ratios for a large class of problems [108].

Not depending on the Unique Games Conjecture, there is a NP-hardness of approximation of $\frac{16}{17}$ for `Max-Cut` [66].

Remark 7.3. Note that a simple greedy method that assigns membership to each vertex as to maximize the number of edges cut involving vertices already assigned achieves an approximation ratio of $\frac{1}{2}$ (even of $\frac{1}{2}$ of the total number of edges, not just of the optimal cut).

7.1 A Sums-of-Squares interpretation

We now give a different interpretation to the approximation ratio obtained above. Let us first slightly reformulate the problem (recall that $w_{ii} = 0$).

Recall from Proposition 4.3 that a cut can be rewritten as a quadratic form involving the graph Laplcian. We can rewrite (7.1) as

$$\begin{aligned} \max\ & \tfrac{1}{4} y^T L_G y \\ & y_i = \pm 1,\ i = 1, \ldots, n. \end{aligned} \tag{7.7}$$

Similarly, (7.3) can be written (by taking $X = yy^T$) as

$$\begin{aligned} \max\ & \tfrac{1}{4} \operatorname{Tr}(L_G X) \\ s.t.\ & X \succeq 0 \\ & X_{ii} = 1,\ i = 1, \ldots, n. \end{aligned} \tag{7.8}$$

In Chapter 8 we will derive the the dual program to (7.8) in the context of recovery in the Stochastic Block Model. Here we will simply state it, and show weak duality as it will be important for the argument that follows.

$$\begin{aligned} \min\ & \operatorname{Tr}(D) \\ s.t.\ & D \text{ is a diagonal matrix} \\ & D - \tfrac{1}{4} L_G \succeq 0. \end{aligned} \tag{7.9}$$

Indeed, if X is a feasible solution to (7.8) and D a feasible solution to (7.9) then, since X and $D - \frac{1}{4} L_G$ are both positive semidefinite $\operatorname{Tr}\left[X\left(D - \frac{1}{4} L_G\right)\right] \geq 0$ which gives

$$0 \leq \operatorname{Tr}\left[X\left(D - \frac{1}{4} L_G\right)\right] = \operatorname{Tr}(XD) - \frac{1}{4}\operatorname{Tr}(L_G X) = \operatorname{Tr}(D) - \frac{1}{4}\operatorname{Tr}(L_G X),$$

since D is diagonal and $X_{ii} = 1$. This shows weak duality, the fact that the value of (7.9) is larger than the one of (7.8).

If certain conditions, the so called Slater conditions [136, 135], are satisfied then the optimal values of both programs are known to coincide, this is known

as strong duality. In this case, the Slater conditions ask whether there is a matrix strictly positive definite that is feasible for (7.8), and the identity is such a matrix. This means that there exists $D^\natural$ feasible for (7.9) such that

$$\mathrm{Tr}(D^\natural) = \mathcal{R}\mathrm{MaxCut}.$$

Hence, for any $y \in \mathbb{R}^n$ we have

$$\frac{1}{4}y^T L_G y = \mathcal{R}\mathrm{MaxCut} - y^T \left(D^\natural - \frac{1}{4}L_G\right)^T y + \sum_{i=1}^n D^\natural_{ii}\left(y_i^2 - 1\right). \qquad (7.10)$$

Note that (7.10) certifies that no cut of G is larger than $\mathcal{R}\mathrm{MaxCut}$. Indeed, if $y \in \{\pm 1\}^2$ then $y_i^2 = 1$ and so

$$\mathcal{R}\mathrm{MaxCut} - \frac{1}{4}y^T L_G y = y^T \left(D^\natural - \frac{1}{4}L_G\right)^T y.$$

Since $D^\natural - \frac{1}{4}L_G \succeq 0$, there exists V such that $D^\natural - \frac{1}{4}L_G = VV^T$ with the columns of V denoted by $v_1, \ldots, v_n$, meaning that $y^T\left(D^\natural - \frac{1}{4}L_G\right)^T y = \left\|V^T y\right\|^2 = \sum_{k=1}^n (v_k^T y)^2$. Hence, for any $y \in \{\pm 1\}^2$,

$$\mathcal{R}\mathrm{MaxCut} - \frac{1}{4}y^T L_G y = \sum_{k=1}^n (v_k^T y)^2.$$

In other words, $\mathcal{R}\mathrm{MaxCut} - \frac{1}{4}y^T L_G y$ is, for y in the hypercube ($y \in \{\pm 1\}^2$), a sum-of-squares of degree 2. This is known as a sum-of-squares certificate [26, 25, 102, 80, 117, 101]; indeed, if a real-valued polynomial is a sum-of-squares naturally it is non-negative.

Note that, by definition, $\mathrm{MaxCut} - \frac{1}{4}y^T L_G y$ is always non-negative on the hypercube. This does not mean, however, that it needs to be a sum-of-squares[2] of degree 2.

The remarkable fact is that sum-of-squares certificates of at most a specified degree can be found using Semidefinite programming [102, 80]. In fact, SDPs (7.8) and (7.9) are finding the smallest real number Λ such that $\Lambda - \frac{1}{4}y^T L_G y$ is a sum-of-squares of degree 2 over the hypercube. The dual SDP is finding a certificate as in (7.10) while the primal is in some sense constraining the degree 2 moments of y $X_{ij} = y_i y_j$ (we recommend [25] for nice lecture notes on sum-of-squares; see also Remark 7.4). Many natural questions remain open towards a precise understanding of the power of SDPs corresponding to higher degree sum-of-squares certificates.

Remark 7.4 (triangular inequalities and Sum of squares level 4). A natural follow-up question is whether the relaxation of degree 4 is actually strictly

[2] This is related with Hilbert's 17th problem [114] and Stengle's Positivstellensatz [118]

tighter than the one of degree 2 for Max-Cut (in the sense of forcing extra constraints). What follows is an interesting set of inequalities that degree 4 enforces and that degree 2 doesn't, known as triangular inequalities. This example helps illustrate the differences between Sum-of-Squares certificates of different degree.

Since $y_i \in \{\pm 1\}$ we naturally have that, for all i, j, k

$$y_i y_j + y_j y_k + y_k y_i \geq -1,$$

this would mean that, for $X_{ij} = y_i y_j$ we would have,

$$X_{ij} + X_{jk} + X_{ik} \geq -1,$$

however it is not difficult to see that the SDP (7.8) of degree 2 is only able to constraint

$$X_{ij} + X_{jk} + X_{ik} \geq -\frac{3}{2},$$

which is considerably weaker. There are a few different ways of thinking about this, one is that the three vector u_i, u_j, u_k in the relaxation may be at an angle of 120 degrees with each other. Another way of thinking about this is that the inequality $y_i y_j + y_j y_k + y_k y_i \geq -\frac{3}{2}$ can be proven using sum-of-squares proof with degree 2:

$$(y_i + y_j + y_k)^2 \geq 0 \quad \Rightarrow \quad y_i y_j + y_j y_k + y_k y_i \geq -\frac{3}{2}$$

However, the stronger constraint cannot.

On the other hand, if degree 4 monomials are involved, (let's say $X_S = \prod_{s \in S} y_s$, note that $X_\emptyset = 1$ and $X_{ij} X_{ik} = X_{jk}$) then the constraint

$$\begin{bmatrix} X_\emptyset \\ X_{ij} \\ X_{jk} \\ X_{ki} \end{bmatrix} \begin{bmatrix} X_\emptyset \\ X_{ij} \\ X_{jk} \\ X_{ki} \end{bmatrix}^T = \begin{bmatrix} 1 & X_{ij} & X_{jk} & X_{ki} \\ X_{ij} & 1 & X_{ik} & X_{jk} \\ X_{jk} & X_{ik} & 1 & X_{ij} \\ X_{ki} & X_{jk} & X_{ij} & 1 \end{bmatrix} \succeq 0$$

implies $X_{ij} + X_{jk} + X_{ik} \geq -1$ just by taking

$$\mathbf{1}^T \begin{bmatrix} 1 & X_{ij} & X_{jk} & X_{ki} \\ X_{ij} & 1 & X_{ik} & X_{jk} \\ X_{jk} & X_{ik} & 1 & X_{ij} \\ X_{ki} & X_{jk} & X_{ij} & 1 \end{bmatrix} \mathbf{1} \geq 0.$$

Also, note that the inequality $y_i y_j + y_j y_k + y_k y_i \geq -1$ can indeed be proven using sum-of-squares proof with degree 4 (recall that $y_i^2 = 1$):

$$(1 + y_i y_j + y_j y_k + y_k y_i)^2 \geq 0 \quad \Rightarrow \quad y_i y_j + y_j y_k + y_k y_i \geq -1.$$

Interestingly, it is known [76] that these extra inequalities alone will not increase the approximation power (in the worst case) of (7.3).

8

Community Detection and the Power of Convex Relaxations

The problem of detecting communities in network data is a central problem in data science, examples of interest include social networks, the internet, or biological and ecological networks. In Chapter 4 we discussed clustering in the context of graphs, and described performance guarantees for spectral clustering (based on Cheeger's Inequality) that made no assumptions on the underlying graph. While these guarantees are remarkable, they are worst-case and hence pessimistic in nature. In an effort to understand the performance of some of these approaches on more realistic models of data, we will now analyze a generative model for graphs with community structure, the stochastic block model. On the methodology side, we will focus on convex relaxations, based on semidefinite programming (as in Chapter 7), and will show that this approach achieves exact recovery of the communities on graphs drawn from this model. The techniques developed to prove these guarantees mirror the ones used to prove analogous guarantees for a variety of other problems where convex relaxations yield exact recovery.

8.1 The Stochastic Block Model

The Stochastic Block Model is a random graph model that produces graphs with a community structure. While, as with any model, we do not expect it to capture all properties of a real world network (examples include network hubs, power-law degree distributions, and other structures) the goal is to study a simple graph model that produces community structure, as a test bed for understanding fundamental limits of community detection and analyzing the performance of recovery algorithms.

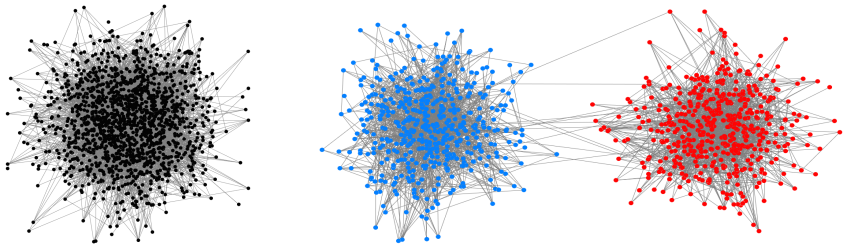
Definition 8.1 (Stochastic Block Model). *Let n and k be positive integers representing respectively the number of nodes and communities, $c \in [k]^n$ be the vector of community memberships for the different nodes, and $P \in [0,1]^{k \times k}$ a symmetric matrix of connectivity probabilities. A graph G is said to be drawn*

from the Stochastic Block Model on n nodes, when for each pair of nodes (i, j) the probability that $(i, j) \in E$ is independent from all other edges and given by P_{c_i,c_j}.

We will focus on the special case of the two communities ($k = 2$) balanced symmetric block model where n is even, both communities are of the same size, and

$$P = \begin{bmatrix} p & q \\ q & p \end{bmatrix},$$

where $p, q \in [0, 1]$ are constants, cf. Figure 8.1. Furthermore, we will focus on the associative case ($p > q$), while noting that all that follows can be easily adapted to the disassociate case ($q > p$). . We note also that when $p = q$ this model reduces to the classical Erdős-Renyí model described in Chapter 4. Since there are only two communities we will identify their membership labels with $+1$ and -1.



(a) (b)

Fig. 8.1: A graph generated form the stochastic block model with 600 nodes and 2 communities, scrambled in Fig. 8.1(a), clustered and color-coded in Fig. 8.1(b). Nodes in this graph connect with probability $p = 6/600$ within communities and $q = 0.1/600$ across communities. (Image courtesy of Emmanuel Abbe.)

Many fascinating questions can be asked in the context of this model. Natural questions include to characterize statistics of the model, such as number of triangles or larger cliques. In this chapter, motivated by the problem of community detection, we are interested in understanding when is it possible to reconstruct, or estimate, the community memberships from an observation of the graph, and what efficient algorithms succeed at this inference task.

Before proceeding we note that the difficulty of this problem should certainly depend on the value of p and q. As illustrative examples, this problem is trivial when $p = 1$ and $q = 0$ and hopeless when $p = q$ (notice that even in the easy case the actual membership can only be determined up to a re-labeling of the communities). As $p > q$, we will attempt to recover the original partition

by trying to compute the minimum bisection of the graph; while related to the Max-Cut problem described in Chapter 7, notice how the objective here is to produce the minimum balanced cut.

8.2 Spike Model Prediction

A natural approach is to draw motivation from Chapter 4 and to use a form of spectral clustering to attempt to partition the graph.

Let A be the adjacency matrix of G,

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise.} \end{cases} \tag{8.1}$$

Note that in our model, A is a random matrix. We would like to solve

$$\begin{aligned} \max \ & \sum_{i,j} A_{ij} x_i x_j \\ \text{s.t. } & x_i = \pm 1, \forall_i \\ & \sum_j x_j = 0, \end{aligned} \tag{8.2}$$

The optimal solution x of (8.2) takes the value $+1$ on one side of a partition and -1 on the other side, where the partition is balanced and achieves the minimum cut between the resulting clusters.

Relaxing the condition $x_i = \pm 1, \ \forall_i$ to $\|x\|_2^2 = n$ would yield a spectral method

$$\begin{aligned} \max \ & \sum_{i,j} A_{ij} x_i x_j \\ \text{s.t. } & \|x\|_2 = \sqrt{n} \\ & \mathbf{1}^T x = 0 \end{aligned} \tag{8.3}$$

The solution of (8.3) corresponds to the leading eigenvector of the matrix obtained by projecting A on the orthogonal complement of the all-ones vector $\mathbf{1}$.

The matrix A is a random matrix whose expectation is given[1] by

$$\mathbb{E}[A] = \begin{cases} p & \text{if } i \text{ and } j \text{ are in the same community} \\ q & \text{otherwise.} \end{cases}$$

[1]For simplicity we assume that self-loops also have probability p. This does not affect any of the conclusions, as it does not give information about the community memberships.

Let g denote the vector corresponding to the true community memberships, with entries $+1$ and -1; note that this is the vector we want to recover.[2] We can write

$$\mathbb{E}[A] = \frac{p+q}{2}\mathbf{1}\mathbf{1}^T + \frac{p-q}{2}gg^T,$$

and

$$A = \big(A - \mathbb{E}[A]\big) + \frac{p+q}{2}\mathbf{1}\mathbf{1}^T + \frac{p-q}{2}gg^T.$$

In order to remove the term $\frac{p+q}{2}\mathbf{1}\mathbf{1}^T$ we consider the random matrix

$$\mathcal{A} = A - \frac{p+q}{2}\mathbf{1}\mathbf{1}^T.$$

It is easy to see that

$$\mathcal{A} = \big(\mathcal{A} - \mathbb{E}[\mathcal{A}]\big) + \frac{p-q}{2}gg^T.$$

This means that $\mathcal{A}$ is the sum of a random matrix whose expected value is zero and a rank-1 matrix, i.e.

$$\mathcal{A} = W + \lambda vv^T$$

where $W = \big(\mathcal{A} - \mathbb{E}[\mathcal{A}]\big)$ and $\lambda vv^T = \frac{p-q}{2}n\left(\frac{g}{\sqrt{n}}\right)\left(\frac{g}{\sqrt{n}}\right)^T$. In Chapter 3 we saw that for a large enough rank-1 additive perturbation to a Wigner matrix, there is an eigenvalue associated with the perturbation that pops outside of the distribution of eigenvalues of a Wigner Gaussian matrix W. Moreover, whenever this happens, we saw that the leading eigenvector has a non-trivial correlation with g.

Since $\mathcal{A}$ is simply A minus a multiple of $\mathbf{1}\mathbf{1}^T$, problem (8.3) is equivalent to

$$\begin{aligned} \max \ & \sum_{i,j} \mathcal{A}_{ij}x_ix_j \\ \text{s.t. } & \|x\|_2 = \sqrt{n} \\ & \mathbf{1}^T x = 0 \end{aligned} \tag{8.4}$$

Since we have subtracted a multiple of $\mathbf{1}\mathbf{1}^T$, we will drop the constraint $\mathbf{1}^T x = 0$. Notice how a deviation from $\mathbf{1}^T x = 0$ would be penalized in the new objective, the fact that the multiple we subtracted is sufficient for us to drop the constraint will be confirmed by the success of the new optimization problem, now given by

[2] We want to recover either g or $-g$, as they correspond to different labelings of the same community structure.

$$\max \sum_{i,j} \mathcal{A}_{ij} x_i x_j$$
$$\text{s.t. } \|x\|_2 = \sqrt{n}, \tag{8.5}$$

whose solution corresponds to the leading eigenvector of $\mathcal{A}$.

Recall that if $\mathcal{A} - \mathbb{E}[\mathcal{A}]$ is a Wigner matrix with i.i.d. entries with zero mean and variance σ^2 then its empirical spectral density follows the semicircle law and it is essentially supported in $[-2\sigma\sqrt{n}, 2\sigma\sqrt{n}]$. We would then expect the top eigenvector of $\mathcal{A}$ to correlate with g as long as

$$\frac{p-q}{2} n > \frac{2\sigma\sqrt{n}}{2}. \tag{8.6}$$

Unfortunately $\mathcal{A} - \mathbb{E}[\mathcal{A}]$ is not a Wigner matrix in general. In fact, half of its entries have variance $p(1-p)$ while the variance of the other half is $q(1-q)$.

Putting rigor aside for a second, if we were to take $\sigma^2 = \frac{p(1-p)+q(1-q)}{2}$ then (8.6) would suggest that the leading eigenvector of $\mathcal{A}$ correlates with the true partition vector g as long as

$$\frac{p-q}{2} > \frac{1}{\sqrt{n}} \sqrt{\frac{p(1-p)+q(1-q)}{2}}, \tag{8.7}$$

This argument is of course not valid, because the matrix in question is not a Wigner matrix. The special case $q = 1-p$ can be easily salvaged, since all entries of $\mathcal{A} - \mathbb{E}[\mathcal{A}]$ have the same variance and they can be made to be identically distributed by conjugating with gg^T. This is still an impressive result, it says that if $p = 1-q$ then $p-q$ needs only to be around $\frac{1}{\sqrt{n}}$ to be able to make an estimate that correlates with the original partitioning!

An interesting regime (motivated, for example, by friendship networks in social sciences) is when the average degree of each node is constant. This can be achieved by taking $p = \frac{a}{n}$ and $q = \frac{b}{n}$ for constants a and b. While the argument presented to justify condition (8.7) is not valid in this setting, it nevertheless suggests that the condition on a and b needed to be able to make an estimate that correlates with the original partition, often referred to as partial recovery, is

$$(a-b)^2 > 2(a+b). \tag{8.8}$$

Remarkably this was posed as a conjecture by Decelle et al. [45] and proved in a series of works by Mossel et al. [98, 97] and Massoulie [92]. While describing the proof of this conjecture is outside the scope of this book, we note that the conjectures were obtained by studying fixed points of a certain linearization of belief propagation using techniques from statistical physics. The lower bound can be proven by showing contiguity between the two models below the phase transition, and the upper bound is obtained by analyzing an algorithm that is an adaptation of belief propagation and studying the so-called non-backtracking operator. We refer the reader to the excellent survey of Abbe [4] and references therein for further reading.

Remark 8.2 (More than three communities). The balanced symmetric stochastic block model with $k > 3$ communities is conjectured to have a fascinating statistical-to-computational gap. In the sparse regime of inner probability $p = \frac{a}{n}$ and outer probability $q = \frac{b}{n}$ it is believed that, for $k > 3$ there is a regime of the parameters a and b such that the problem of partially recovering the community memberships is statistically, or information-theoretically, possible but that there does not exist a polynomial-time algorithm to perform this task. These conjectures are based on insight obtained with tools from statistical physics. We refer the reader to [45, 140, 59, 3] for further reading.

8.3 Exact recovery

We now turn our attention to the problem of recovering the cluster membership of every single node correctly, not simply having an estimate that correlates with the true labels. We will keep our focus on the balanced, symmetric, two communities setting and briefly describe extensions later. If the inner-probability is $p = \frac{a}{n}$ then it is not hard to show that each cluster will have isolated nodes, making it impossible to recover the membership of every possible node correctly. In fact this is the case whenever $p \leq \frac{(2-\varepsilon)\log n}{n}$, for some $\varepsilon > 0$. For that reason we focus on the regime

$$p = \frac{\alpha \log(n)}{n} \text{ and } q = \frac{\beta \log(n)}{n}, \tag{8.9}$$

for some constants $\alpha > \beta$.

A natural algorithm would be to compute the minimum bisection (8.2) which corresponds to the Maximum Likelihood Estimator, and also the Maximum a Posteriori Estimator when the community memberships are drawn uniformly at random. In fact, it is known (see [1] for a proof) that if

$$\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}, \tag{8.10}$$

then, with high probability, (8.2) recovers the true partition. Moreover, if

$$\sqrt{\alpha} - \sqrt{\beta} < \sqrt{2},$$

no algorithm can, with high probability, recover the true partition.

In this section we will analyze a semidefinite programming relaxation, analogous to the ones described in Chapter 7 for Max-Cut. By making use of convex duality, we will derive conditions for exact recovery with this particular algorithm, reducing the problem to a problem in random matrix theory. We will present a solution to the resulting random matrix question, using the matrix concentration tools developed in Chapter 6. While not providing a the strongest known guarantee, this approach is extremely adaptable and can be used to solve a large number of similar questions.

8.4 A semidefinite relaxation

Let $x \in \mathbb{R}^n$ with $x_i = \pm 1$ represent a partition of the nodes (recall that there is an ambiguity in the sense that x and $-x$ represent the same partition). Note that if we remove the constraint $\sum_j x_j = 0$ in (8.2) then the optimal solution becomes $x = \mathbf{1}$. Let us define $B = 2A - (\mathbf{1}\mathbf{1}^T - I)$, meaning that

$$B_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } (i,j) \in E(G) \\ -1 & \text{otherwise} \end{cases} \tag{8.11}$$

It is clear that the problem

$$\begin{aligned} \max \ & \sum_{i,j} B_{ij} x_i x_j \\ \text{s.t. } & x_i = \pm 1, \forall_i \\ & \sum_j x_j = 0 \end{aligned} \tag{8.12}$$

has the same solution as (8.2). However, when the constraint is dropped,

$$\begin{aligned} \max \ & \sum_{i,j} B_{ij} x_i x_j \\ \text{s.t. } & x_i = \pm 1, \forall_i, \end{aligned} \tag{8.13}$$

$x = \mathbf{1}$ is no longer an optimal solution. As with (8.5) above, the penalization created by subtracting a large multiple of $\mathbf{1}\mathbf{1}^T$ will be enough to discourage unbalanced partitions (the reader may notice the connection with Lagrangian duality). In fact, (8.13) is the problem we will set ourselves to solve.

Unfortunately (8.13) is in general NP-hard (one can encode, for example, *Max-Cut* by picking an appropriate B). We will relax it to an easier problem by the same technique used to approximate the Max-Cut problem in the previous section (this technique is often known as *matrix lifting*). If we write $X = xx^T$ then we can formulate the objective of (8.13) as

$$\sum_{i,j} B_{ij} x_i x_j = x^T B x = \mathrm{Tr}(x^T B x) = \mathrm{Tr}(B x x^T) = \mathrm{Tr}(BX)$$

Also, the condition $x_i = \pm 1$ implies $X_{ii} = x_i^2 = 1$. This means that (8.13) is equivalent to

$$\begin{aligned} \max \quad & \mathrm{Tr}(BX) \\ \text{s.t.} \quad & X_{ii} = 1, \forall_i \\ & X = xx^T \text{ for some } x \in \mathbb{R}^n. \end{aligned} \tag{8.14}$$

The fact that $X = xx^T$ for some $x \in \mathbb{R}^n$ is equivalent to $\text{rank}(X) = 1$ and $X \succeq 0$. This means that (8.13) is equivalent to

$$\begin{aligned} \max \quad & \text{Tr}(BX) \\ \text{s.t.} \quad & X_{ii} = 1, \forall_i \\ & X \succeq 0 \\ & \text{rank}(X) = 1. \end{aligned} \tag{8.15}$$

We now relax the problem by removing the non-convex rank constraint

$$\begin{aligned} \max \quad & \text{Tr}(BX) \\ \text{s.t.} \quad & X_{ii} = 1, \forall_i \\ & X \succeq 0. \end{aligned} \tag{8.16}$$

This is an SDP that can be solved (up to arbitrary precision) in polynomial time [135].

Since we removed the rank constraint, the solution to (8.16) is no longer guaranteed to be rank-1. We will take a different approach from the one we used in Chapter 7 to obtain an approximation ratio for `Max-Cut`, which was a worst-case approximation ratio guarantee. What we will show is that, for some values of α and β, with high probability, the solution to (8.16) not only satisfies the rank constraint but it coincides with $X = gg^T$ where g corresponds to the true partition. From X one can compute g by simply calculating its leading eigenvector.

8.5 Convex Duality

A standard technique to show that a candidate solution is the optimal one for a convex problem is to use convex duality.

We will describe duality with a game theoretical intuition in mind. The idea will be to rewrite (8.16) without imposing constraints on X but rather have the constraints be implicitly enforced. Consider the following optimization problem.

$$\max_{X} \quad \min_{\substack{Z,\, Q \\ Z \text{ is diagonal} \\ Q \succeq 0}} \quad \text{Tr}(BX) + \text{Tr}(QX) + \text{Tr}\left(Z\left(I_{n\times n} - X\right)\right). \tag{8.17}$$

Let us provide a game theoretical interpretation for (8.17). Suppose that there is a *primal player* (picking X) whose objective is to maximize the objective and a *dual player* (picking Z and Q after seeing X) trying to make the objective as small as possible. If the primal player does not pick X satisfying the constraints of (8.16) then we claim that the dual player is capable of driving the objective to $-\infty$. Indeed, if there is an i for which $X_{ii} \neq 1$ then the

dual player can simply pick $Z_{ii} = -c\frac{1}{1-X_{ii}}$ and make the objective as small as desired by taking a large enough c. Similarly, if X is not positive semidefinite, then the dual player can take $Q = cvv^T$ where v is such that $v^T X v < 0$. If, on the other hand, X satisfies the constraints of (8.16) then

$$\operatorname{Tr}(BX) \leq \min_{\substack{Z,\ Q \\ Z \text{ is diagonal} \\ Q \succeq 0}} \operatorname{Tr}(BX) + \operatorname{Tr}(QX) + \operatorname{Tr}\left(Z\left(I_{n\times n} - X\right)\right).$$

Since equality can be achieved if for example the dual player picks $Q = 0_{n\times n}$, then it is evident that the values of (8.16) and (8.17) are the same:

$$\max_{\substack{X, \\ X_{ii} \leq 1\ \forall_i \\ X \succeq 0}} \operatorname{Tr}(BX) = \max_{X} \min_{\substack{Z,\ Q \\ Z \text{ is diagonal} \\ Q \succeq 0}} \operatorname{Tr}(BX) + \operatorname{Tr}(QX) + \operatorname{Tr}\left(Z\left(I_{n\times n} - X\right)\right)$$

With this game theoretical intuition in mind, it is clear that if we change the "rules of the game" and have the dual player decide their variables before the primal player (meaning that the primal player can pick X knowing the values of Z and Q) then it is clear that the objective can only increase, which means that:

$$\max_{\substack{X, \\ X_{ii} \leq 1\ \forall_i \\ X \succeq 0}} \operatorname{Tr}(BX) \leq \min_{\substack{Z,\ Q \\ Z \text{ is diagonal} \\ Q \succeq 0}} \max_{X} \operatorname{Tr}(BX) + \operatorname{Tr}(QX) + \operatorname{Tr}\left(Z\left(I_{n\times n} - X\right)\right).$$

Note that we can rewrite

$$\operatorname{Tr}(BX) + \operatorname{Tr}(QX) + \operatorname{Tr}\left(Z\left(I_{n\times n} - X\right)\right) = \operatorname{Tr}\left(\left(B + Q - Z\right)X\right) + \operatorname{Tr}(Z).$$

When playing:

$$\min_{\substack{Z,\ Q \\ Z \text{ is diagonal} \\ Q \succeq 0}} \max_{X} \operatorname{Tr}\left(\left(B + Q - Z\right)X\right) + \operatorname{Tr}(Z),$$

if the dual player does not set $B + Q - Z = 0_{n\times n}$ then the primal player can drive the objective value to $+\infty$, this means that the dual player is forced to chose $Q = Z - B$ and so we can write

$$\min_{\substack{Z,\ Q \\ Z \text{ is diagonal} \\ Q \succeq 0}} \max_{X} \operatorname{Tr}\left(\left(B + Q - Z\right)X\right) + \operatorname{Tr}(Z) = \min_{\substack{Z, \\ Z \text{ is diagonal} \\ Z - B \succeq 0}} \max_{X} \operatorname{Tr}(Z),$$

which clearly does not depend on the choices of the primal player. This means that

$$\max_{\substack{X, \\ X_{ii} \leq 1\ \forall_i \\ X \succeq 0}} \operatorname{Tr}(BX) \leq \min_{\substack{Z, \\ Z \text{ is diagonal} \\ Z - B \succeq 0}} \operatorname{Tr}(Z).$$

This is known as weak duality (strong duality says that, under some conditions the two optimal values actually match, see for example [135], recall that we used strong duality when giving a sum-of-squares interpretation to the Max-Cut approximation ratio, a similar interpretation can be given in this problem, see [18]).

Also, the problem

$$\begin{array}{ll} \min & \mathrm{Tr}(Z) \\ \text{s.t.} & Z \text{ is diagonal} \\ & Z - B \succeq 0 \end{array} \tag{8.18}$$

is called the dual problem of (8.16).

The derivation above explains why the objective value of the dual problem is always greater or equal to the primal problem. Nevertheless, there is a much simpler proof (although not as enlightening): let X, Z be a feasible point of (8.16) and (8.18), respectively. Since Z is diagonal and $X_{ii} = 1$, it follows that $\mathrm{Tr}(ZX) = \mathrm{Tr}(Z)$. Also, $Z - B \succeq 0$ and $X \succeq 0$, therefore $\mathrm{Tr}[(Z - B)X] \geq 0$. Altogether,

$$\mathrm{Tr}(Z) - \mathrm{Tr}(BX) = \mathrm{Tr}[(Z - B)X] \geq 0,$$

as stated.

Recall that we want to show that gg^T is the optimal solution of (8.16). Then, if we find Z diagonal, such that $Z - B \succeq 0$ and

$\mathrm{Tr}[(Z - B)gg^T] = 0$, (this condition is known as complementary slackness)

then $X = gg^T$ must be an optimal solution of (8.16). To ensure that gg^T is the unique solution we just have to ensure that the nullspace of $Z - B$ only has dimension 1 (which corresponds to multiples of g). Essentially, if this is the case, then for any other possible solution X one could not satisfy complementary slackness.

This means that if we can find Z with the following properties:

(1) Z is diagonal
(2) $\mathrm{Tr}[(Z - B)gg^T] = 0$
(3) $Z - B \succeq 0$
(4) $\lambda_2(Z - B) > 0$,

then gg^T is the unique optimum of (8.16) and so recovery of the true partition is possible (with an efficient algorithm). Z is known as the dual certificate, or dual witness.

8.6 Building the dual certificate

The idea to build Z is to construct it to satisfy properties (1) and (2) and try to show that it satisfies (3) and (4) using concentration. In fact, since Z is

diagonal this design problem has n free variables. If $Z - B \succeq 0$ then condition (2) is equivalent to $(Z - B)g = 0$ which provides n equations, as the resulting linear system is non-singular, the candidate arising from using conditions (1) and (2) will be unique.

A couple of preliminary definitions will be useful before writing out the candidate Z. Recall that the degree matrix D of a graph G is a diagonal matrix where each diagonal coefficient D_{ii} corresponds to the number of neighbors of vertex i and that $\lambda_2(M)$ is the second smallest eigenvalue of a symmetric matrix M.

Definition 8.3. *Let $\mathcal{G}_+$ (resp. $\mathcal{G}_-$) be the subgraph of G that includes the edges that link two nodes in the same community (resp. in different communities) and A the adjacency matrix of G. We denote by $D^+_{\mathcal{G}}$ (resp. $D^-_{\mathcal{G}}$) the degree matrix of $\mathcal{G}_+$ (resp. $\mathcal{G}_-$) and define the Stochastic Block Model Laplacian to be*

$$L_{SBM} = D^+_{\mathcal{G}} - D^-_{\mathcal{G}} - A.$$

Note that the inclusion of self loops does not change L_{SBM}. Also, we point out that L_{SBM} is not in general positive-semidefinite.

Now we are ready to construct the candidate Z. Condition (2) implies that we need $Z_{ii} = \frac{1}{g_i} B[i,:]g$. Since $B = 2A - (\mathbf{1}\mathbf{1}^T - I)$ we have

$$Z_{ii} = \frac{1}{g_i}(2A - (\mathbf{1}\mathbf{1}^T - I))[i,:]g = 2\frac{1}{g_i}(Ag)_i + 1,$$

meaning that

$$Z = 2(D^+_{\mathcal{G}} - D^-_{\mathcal{G}}) + I.$$

This is our candidate dual witness. As a result

$$Z - B = 2(D^+_{\mathcal{G}} - D^-_{\mathcal{G}}) - I - \left[2A - (\mathbf{1}\mathbf{1}^T - I)\right] = 2L_{SBM} + 11^T.$$

It trivially follows (by construction) that

$$(Z - B)g = 0.$$

This immediately gives the following lemma.

Lemma 8.4. *Let L_{SBM} denote the Stochastic Block Model Laplacian as defined in Definition 8.3. If*

$$\lambda_2(2L_{SBM} + 11^T) > 0, \tag{8.19}$$

then the relaxation (8.14) *recovers the true partition.*

Note that $2L_{\text{SBM}}+11^T$ is a random matrix and so this reduces to "an exercise" in random matrix theory.

8.7 Matrix Concentration

In this section we show how the resulting question amounts to controlling the largest eigenvalue of a random matrix, which can be done with the matrix concentration tools described in Chapter 6.

Let us start by noting that

$$\mathbb{E}\left[2L_{\mathrm{SBM}} + 11^T\right] = 2\mathbb{E}L_{\mathrm{SBM}} + 11^T = 2\mathbb{E}D_{\mathcal{G}}^{+} - 2\mathbb{E}D_{\mathcal{G}}^{-} - 2\mathbb{E}A + 11^T,$$

and $\mathbb{E}D_{\mathcal{G}}^{+} = \frac{n}{2}\frac{\alpha\log(n)}{n}I$, $\mathbb{E}D_{\mathcal{G}}^{-} = \frac{n}{2}\frac{\beta\log(n)}{n}I$. Moreover, $\mathbb{E}A$ is a matrix with four $\frac{n}{2}\times\frac{n}{2}$ blocks where the diagonal blocks have entries $\frac{\alpha\log(n)}{n}$ and the off-diagonal blocks have entries $\frac{\beta\log(n)}{n}$.[3] In other words

$$\mathbb{E}A = \frac{1}{2}\left(\frac{\alpha\log(n)}{n} + \frac{\beta\log(n)}{n}\right)11^T + \frac{1}{2}\left(\frac{\alpha\log(n)}{n} - \frac{\beta\log(n)}{n}\right)gg^T.$$

This means that

$$\mathbb{E}\left[2L_{\mathrm{SBM}} + 11^T\right] = \left((\alpha-\beta)\log n\right)I + \left(1 - (\alpha+\beta)\frac{\log n}{n}\right)11^T - (\alpha-\beta)\frac{\log n}{n}gg^T.$$

Since $2L_{\mathrm{SBM}}g = 0$ we can safely ignore what happens in the span of g, and it is not hard to see that

$$\lambda_2\left(\mathbb{E}\left[2L_{\mathrm{SBM}} + 11^T\right]\right) = (\alpha-\beta)\log n.$$

Thus, it is enough to show that

$$\|L_{\mathrm{SBM}} - \mathbb{E}\left[L_{\mathrm{SBM}}\right]\| < \frac{\alpha-\beta}{2}\log n, \tag{8.20}$$

which is a large deviation inequality; recall that $\|\cdot\|$ denotes operator norm.

The idea is to write $L_{\mathrm{SBM}} - \mathbb{E}\left[L_{\mathrm{SBM}}\right]$ as a sum of independent random matrices and use the Matrix Bernstein Inequality (Theorem 6.1). This gives an illustrative example of the applicability of matrix concentration tools, as many random matrices of interest can be rewritten as sums of independent matrices.

Let us start by defining, for i and j in the same community (i.e. $g_i = g_j$),

$$\gamma_{ij}^{+} = \begin{cases} 1 & \text{if } (i,j)\in E \\ 0 & \text{otherwise,} \end{cases}$$

and

[3] For simplicity we assume the possibility of self-loops; notice however that the matrix in question does not depend on this, only its decomposition in the degree matrices and A.

$$\Delta_{ij}^{+} = (e_i - e_j)(e_i - e_j)^T,$$

where e_i (resp. e_j) is the vector of all zeros except the i^{th} (resp. j^{th}) coefficient which is 1.

For i and j in different communities (i.e. $g_i \neq g_j$), define

$$\gamma_{ij}^{-} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\Delta_{ij}^{-} = -(e_i + e_j)(e_i + e_j)^T.$$

We have

$$L_{\text{SBM}} = \sum_{i<j : g_i = g_j} \gamma_{ij}^{+} \Delta_{ij}^{+} + \sum_{i<j : g_i \neq g_j} \gamma_{ij}^{-} \Delta_{ij}^{-}.$$

We note how $(\gamma_{ij}^{+})_{i,j}$ and $(\gamma_{ij}^{-})_{i,j}$ are jointly independent random variables with expectations $\mathbb{E}(\gamma_{ij}^{+}) = \frac{\alpha \log n}{n}$ and $\mathbb{E}(\gamma_{ij}^{-}) = \frac{\beta \log n}{n}$. Δ_{ij}^{+} and Δ_{ij}^{-} are deterministic matrices. This means that

$$L_{\text{SBM}} - \mathbb{E}L_{\text{SBM}} = \sum_{\substack{i<j: \\ g_i = g_j}} \left(\gamma_{ij}^{+} - \frac{\alpha \log n}{n} \right) \Delta_{ij}^{+} + \sum_{i<j:\ g_i \neq g_j} \left(\gamma_{ij}^{-} - \frac{\beta \log n}{n} \right) \Delta_{ij}^{-}.$$

We can then use Theorem 6.1 by setting

$$\sigma^2 = \left\| \text{Var}\left[\gamma^{+}\right] \sum_{i<j:\ g_i = g_j} \left(\Delta_{ij}^{+}\right)^2 + \text{Var}\left[\gamma^{-}\right] \sum_{i<j:\ g_i \neq g_j} \left(\Delta_{ij}^{-}\right)^2 \right\|, \tag{8.21}$$

and $R = 2$, since $\left\|\Delta_{ij}^{+}\right\| = \left\|\Delta_{ij}^{-}\right\| = 2$ and both $(\gamma_{ij}^{+})_{i,j}$ and $(\gamma_{ij}^{-})_{i,j}$ take values in $[-1, 1]$. Note how this bound is for the spectral norm of the summands, not just the largest eigenvalue, as our goal is to bound the spectral norm of the random matrix. In order to compute σ^2, we write

$$\sum_{i<j:\ g_i = g_j} \left(\Delta_{ij}^{+}\right)^2 = nI - \left(\mathbf{1}\mathbf{1}^T + gg^T\right),$$

and

$$\sum_{i<j:\ g_i \neq g_j} \left(\Delta_{ij}^{-}\right)^2 = nI + \left(\mathbf{1}\mathbf{1}^T - gg^T\right)$$

.

Since $\text{Var}\left[\gamma^{+}\right] \leq \frac{\alpha \log n}{n}$, $\text{Var}\left[\gamma^{-}\right] \leq \frac{\beta \log n}{n}$, and all the summands are positive semidefinite we have

$$\sigma^2 \leq \left\| \frac{(\alpha + \beta)\log n}{n} \left(nI - gg^T\right) - \frac{(\alpha - \beta)\log n}{n} \mathbf{1}\mathbf{1}^T \right\| = (\alpha + \beta)\log n.$$

Using Theorem 6.1 for $t = \frac{\alpha-\beta}{2}\log n$ on both the largest and smallest eigenvalue gives

$$\begin{aligned}\mathbb{P}\left\{\|L_{\mathrm{SBM}} - \mathbb{E}\left[L_{\mathrm{SBM}}\right]\| < \frac{\alpha-\beta}{2}\log n\right\} \leq \\ &\leq 2n \cdot \exp\left(\frac{-\left(\frac{\alpha-\beta}{2}\log n\right)^2}{2\left(\alpha+\beta\right)\log n + \frac{4}{3}\left(\frac{\alpha-\beta}{2}\log n\right)}\right) \\ &= 2 \cdot \exp\left(-\frac{(\alpha-\beta)^2 \log n}{8\left(\alpha+\beta\right) + \frac{8}{3}\left(\alpha-\beta\right)} + \log n\right) \\ &= 2n^{-\left(\frac{(\alpha-\beta)^2}{8(\alpha+\beta)+\frac{8}{3}(\alpha-\beta)}-1\right)}.\end{aligned}$$

Together with Lemma 8.4, this implies that as long as

$$(\alpha-\beta)^2 > 8\left(\alpha+\beta\right) + \frac{8}{3}\left(\alpha-\beta\right), \tag{8.22}$$

the semidefinite programming relaxation (8.14) recovers the true partition, with probability tending to 1 as n increases.

While it is possible to obtain a stronger guarantee for this relaxation, the derivation above illustrates the matrix concentration technique in a simple, yet powerful, instance. In fact, the analysis in [1] uses the same technique. In order to obtain a sharp guarantee (Theorem 8.5 below) one needs more specialized tools. We refer the interested reader to [17] or [64] for a discussion and proof of Theorem 8.5; the main idea is to separate the diagonal from the non-diagonal part of $L_{\mathrm{SBM}} - \mathbb{E}\left[L_{\mathrm{SBM}}\right]$, treating the former with scalar concentration inequalities, and the latter with specialized matrix concentration inequalities such as the ones in [24].

Theorem 8.5. *Let G be a random graph with n nodes drawn according to the stochastic block model on two communities with edge probabilities p and q. Let $p = \frac{\alpha\log n}{n}$ and $q = \frac{\beta\log n}{n}$, where $\alpha > \beta$ are constants. Then, as long as*

$$\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}, \tag{8.23}$$

the semidefinite program considered above coincides with the true partition with high probability.

Note that, if

$$\sqrt{\alpha} - \sqrt{\beta} < \sqrt{2}, \tag{8.24}$$

then exact recovery of the communities is impossible, meaning that the SDP algorithm is optimal. Furthermore, in this regime (8.24), one can show that there will be a node on each community that is more connected to the other community than to its own, meaning that a partition that swaps them would

have more likelihood. The fact that the SDP will start working essentially when this starts happening appears naturally in the analysis in [17]. More recently it has been proven that the spectral method (8.3), followed by a simple thresholding step, also gives exact recovery of the communities [2]. An analogous analysis has recently been obtained for the (normalized or unnormalized) graph Laplacian in place of the adjacency matrix, see [46]. However, the proof techniques for the graph Laplacian are different and a bit more involved, since—unlike the adjacency matrix—the graph Laplacian does not exhibit row/column-wise independence.

Remark 8.6. An important advantage of semidefinite relaxations is that they are often able to produce certificates of optimality. Indeed, if the solution of the relaxation (8.14) is rank 1 then the user is sure that it must be the solution of (8.13). These advantages, and ways of producing such certificates while bypassing the need to solve the semidefinite program are discussed in [18].

9

Linear Dimension Reduction via Random Projections

In Chapters 3 and 5 we saw both linear and non-linear methods for dimension reduction. In this chapter we will see one of the most fascinating consequences of the phenomenon of concentration of measure in high dimensions, one of the *blessings of high dimensions* described in Chapter 2. When given a data set in high dimensions, we will see that it is sometimes the case that a projection to a lower dimensional space, taken at random, preserves certain geometric features of the data set. The remarkable aspect here is that this "lower" dimension can be strikingly lower. This allows for significant computational savings in many data processing tasks by including a random projection as a pre-processing step. There is however another less obvious implication of this phenomenon with important practical implications: since the projection is agnostic of the data, it can be leveraged even when the data set is not explicit, such as the set of all *natural images* or the set of all "possible" *brain scans*; this is at the heart of *Compressed Sensing*.

9.1 The Johnson-Lindenstrauss Lemma

Suppose one has n points, $X = \{x_1, \ldots, x_n\}$, in $\mathbb{R}^p$ (with p large). If $d > n$, the points actually lie in a subspace of dimension n, so the projection $f : \mathbb{R}^p \to \mathbb{R}^n$ of the points to that subspace acts without distorting the geometry of X. In particular, for every x_i and x_j, $\|f(x_i) - f(x_j)\|^2 = \|x_i - x_j\|^2$, meaning that f is an isometry in X. Suppose instead we allow a bit of distortion, and look for a map $f : \mathbb{R}^p \to \mathbb{R}^d$ that is an $\varepsilon-$isometry, meaning that

$$(1 - \varepsilon)\|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1 + \varepsilon)\|x_i - x_j\|^2. \tag{9.1}$$

Can we do better than $d = n$?

In 1984, Johnson and Lindenstrauss [71] showed a remarkable lemma that answers this question affirmatively.

Theorem 9.1 (Johnson-Lindenstrauss Lemma [71]). *For any $0 < \varepsilon < 1$ and for any integer n, let d be such that*

$$d \geq 4 \frac{1}{\varepsilon^2/2 - \varepsilon^3/3} \log n. \tag{9.2}$$

Then, for any set X of n points in $\mathbb{R}^d$, there is a linear map $f : \mathbb{R}^p \to \mathbb{R}^d$ that is an $\varepsilon-$isometry for X (see (9.1)). This map can be found in randomized polynomial time.

We follow [44] for an elementary proof for the Theorem. We need a few concentration of measure bounds, we will omit the proof of those but they are available in [44] and are essentially the same ideas as those used to show the concentration inequalities in Chapter 2.

Lemma 9.2 (see [44]). *Let $y_1, \ldots, y_p$ be i.i.d standard Gaussian random variables and $Y = (y_1, \ldots, y_d)$. Let $g : \mathbb{R}^p \to \mathbb{R}^d$ be the projection into the first d coordinates and $Z = g\left(\frac{Y}{\|Y\|}\right) = \frac{1}{\|Y\|}(y_1, \ldots, y_d)$ and $L = \|Z\|^2$. It is clear that $\mathbb{E}L = \frac{d}{p}$. In fact, L is very concentrated around its mean*

- *If $\beta < 1$,*
$$\Pr\left[L \leq \beta \frac{d}{p}\right] \leq \exp\left(\frac{d}{2}(1 - \beta + \log \beta)\right).$$
- *If $\beta > 1$,*
$$\Pr\left[L \geq \beta \frac{d}{p}\right] \leq \exp\left(\frac{d}{2}(1 - \beta + \log \beta)\right).$$

Proof. [of Johnson-Lindenstrauss Lemma]

We will start by showing that, given a pair x_i, x_j a projection onto a random subspace of dimension k will satisfy (after appropriate scaling) property (9.1) with high probability. Without loss of generality we can assume that $u = x_i - x_j$ has unit norm. Understanding what is the norm of the projection of u on a random subspace of dimension d is the same as understanding the norm of the projection of a (uniformly) random point on S^{p-1} the unit sphere in $\mathbb{R}^p$ on a specific d-dimensional subspace—let us say the one generated by the first d canonical basis vectors.

This means that we are interested in the distribution of the norm of the first k entries of a random vector drawn from the uniform distribution over S^{p-1} – this distribution is the same as taking a standard Gaussian vector in $\mathbb{R}^p$ and normalizing it to the unit sphere.

Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be the projection on a random $k-$dimensional subspace and let $f : \mathbb{R}^d \to \mathbb{R}^k$ defined as $f = \sqrt{\frac{d}{k}} g$. Then (by the above discussion), given a pair of distinct x_i and x_j, $\frac{\|f(x_i) - f(x_j)\|^2}{\|x_i - x_j\|^2}$ has the same distribution as $\frac{d}{k} L$, as defined in Lemma 9.2. Using Lemma 9.2, we have, given a pair x_i, x_j,

$$\Pr\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \leq (1-\varepsilon)\right] \leq \exp\left(\frac{d}{2}(1-(1-\varepsilon)+\log(1-\varepsilon))\right),$$

since for $\varepsilon \geq 0$, $\log(1-\varepsilon) \leq -\varepsilon - \varepsilon^2/2$ we have

$$\begin{aligned}\Pr\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \leq (1-\varepsilon)\right] &\leq \exp\left(-\frac{k\varepsilon^2}{4}\right)\\ &\leq \exp\left(-2\log n\right) = \frac{1}{n^2}.\end{aligned}$$

On the other hand,

$$\Pr\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \geq (1+\varepsilon)\right] \leq \exp\left(\frac{k}{2}(1-(1+\varepsilon)+\log(1+\varepsilon))\right).$$

since for $\varepsilon \geq 0$, $\log(1+\varepsilon) \leq \varepsilon - \varepsilon^2/2 + \varepsilon^3/3$ we have

$$\begin{aligned}\mathbb{P}\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \geq (1+\varepsilon)\right] &\leq \exp\left(-\frac{k\left(\varepsilon^2-2\varepsilon^3/3\right)}{4}\right)\\ &\leq \exp\left(-2\log n\right) = \frac{1}{n^2}.\end{aligned}$$

By the union bound it follows that

$$\Pr\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \notin [1-\varepsilon, 1+\varepsilon]\right] \leq \frac{2}{n^2}.$$

Since there exist $\binom{n}{2}$ such pairs, again, a simple union bound gives

$$\Pr\left[\exists_{i,j} : \frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \notin [1-\varepsilon, 1+\varepsilon]\right] \leq \frac{2}{n^2}\frac{n(n-1)}{2} = 1-\frac{1}{n}.$$

Therefore, choosing f as a properly scaled projection onto a random k-dimensional subspace gives an ε-isometry on X (see (9.1)) with probability at least $\frac{1}{n}$. We can achieve any desirable constant probability of success by trying $\mathcal{O}(n)$ such random projections, meaning we can find an $\varepsilon-$isometry in randomized polynomial time.

□

Note that by considering k slightly larger one can get a good projection on the first random attempt with high confidence. In fact, it is trivial to adapt the proof above to obtain the following lemma:

Lemma 9.3. *For any $0 < \varepsilon < 1$, $\tau > 0$, and for any integer n, let k be such that*

$$d \geq \frac{2(2+\tau)}{\varepsilon^2/2 - \varepsilon^3/3}\log n.$$

Then, for any set X of n points in $\mathbb{R}^p$, take $f : \mathbb{R}^p \to \mathbb{R}^d$ to be a suitably scaled projection on a random subspace of dimension d, then f is an $\varepsilon-$isometry for X (see (9.1)) with probability at least $1 - \frac{1}{n^\tau}$.

Lemma 9.3 is quite remarkable. Consider the situation where we are given a high-dimensional data set in a streaming fashion – meaning that we get each data point at a time, consecutively. To run a dimension-reduction technique like PCA or Diffusion maps we would need to wait until we received the last data point and then compute the dimension reduction map (both PCA and Diffusion Maps are, in some sense, data adaptive). Using Lemma 9.3 one can just choose a projection at random in the beginning of the process (all one needs to know is an estimate of the logarithm of the size of the data set) and just map each point using this projection matrix which can be done online – we do not need to see the next point to compute the projection of the current data point. Lemma 9.3 ensures that this (seemingly naïve) procedure will, with high probably, not distort the data by more than ε.

One might wonder if such low-dimensional embeddings as provided by the Johson-Lindenstrauss Lemma also extend to other norms besides the Euclidean norm. For the ℓ_1-norm there exist lower bounds which prevent such a dramatic dimension reduction (see [84]), and for the ℓ_∞-norm one can easily construct examples that demonstrate the impossibility of dimension reduction. Hence, the Johnson-Lindenstrauss Lemma seems to be a subtle result about the Euclidean norm.

9.1.1 The Fast Johnson-Lindenstrauss transform and optimality

Let us continue thinking about our example of high-dimensional streaming data. After we draw the random projection matrix[1], say M, for each data point x, we still have to compute Mx which has a computational cost of $\mathcal{O}(\varepsilon^{-2}\log(n)p)$ since M has $\mathcal{O}(\varepsilon^{-2}\log(n)p)$ entries (since M is a random matrix, generically it will be a dense matrix). In some applications this might be too expensive, raising the natural question of whether one can do better. Moreover, storing a large-scale dense matrix M is not very desirable either. There is no hope of significantly reducing the number of rows in general, as it is known that the Johnson-Lindenstrauss Lemma is orderwise optimal [8, 79].

We might hope to replace the dense random matrix M by a sparse matrix S to speed up the matrix-vector multiplication and to reduce the storage requirements. This method was proposed and analyzed in [6]. Here we discuss a slightly simplified version, see also [43].

We let S be a very sparse $k \times d$ matrix, where each row of S has just one single non-zero entry of value $\sqrt{d/p}$ at a location drawn uniformly at random. Then, for any vector $x \in \mathbb{R}^p$

$$\mathop{\mathbb{E}}_{i}[(Sx)_i^2] = \sum_{j=1}^{d} \mathbb{P}(i=j) \cdot \frac{d}{k} \cdot x_j^2 = \frac{1}{k}\|x\|_2^2,$$

[1] An orthogonal projection P must satisfy $P = P^*$ and $P^2 = P$. Here, it is not M that represents a projection, but M^*M, yet for our purposes of approximate norm-preserving dimension reduction it suffices to apply M instead of M^*M. However, with a slight abuse of terminology, we still refer to M as projection.

hence $\mathbb{E}[\|Sx\|_2^2] = \mathbb{E}[\sum_{i=1}^{k}(Sx_i)^2] = \|x\|_2^2$. This result is satisfactory with respect to expectation (even for $k = 1$), but not with respect to the variance of $\|Sx\|_2^2$. For instance, if x has only one non-zero entry we need $k \sim \mathcal{O}(p)$ to ensure that $\|Sx\|_2^2 \neq 0$. More generally, if one coordinate of x is much larger (in absolute value) than all its other coordinates, then we will need a rather large value for d to guarantee that $\|Sx\|_2 \approx \|x\|_2$.

A natural way to quantify the "peakiness" of a vector x is via the *peak-to-average ratio*[2] measured by the quantity $\|x\|_\infty/\|x\|_2$. It is easy to see that we have (assuming x is not the zero-vector)

$$\frac{1}{\sqrt{p}} \leq \frac{\|x\|_\infty}{\|x\|_2} \leq 1.$$

The upper bounds is achieved by vectors with only one non-zero entry, while the lower bound is met by constant-modulus vectors. Thus, if

$$\frac{\|x\|_\infty}{\|x\|_2} \approx \frac{1}{\sqrt{p}}, \tag{9.3}$$

we can hope that sparse subsampling of x will still preserve its Euclidean norm.

Thus, this suggests to include a preprocessing step by applying a rotation so that sparse vectors become non-sparse in the new basis, thereby reducing their ∞-norm (while their 2-norm remains invariant under rotation). Two natural choices for such a rotation are the Discrete Fourier transform (which maps unit-vectors into constant modulus vectors) and its $\mathbb{Z}_2$-cousin, the Walsh-Hadamard matrix[3]. But since the Fast Johnson-Lindenstrauss Transform (FJLT) has to work for all vectors, we need to avoid that this rotation maps dense vectors into sparse vectors. We can address this issue by "randomizing" the rotation, thereby ensuring with overwhelming probability that dense vectors are not mapped into sparse vectors. This can be accomplished in a numerically efficient manner (thus maintaining our overall goal of numerical efficiency) by first randomizing the signs of x before applying the rotation. Putting these steps together we arrive at the following map.

Definition 9.4. *The* Fast Johnson-Lindenstrauss Transform *is the map* $\Psi : \mathbb{R}^p \to \mathbb{R}^d$*, defined by* $\Psi := SFD$*, where* S *and* D *are random matrices and* F *is a deterministic matrix. In particular,*

[2]This quantity also plays an important role in wireless communications. There, one tries to avoid transmitting signals with a large peak-to-average ratio, since such signals would suffer from nonlinear distortions when they are passing through those cheap power amplifiers that are usually installed in cell phones. The potentially large peak-to-average ratio of OFDM signals is one of the alleged reasons why CDMA was dominant over OFDM for such a long time.

[3]Hadamard matrices do not exist for all dimension d. But we can always pad x with zeroes to achieve the desired length.

- *S is a $d \times p$ matrix, where each row of S has just one single non-zero entry of value $\sqrt{d/p}$ at a location drawn uniformly at random.*
- *F is either the $p \times p$ DFT matrix or the $p \times p$ Hadamard matrix (if it exists), in each case normalized by $1/\sqrt{p}$ to obtain a unitary matrix.*
- *D is a $p \times p$ diagonal matrix whose entries are drawn independently from $\{-1, +1\}$ with probability $1/2$.*

We can carry out the matrix-vector multiplication by the DFT matrix via the Fast Fourier Transform (FFT) in $\mathcal{O}(p \log p)$ operations; a similar algorithm exists for the Walsh-Hadamard matrix. The FJLT allows for a dimension reduction that is competitive with the Johnson-Lindenstrauss Lemma as manifested by the following theorem.

Theorem 9.5 (Fast Johnson-Lindenstrauss Transform). *There is a random matrix Ψ of size $d \times p$ with $d = \mathcal{O}\big(\log(d/\delta)\log(1/\delta)/\varepsilon^2\big)$ such that, for each $x \in \mathbb{R}^p$,*

$$\|\Psi x\|_2 \in [1-\varepsilon, 1+\varepsilon] \cdot \|x\|_2$$

holds with probability at least $1-\delta$. Matrix-vector multiplication with Ψ takes $\mathcal{O}(p \log p + d)$ operations.

The proof of this theorem follows from the two lemmas below. We first show that with high probability the random rotation FD produces vectors with a sufficiently low peak-to-average ratio.

Lemma 9.6. *Let $y = FDx$, where F and D are as in Definition 9.4. Then*

$$\mathop{\mathbb{P}}_{D}\left(\frac{\|y\|_\infty}{\|y\|_2} \geq \frac{2\log(4p/\delta)}{p}\right) \leq \frac{\delta}{2}. \tag{9.4}$$

Proof. Since FD is unitary, the quantity $\|FDx\|_\infty / \|FDx\|_2$ is invariant under rescaling of x and therefore we can assume $\|x\|_2 = 1$.

Let $\xi_i = \pm 1$ be the i-th diagonal entry of D. We have $y_i = \sum_{j=1}^d \varepsilon_j F_{ij} x_j$ and note that the terms of this sum are i.i.d. bounded random variables. We thus can apply Hoeffding's inequality. In the notation of Theorem 2.14, let $X_j = \varepsilon_j F_{ij} x_j$. We note that $X_j = \pm F_{ij} x_j$, hence $\mathbb{E}[X_j] = 0$ and $|X_j| \leq a_j$, where $a_j = F_{ij} x_j$. It holds that

$$\sum_{j=1}^d a_j^2 = \sum_{j=1}^d F_{ij}^2 x_j^2 = \sum_{j=1}^d \frac{1}{d} x_j^2 = \frac{\|x\|_2^2}{d} = \frac{1}{d}.$$

We can now use Theorem 2.14 with $t = \sqrt{2\log(4p/\delta)/p}$ and obtain

$$\mathbb{P}\left(|y_i| > \sqrt{\frac{2\log(4p/\delta)}{p}}\right) \leq 2\exp\left(-\frac{2\log(4p/\delta)/p}{2/p}\right) = \frac{\delta}{2d}.$$

Applying the union bound finishes the proof. □

Lemma 9.7. *Conditioned on the event that* $\|y\|_\infty \gtrsim \frac{2\log(4p/\delta)}{p}$*, it holds that*

$$\mathbb{P}\big(\|Sy\|_2^2 - 1| \le \varepsilon\big) \le 1 - \frac{\delta}{2}.$$

Proof. We use the following Chernoff bound: Given independent random variables $X_1, \dots, X_n$, $X = \sum_i X_i$, $\mu = \mathbb{E}X, \sigma^2 = \mathbb{E}[(X - \mathbb{E}X)^2]$, and $|X_i| \le K$ with probability 1,

$$\mathbb{P}(|X - \mu| > t) \lesssim \max\{e^{-ct^2/\sigma^2}, e^{-ct/K}\}.$$

We denote $S_{ji} = \sqrt{d/k}\delta_{ji}$, where $\delta_{ji} \in \{0, 1\}$ is our random sample of the columns for row j. Hence for all j, $\sum_{i=1}^p \delta_{j,i} = 1$. We write $z := Sy$ and compute

$$\begin{aligned} q_j &= z_j^2 \\ &= \frac{d}{k}\left(\sum_{i=1}^d \delta_{ji} y_i\right)^2 \\ &= \frac{d}{k}\left(\sum_i \delta_{ji} y_i^2 + \sum_{i \ne \ell} \delta_{ji}\delta_{j\ell} y_i y_\ell\right) \\ &= \frac{d}{k}\sum_i \delta_{ji}. \end{aligned}$$

We care about $z^2 = \sum_j q_j$. Since the q_j's are independent, we can apply the aforementioned Chernoff bound, provided we bound σ^2 and K, which we will do now.

$$K \le \frac{p}{d}\|y\|_\infty^2 \lesssim \frac{\log(p/\delta)}{d}.$$

$$\begin{aligned} \sigma^2 \le k\mathbb{E}[q_1^2] &= k\mathbb{E}\left[\frac{d^2}{k^2}\sum_i \delta_{ij} y_i^4\right] \\ &= \frac{p}{d}\sum_i y_i^4 \\ &= \|y\|_\infty^2 \frac{p}{d}\sum_i y_i^2 \\ &= \frac{p}{d}\|y\|_\infty \\ &\lesssim \frac{\log(p/\delta)}{d}. \end{aligned}$$

Plugging these terms into the Chernoff bound yields

$$\mathbb{P}\big(|\|Sy\|^2 - 1| > \varepsilon\big) = \mathbb{P}\left(\Big|\sum_j q_j - 1\Big| > \varepsilon\right) \lesssim \max\left\{e^{-\frac{c\varepsilon^2 d}{\log(p/\delta)}}, e^{-\frac{c\varepsilon d}{\log(p/\delta)}}\right\}.$$

Since the first term in the right hand side above dominates, we can choose $k \sim 1/\varepsilon^2 \log(p/\delta)\log(1/\delta)$ to get the desired $\delta/2$-bound. □

Combining Lemma 9.6 and Lemma 9.7 establishes Theorem 9.5.

Besides the potential speedup, another advantage of the FJLT is that it requires significantly less memory compared to storing an unstructured random projection matrix as is the case for the standard Johnson-Lindenstrauss approach.

9.2 Gordon's Theorem

In the last section we showed that in order to approximately preserve the distances (up to $1 \pm \varepsilon$) between n points, it suffices to randomly project them to $\Theta\left(\varepsilon^{-2}\log n\right)$ dimensions. The key argument was that a random projection approximately preserves the norm of every point in a set S, in this case the set of differences between pairs of n points. What we showed is that in order to approximately preserve the norm of every point in S, it is enough to project to $\Theta\left(\varepsilon^{-2}\log|S|\right)$ dimensions. The question this section is meant to answer is: can this be improved if S has a special structure? Given a set S, what is the measure of complexity of S that explains how many dimensions one needs to project on to still approximately preserve the norms of points in S. We shall see below that this will be captured—via Gordon's Theorem—by the so called *Gaussian width* of S.

Definition 9.8 (Gaussian width). *Given a closed set $S \subset \mathbb{R}^p$, its* Gaussian width $\omega(S)$ *is define as:*

$$\omega(S) = \mathbb{E}\max_{x\in S}\left[g_p^T x\right],$$

where $g_p \sim \mathcal{N}(0, I_{p\times p})$.

Similarly to the proof of Theorem 9.1 we will restrict our attention to sets S of unit norm vectors, meaning that $S \subset \mathbb{S}^{p-1}$.

Also, we will focus our attention not in random projections but in the similar model of random linear maps $G : \mathbb{R}^p \to \mathbb{R}^d$ that are given by matrices with i.i.d. Gaussian entries. For this reason the following proposition will be useful:

Proposition 9.9. *Let $g_d \sim \mathcal{N}(0, I_{d\times d})$, and define*

$$a_d := \mathbb{E}\|g_d\|.$$

Then $\sqrt{\frac{d}{d+1}}\sqrt{d} \le a_d \le \sqrt{d}$.

We are now ready to present Gordon's Theorem.

Theorem 9.10 (Gordon's Theorem [63]). *Let $G \in \mathbb{R}^{d\times p}$ a random matrix with independent $\mathcal{N}(0,1)$ entries and $S \subset \mathbb{S}^{p-1}$ be a closed subset of the unit sphere in p dimensions. Let $x \in \mathbb{R}^p$. Then*

$$\mathbb{E}\max_{x\in S}\left\|\frac{1}{a_d}Gx\right\| \leq 1 + \frac{\omega(S)}{a_d},$$

and

$$\mathbb{E}\min_{x\in S}\left\|\frac{1}{a_d}Gx\right\| \geq 1 - \frac{\omega(S)}{a_d},$$

where $a_d = \mathbb{E}\|g_d\|$ and $\omega(S)$ is the Gaussian width of S. Recall that $\sqrt{\frac{d}{d+1}}\sqrt{d} \leq a_d \leq \sqrt{d}$.

Before proving Gordon's Theorem we will note some of its direct implications. The theorem suggests that $\frac{1}{a_d}G$ preserves the norm of the points in S up to $1 \pm \frac{\omega(S)}{a_d}$, indeed we can make this precise with Gaussian concentration (Theorem 6.2).

Note that the function $F(G) = \max_{x\in S}\|Gx\|$ is 1-Lipschitz. Indeed

$$\begin{aligned}\left|\max_{x_1\in S}\|G_1x_1\| - \max_{x_2\in S}\|G_2x_2\|\right| &\leq \max_{x\in S}|\|G_1x\| - \|G_2x\|| \leq \max_{x\in S}\|(G_1-G_2)\,x\| \\ &= \|G_1 - G_2\| \leq \|G_1 - G_2\|_F.\end{aligned}$$

Similarly, one can show that $F(G) = \min_{x\in S}\|Gx\|$ is 1-Lipschitz. Thus, one can use Gaussian concentration to get

$$\mathbb{P}\left\{\max_{x\in S}\|Gx\| \geq a_d + \omega(S) + t\right\} \leq \exp\left(-\frac{t^2}{2}\right), \tag{9.5}$$

and

$$\mathbb{P}\left\{\min_{x\in S}\|Gx\| \leq a_d - \omega(S) - t\right\} \leq \exp\left(-\frac{t^2}{2}\right). \tag{9.6}$$

This gives us the following theorem.

Theorem 9.11. *Let $G \in \mathbb{R}^{d\times p}$ a random matrix with independent $\mathcal{N}(0,1)$ entries and $S \subset \mathbb{S}^{p-1}$ be a closed subset of the unit sphere in p dimensions. Then, for $\varepsilon > \sqrt{\frac{\omega(S)^2}{a_d^2}}$, with probability $\geq 1 - 2\exp\left[-\frac{a_d^2}{2}\left(\varepsilon - \frac{\omega(S)}{a_d}\right)^2\right]$:*

$$(1-\varepsilon)\|x\| \leq \left\|\frac{1}{a_d}Gx\right\| \leq (1+\varepsilon)\|x\|,$$

for all $x \in S$.

Recall that $d\frac{d}{d+1} \leq a_d^2 \leq k$.

Proof. This is readily obtained by taking $\varepsilon = \frac{\omega(S)+t}{a_d}$, using (9.5) and (9.6). □

Remark 9.12. Note that a simple use of a union bound[4] shows that $\omega(S) \lesssim \sqrt{2\log|S|}$, which means that taking d to be of the order of $\log|S|$ suffices to ensure that $\frac{1}{a_d}G$ to have the Johnson Lindenstrauss property. This observation shows that Theorem 9.11 essentially directly implies Theorem 9.1 (although not exactly, since $\frac{1}{a_d}G$ is not a projection).

9.2.1 Gordon's Escape Through a Mesh Theorem

Theorem 9.11 suggests that, if $\omega(S) \leq a_d$, a uniformly chosen random subspace of $\mathbb{R}^n$ of dimension $(n-d)$ (which can be seen as the nullspace of G) avoids a set S with high probability. This is indeed the case and is known as Gordon's Escape Through a Mesh Theorem (Corollary 3.4. in Gordon's original paper [63]). See also [95] for a description of the proof. We include the Theorem below for the sake of completeness.

Theorem 9.13 (Corollary 3.4. in [63]). *Let $S \subset \mathbb{S}^{p-1}$ be a closed subset of the unit sphere in p dimensions. If $\omega(S) < a_d$, then for a $(p-d)$-dimensional subspace Λ drawn uniformly from the Grassmanian manifold we have*

$$\mathbb{P}\left\{\Lambda \cap S \neq \emptyset\right\} \leq \frac{7}{2}\exp\left(-\frac{1}{18}\left(a_d - \omega(S)\right)^2\right),$$

where $\omega(S)$ is the Gaussian width of S and $a_d = \mathbb{E}\|g_d\|$ where $g_k \sim \mathcal{N}(0, I_{d\times d})$.

9.2.2 Proof of Gordon's Theorem

In order to prove this Theorem we will use extensions of the Slepian's Comparison Lemma. This, and the closely related Sudakov-Fernique inequality, are crucial tools to compare Gaussian processes. A Gaussian process is a family of Gaussian random variables indexed by some set T, $\{X_t\}_{t\in T}$ (if T is finite this is simply a Gaussian vector). Given a Gaussian process X_t, a particular quantity of interest is $\mathbb{E}\left[\max_{t\in T} X_t\right]$. Intuitively, if we have two Gaussian processes X_t and Y_t with mean zero $\mathbb{E}\left[X_t\right] = \mathbb{E}\left[Y_t\right] = 0$, for all $t \in T$, and the same variance, then the process that has the "least correlations" should have a larger maximum (think the maximum entry of vector with i.i.d. Gaussian entries versus one always with the same Gaussian entry). The following inequality makes this intuition precise and extends it to processes with different variances.[5]

[4] This follows from the fact that the maximum of n standard Gaussian random variables is $\lesssim \sqrt{2\log n}$.

[5] Although intuitive in some sense, this turns out to be a delicate statement about Gaussian random variables, as it does not hold in general for other distributions.

Theorem 9.14 (Slepian/Sudakov-Fernique inequality).
Let $\{X_u\}_{u\in U}$ and $\{Y_u\}_{u\in U}$ be two (almost surely bounded) centered Gaussian processes indexed by the same (compact) set U. If, for every $u_1, u_2 \in U$:

$$\mathbb{E}\left[X_{u_1} - X_{u_2}\right]^2 \leq \mathbb{E}\left[Y_{u_1} - Y_{u_2}\right]^2, \tag{9.7}$$

then

$$\mathbb{E}\left[\max_{u\in U} X_u\right] \leq \mathbb{E}\left[\max_{u\in U} Y_u\right].$$

The following extension is due to Gordon [62, 63].

Theorem 9.15. *[Theorem A in [63]] Let $\{X_{t,u}\}_{(t,u)\in T\times U}$ and $\{Y_{t,u}\}_{(t,u)\in T\times U}$ be two (almost surely bounded) centered Gaussian processes indexed by the same (compact) sets T and U. If, for every $t_1, t_2 \in T$ and $u_1, u_2 \in U$:*

$$\mathbb{E}\left[X_{t_1,u_1} - X_{t_1,u_2}\right]^2 \leq \mathbb{E}\left[Y_{t_1,u_1} - Y_{t_1,u_2}\right]^2, \tag{9.8}$$

and, for $t_1 \neq t_2$,

$$\mathbb{E}\left[X_{t_1,u_1} - X_{t_2,u_2}\right]^2 \geq \mathbb{E}\left[Y_{t_1,u_1} - Y_{t_2,u_2}\right]^2, \tag{9.9}$$

then

$$\mathbb{E}\left[\min_{t\in T}\max_{u\in U} X_{t,u}\right] \leq \mathbb{E}\left[\min_{t\in T}\max_{u\in U} Y_{t,u}\right].$$

Note that Theorem 9.14 easily follows by setting $|T| = 1$.

We are now ready to prove Gordon's Theorem.

Proof. [of Theorem 9.10]

Let $G \in \mathbb{R}^{d\times p}$ with i.i.d. $\mathcal{N}(0,1)$ entries. We define two Gaussian processes: For $v \in S \subset \mathbb{S}^{p-1}$ and $u \in \mathbb{S}^{d-1}$ let $g \sim \mathcal{N}\left(0, I_{d\times d}\right)$ and $h \sim \mathcal{N}\left(0, I_{p\times p}\right)$ and define the following processes:

$$A_{u,v} = g^T u + h^T v,$$

and

$$B_{u,v} = u^T G v.$$

For all $v, v' \in S \subset \mathbb{S}^{p-1}$ and $u, u' \in \mathbb{S}^{d-1}$,

$$\begin{aligned}
\mathbb{E}\left|A_{v,u} - A_{v',u'}\right|^2 - \mathbb{E}\left|B_{v,u} - B_{v',u'}\right|^2 &= 4 - 2\left(u^T u' + v^T v'\right) - \sum_{ij}\left(v_i u_j - v_i' u_j'\right)^2 \\
&= 4 - 2\left(u^T u' + v^T v'\right) - \left[2 - 2\left(v^T v'\right)\left(u^T u'\right)\right] \\
&= 2 - 2\left(u^T u' + v^T v' - u^T u' v^T v'\right) \\
&= 2\left(1 - u^T u'\right)\left(1 - v^T v'\right).
\end{aligned}$$

This means that $\mathbb{E}\left|A_{v,u} - A_{v',u'}\right|^2 - \mathbb{E}\left|B_{v,u} - B_{v',u'}\right|^2 \geq 0$ and $\mathbb{E}\left|A_{v,u} - A_{v',u'}\right|^2 - \mathbb{E}\left|B_{v,u} - B_{v',u'}\right|^2 = 0$ if $v = v'$. This implies that we can use Theorem 9.15 with $X = A$ and $Y = B$, to get

$$\mathbb{E}\min_{v\in S}\max_{u\in\mathbb{S}^{kd1}} A_{v,u} \leq \mathbb{E}\min_{v\in S}\max_{u\in\mathbb{S}^{d-1}} B_{v,u}.$$

Noting that

$$\mathbb{E}\min_{v\in S}\max_{u\in\mathbb{S}^{k-1}} B_{v,u} = \mathbb{E}\min_{v\in S}\max_{u\in\mathbb{S}^{k-1}} u^T G v = \mathbb{E}\min_{v\in S}\|Gv\|,$$

and

$$\begin{aligned}\mathbb{E}\left[\min_{v\in S}\max_{u\in\mathbb{S}^{k-1}} A_{v,u}\right] &= \mathbb{E}\max_{u\in\mathbb{S}^{k-1}} g^T u + \mathbb{E}\min_{v\in S} h^T v \\ &= \mathbb{E}\max_{u\in\mathbb{S}^{k-1}} g^T u - \mathbb{E}\max_{v\in S}(-h^T v) = a_k - \omega(S),\end{aligned}$$

gives the second part of the theorem.

On the other hand, since $\mathbb{E}\left|A_{v,u} - A_{v',u'}\right|^2 - \mathbb{E}\left|B_{v,u} - B_{v',u'}\right|^2 \geq 0$ then we can similarly use Theorem 9.14 with $X = B$ and $Y = A$, to get

$$\mathbb{E}\max_{v\in S}\max_{u\in\mathbb{S}^{d-1}} A_{v,u} \geq \mathbb{E}\max_{v\in S}\max_{u\in\mathbb{S}^{d-1}} B_{v,u}.$$

Noting that

$$\mathbb{E}\max_{v\in S}\max_{u\in\mathbb{S}^{d-1}} B_{v,u} = \mathbb{E}\max_{v\in S}\max_{u\in\mathbb{S}^{d-1}} u^T G v = \mathbb{E}\max_{v\in S}\|Gv\|,$$

and

$$\mathbb{E}\left[\max_{v\in S}\max_{u\in\mathbb{S}^{d-1}} A_{v,u}\right] = \mathbb{E}\max_{u\in\mathbb{S}^{d-1}} g^T u + \mathbb{E}\max_{v\in S} h^T v = a_d + \omega(S),$$

concludes the proof of the theorem.

□

9.3 Random projections and Compressed Sensing: Sparse vectors and Low-rank matrices

A remarkable application of Gordon's Theorem is that one can use it for abstract sets S such as the set of all *natural images* or the set of all plausible *user-product* ranking matrices. In these cases Gordon's Theorem suggests that a measurements corresponding just to a random projection may be enough to keep geometric properties of the data set in question, in particular, it may allow for reconstruction of the data point from just the projection. These phenomenon and the sensing savings that arrises from it is at the heart of Compressed Sensing and several recommendation system algorithms, among many other data processing techniques. Motivated by these two applications we will focus in this section on understanding which projections are expected to preserve the norm of sparse vectors and low-rank matrices. Both Compressed Sensing and Low-rank Matrix Modelling will be discussed in length in, respectively, Chapters 10 and **??**.

9.3.1 Gaussian width of s-sparse vectors

Let $x \in \mathbb{R}^p$ represent a signal (or image) that we wish to acquire via linear measurements $y_i = a_i^T x$, for $a_i \in \mathbb{R}^p$. In general, one would need p linear one-dimensional measurements to find x, one for each coordinate. The idea behind *Compressed Sensing* [31, 49] is that one may be able to significantly decrease the number of measurements needed if we know more about the structure of x, a prime example is when x is known to be sparse, i.e. to have few non-zero entries. Sparse signals arise in countless applications: for example, natural images tend to be sparse in the wavelet basis[6], while audio signals tend to be sparse in local Fourier-type expansions[7].

We will revisit sparse recovery and Compressed Sensing on Chapter 10. For now, we will see how Gordon's Theorem can suggest how many linear measurements are needed in order to reconstruct a sparse vector. An efficient way of representing the measurements is to use a matrix

$$A = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix},$$

and represent the linear measurements as

$$y = Ax.$$

In order to be able to reconstruct x from y we need at the very least that A is injective on sparse vectors. Let us assume that x is s-sparse, meaning that x has at most s non-zero entries (often written as $\|x\|_0 \leq s$, where $\|\cdot\|_0$ is called the 0-norm and counts the number of non-zero entries in a vector[8]). Furthermore, in order for reconstruction to be stable, one should ask not only A is injective in s-sparse vectors but actually that it is almost an isometry, meaning that the ℓ_2 distance between Ax_1 and Ax_2 should be comparable to the distances between x_1 and x_2, if they are s-sparse. Since the difference between two s-sparse vectors is a $2s$-sparse vector, we can alternatively ask for A to approximately keep the norm of $2s$ sparse vectors. Gordon's Theorem above suggests that we can take $A \in \mathbb{R}^{m \times p}$ to have i.i.d. Gaussian entries and to take $m \approx \omega^2(\mathcal{S}_{2s})$, where $\mathcal{S}_{2s} = \left\{x : x \in \mathbb{S}^{p-1}, \|x\|_0 \leq 2s\right\}$ is the set of $2s$ sparse vectors, and $\omega(\mathcal{S}_{2s})$ the Gaussian width of $\mathcal{S}_{2s}$.

Proposition 9.16. *If $s \leq p$, the Gaussian width $\omega(\mathcal{S}_s)$ of $\mathcal{S}_s$, the set of unit-norm vectors that are at most s sparse, satisfies*

[6] The approximate sparsity of natural images in the wavelet bases is leveraged in the JPEG2000 compression method.

[7] This approximate sparsity is utilized in MP3 audio compression.

[8] It is important to note that $\|\cdot\|_0$ is not actually a norm, as it does not necessarily rescale linearly with a rescaling of x.

$$\omega\left(\mathcal{S}_s\right)^2 \lesssim s \log\left(\frac{p}{s}\right).$$

Proof.

$$\omega\left(\mathcal{S}_s\right) = \mathbb{E} \max_{v \in \mathbb{S}^{p-1},\, \|v\|_0 \leq s} g^T v,$$

where $g \sim \mathcal{N}(0, I_{p\times p})$. We have

$$\omega\left(\mathcal{S}_s\right) = \mathbb{E} \max_{\Gamma \subset [p],\, |\Gamma| = s} \|g_\Gamma\|,$$

where g_Γ is the restriction of g to the set of indices Γ.

Given a set Γ, Theorem 6.23 yields

$$\mathbb{P}\left\{\|g_\Gamma\|^2 \geq s + 2\sqrt{s}\sqrt{t} + 2t\right\} \leq \exp(-t).$$

Union bounding over all $\Gamma \subset [p]$, $|\Gamma| = s$ gives

$$\mathbb{P}\left\{\max_{\Gamma \subset [p],\, |\Gamma| = s} \|g_\Gamma\|^2 \geq s + 2\sqrt{s}\sqrt{t} + 2t\right\} \leq \binom{p}{s} \exp(-t)$$

Taking u such that $t = su$, gives

$$\mathbb{P}\left\{\max_{\Gamma \subset [p],\, |\Gamma| = s} \|g_\Gamma\|^2 \geq s\left(1 + 2\sqrt{u} + 2u\right)\right\} \leq \exp\left[-su + s\log\left(e\frac{p}{s}\right)\right]. \tag{9.10}$$

Taking $u > \log\left(e\frac{p}{s}\right)$ it can be readily seen that the typical size of $\max_{\Gamma \subset [p],\, |\Gamma| = s} \|g_\Gamma\|^2$ is $\lesssim s\log\left(\frac{p}{s}\right)$. The proof can be finished by integrating (9.10) in order to get a bound of the expectation of $\sqrt{\max_{\Gamma \subset [p],\, |\Gamma| = s} \|g_\Gamma\|^2}$. □

This suggests that $\approx 2s\log\left(\frac{p}{2s}\right)$ measurements suffice to stably identify a $2s$-sparse vector. As we will see in Chapter 10, dedicated to Compressed Sensing, this number of measurement is also sufficient to guarantee that the signal in question can be recover with efficient algorithms.

9.3.2 Gaussian width of rank-r matrices

Another structured set of interest is the set of low rank matrices. Low-rank matrices appear in countless applications, a prime example being recommendation systems such as in the celebrated *Netflix Prize*. In this case the matrix in question is a matrix indexed by users of a service and items, such as movies. Given a user and an item, the corresponding entry of the matrix should correspond to the score that user would attribute to that item. This matrix is believed to be low-rank. The goal is then to estimate the score for user and item pairs that have not been rated yet from the ones that have, by exploiting the low-rank matrix structure. This is known as low-rank matrix completion [35, 37, 110].

In this short section, we will not address the problem of matrix completion but rather make a comment about the problem of low-rank matrix sensing, where instead of observing some of the entries of the matrix $X \in \mathbb{R}^{n_1 \times n_2}$ one has access to linear measurements of it, of the form $y_i = \mathrm{Tr}(A_i^T X)$, the problem of Matrix Completion will be addressed in Chapter **??**.

In order to understand the number of measurements needed for the measurement procedure to be a nearly isometry for rank r matrices, we can estimate the Gaussian width of the set of matrices $X \in \mathbb{R}^{n_1 \times n_2}$ whose rank is smaller or equal to $2r$, and use Gordon's Theorem.

Proposition 9.17.

$$\omega\left(\left\{X : X \in \mathbb{R}^{n_1 \times n_2}, \mathrm{rank}(X) \leq r, \|X\|_F = 1\right\}\right) \lesssim \sqrt{r(n_1+n_2)}.$$

Proof.

$$\omega\left(\left\{X : X \in \mathbb{R}^{n_1 \times n_2}, \mathrm{rank}(X) \leq r, \|X\|_F = 1\right\}\right) = \mathbb{E} \max_{\substack{\|X\|_F=1 \\ \mathrm{rank}(X)\leq r}} \mathrm{Tr}(GX).$$

Let $X = U\Sigma V^T$ be the SVD decomposition of X, then

$$\omega\left(\left\{X : X \in \mathbb{R}^{n_1 \times n_2}, \mathrm{rank}(X) \leq r, \|X\|_F = 1\right\}\right) = \mathbb{E} \max_{\substack{U^TU=V^TV=I_{r\times r} \\ \Sigma\in\mathbb{R}^{r\times r}\|\Sigma\|_F=1 \\ \Sigma \text{ is diagonal}}} \mathrm{Tr}(\Sigma\left(V^TGU\right)).$$

This implies that

$$\omega\left(\left\{X : X \in \mathbb{R}^{n_1 \times n_2}, \mathrm{rank}(X) \leq r, \|X\|_F = 1\right\}\right) \leq (\mathrm{Tr}\,\Sigma)\,(\mathbb{E}\|G\|) \lesssim \sqrt{r}\left(\sqrt{n_1}+\sqrt{n_1}\right),$$

where the last inequality follows from bounds on the largest eigenvalue of a Wishart matrix. □

This suggests that the number of measurements needed to identify an $n_1 \times n_2$ rank r matrix is on the order of $r\,(n_1+n_2)$, rather than the $n_1 n_2$ measurements that would be needed without a low-rank assumption. As we will see in Chapter **??**, these savings play an important role in Matrix Sensing, Matrix Completion, and many recommendation system algorithms.

10

Compressive Sensing and Sparsity

Most of us have noticed how saving an image in JPEG dramatically reduces the space it occupies in our hard drives (as opposed to file types that save the value of each pixel in the image). The idea behind these compression methods is to exploit known structure in the images; although our cameras will record the value (even three values in RGB) for each pixel, it is clear that most collections of pixel values will not correspond to pictures that we would expect to see. Natural images do not correspond to arbitrary arrays of pixel values, but have some specific structure to them. It is this special structure one aims to exploit by choosing a proper representation of the image. Indeed, natural images are known to be approximately sparse in certain bases (such as the wavelet bases) and this is the core idea behind JPEG (actually, JPEG2000; JPEG uses a different basis).

Let us think of $x \in \mathbb{C}^p$ as the signal corresponding to the image already represented in the basis in which it is sparse. The modeling assumption is that x is s-sparse, or $\|x\|_0 \leq s$, meaning that x has at most s non-zero components and, usually, $s \ll p$. The ℓ_0-norm[1] $\|x\|_0$ of a vector x is the number of non-zero entries of x. This means that when we take a picture, our camera makes p measurements (each corresponding to a pixel) but then, after an appropriate change of basis, it keeps only $s \ll p$ non-zero coefficients and drops the others. This seems a rather wasteful procedure and thus motivates the question: "If only a few degrees of freedom are kept after compression, why not in the first place measure in a more efficient way and take considerably less than p measurements?".

The question whether we can carry out data acquisition and compression simultaneously is at the heart of *Compressive Sensing* [31, 32, 33, 34, 49, 57]. It is particularly important in MRI imaging [89, 55], as fewer measurements potentially means shorter data acquisition time. Indeed, current MRI technology based on concepts from compressive sensing can reduce the time needed to collect the data by a factor of six or more [89], which has significant benefits

[1] We recall that the ℓ_0 norm is not actually a norm.

especially in pediatric MR imaging [137]. We recommend the book [57] as a great in-depth reference about compressive sensing.

In mathematical terms, the acquired measurements $y \in \mathbb{C}^m$ are connected to the signal of interest $x \in \mathbb{C}^p$, with $m \ll p$, via

$$\begin{bmatrix} \\ y \\ \\ \end{bmatrix} = \begin{bmatrix} \\ A \\ \\ \end{bmatrix} \begin{bmatrix} \\ x \\ \\ \end{bmatrix}. \tag{10.1}$$

The matrix $A \in \mathbb{C}^{m \times p}$ models the linear measurement (information) process. Classical linear algebra tells us that if $m < p$, then the linear system (10.1) is underdetermined and that there are infinitely many solutions (assuming that there exists at least one solution). In other words, without additional information, it is impossible to recover x from y in the case $m < p$.

In this chapter we assume that x is s-sparse with $s < m \ll p$. The goal is to recover x from this underdetermined system. We emphasize that we *do not know* the location of the non-zero coefficients of x a priori[2], otherwise the task would be trivial.

In the previous chapter we used Gordon's Theorem (Theorem 9.10) to show that when using random Gaussian measurements, on the order of $s \log\left(\frac{p}{s}\right)$ measurements suffice to have all considerably different s-sparse signals correspond to considerably different sets of measurements. This suggests that $m \approx s \log\left(\frac{p}{s}\right)$ may be enough to recover x. While Gordon's Theorem guarantees that this number of measurements will suffice for sparse vectors to be uniquely determined by these random measurements, it does not offer any insight into whether it is possible to recover the signal of interest in a numerically efficient manner. Remarkably, as we will see below, this is indeed possible.

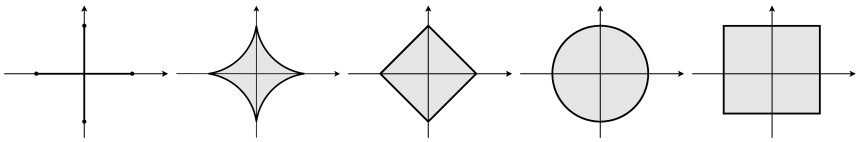
Since the system is underdetermined and we know that x is sparse, the natural approach to try to recover x is to solve

$$\begin{aligned} \min \quad & \|z\|_0 \\ \text{s.t.} \quad & Az = y, \end{aligned} \tag{10.2}$$

and hope that the optimal solution z corresponds to the signal in question x. However the optimization problem (10.2) is NP-hard in general [57]. Instead, the approach usually taken in sparse recovery is to consider a convex surrogate of the ℓ_0 norm, namely the ℓ_1 norm: $\|z\|_1 = \sum_{i=1}^{p} |z_i|$. Figure 10.1 depicts the ℓ_p balls and illustrates how the ℓ_1 norm can be seen as a convex surrogate of

[2] And therein lies the challenge, since s-sparse signals do not form a linear subspace of $\mathbb{R}^p$ (the sum of two s-sparse signals is in general no longer s-sparse but $2s$-sparse).

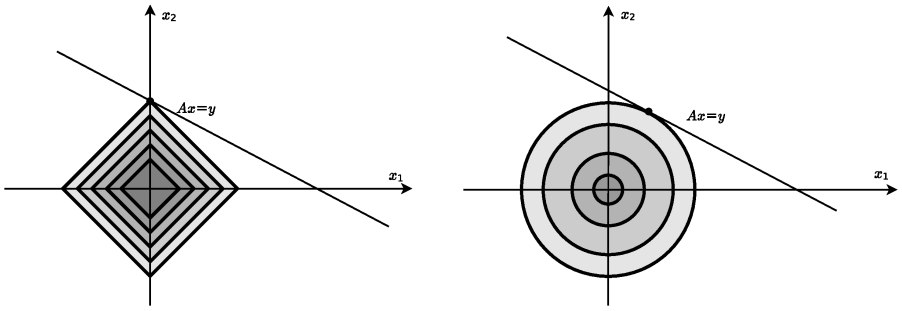
the ℓ_0 norm due to the *pointiness* of the ℓ_1 ball in the direction of the basis vectors, i.e. in "sparse" directions.



(a) $p = 0$ (b) $p = \frac{1}{2}$ (c) $p = 1$ (d) $p = 2$ (e) $p = \infty$

Fig. 10.1: ℓ_p norm unit balls with different values for p

The process of ℓ_p minimization can be understood as inflating the ℓ_p ball until one hits the affine subspace of interest. Figure 10.2 illustrates how ℓ_1 norm minimization promotes sparsity, while ℓ_2 norm minimization does not favor sparse solutions. We have seen in Chapter 2.2.2 that the ℓ_1 ball becomes "increasingly pointy" with increasing dimension. This behavior works in our favor in compressive sensing—another manifestation of the *blessings of dimensionality*.



(a) $p = 1$ (b) $p = 2$

Fig. 10.2: A two-dimensional depiction of ℓ_1 and ℓ_2 minimization. In ℓ_p minimization, one inflates the ℓ_p ball until it hits the affine subspace of interest. This image conveys how the ℓ_1 norm (left) promotes sparsity due to the "pointiness" of the ℓ_1 ball. In contract, ℓ_2 norm minimization (right) does not favor sparse solutions.

This motivates one to consider the following optimization problem (surrogate to (10.2)):

$$\begin{aligned} \min \quad & \|z\|_1 \\ \text{s.t.} \quad & Az = y, \end{aligned} \tag{10.3}$$

In order for (10.3) to be a good procedure for sparse recovery we need two things: for the solution of it to be meaningful (hopefully to coincide with x) and for (10.3) to be efficiently solved.

We will consider for the moment the real-valued case $x \in \mathbb{R}^p, A \in \mathbb{R}^{m \times p}$ and formulate (10.3) as a linear program[3] (and thus show that it is efficiently solvable). Let us think of ω^+ as the positive part of z and ω^- as the symmetric of the negative part of it, meaning that $z = \omega^+ - \omega^-$ and, for each i, either ω_i^- or ω_i^+ is zero. Note that, in that case,

$$\|z\|_1 = \sum_{i=1}^{p} \omega_i^+ + \omega_i^- = \mathbf{1}^T \left(\omega^+ + \omega^-\right).$$

Motivated by this, we consider:

$$\begin{array}{ll} \min & \mathbf{1}^T (\omega^+ + \omega^-) \\ \text{s.t.} & A(\omega^+ - \omega^-) = y \\ & \omega^+ \geq 0 \\ & \omega^- \geq 0, \end{array} \tag{10.4}$$

which is a linear program. It is not difficult to see that the optimal solution of (10.4) will indeed satisfy that, for each i, either ω_i^- or ω_i^+ is zero and it is indeed equivalent to (10.3); if both ω_i^- and ω_i^+ are non-zero, one can lower the objective while keep satisfying the constraints by reducing both variables. Since linear programs are efficiently solvable [136], this implies that the ℓ_1-optimization problem (10.3) is efficiently solvable.

In what follows we will discuss under which circumstances one can guarantee that the solution of (10.3) coincides with the sparse signal of interest. We will discuss a couple of different strategies to show this, as different strategies generalize better to other problems of interest. Later in this chapter we discuss strategies for constructing sensing matrices.

10.1 Null Space Property and Exact Recovery

Given a s-sparse vector x, our goal is to show that under certain conditions x is the unique optimal solution to

$$\begin{array}{ll} \min & \|z\|_1 \\ \text{s.t.} & Az = y, \end{array} \tag{10.5}$$

Let $S = \text{supp}(x)$, with $|S| = s$.[4] If x is not the unique optimal solution of the ℓ_1 minimization problem, there exists $z \neq x$ as optimal solution. Let $v = z - x$, it satisfies

[3] In the complex case, we are dealing with a quadratic program.

[4] If x has support size strictly smaller than s, for what follows, we can simply take a superset of it with size s

$$\|v + x\|_1 \leq \|x\|_1 \quad \text{and} \quad A(v + x) = Ax,$$

this means that $Av = 0$. Also,

$$\|x\|_S = \|x\|_1 \geq \|v + x\|_1 = \|\,(v + x)_S\,\|_1 + \|v_{S^c}\|_1 \geq \|x\|_S - \|v_S\|_1 + \|v\|_{S^c},$$

where the last inequality follows by the triangle inequality. This means that $\|v_S\|_1 \geq \|v_{S^c}\|_1$, but since $|S| \ll N$ it is unlikely for A to have vectors in its nullspace that are so concentrated on such few entries. This motivates the following definition.

Definition 10.1 (Null Space Property). *A is said to satisfy the s-Null Space Property ($A \in$ s-NSP) if, for all v in $\ker(A)$ (the nullspace of A) and all $|S| = s$, we have*

$$\|v_S\|_1 < \|v_{S^c}\|_1.$$

From the argument above, it is clear that if A satisfies the Null Space Property for s, then x will indeed be the unique optimal solution to (10.3). In fact, as the property is described in terms of any set S of size s, it implies recovery for any s-sparse vector.

Theorem 10.2. *Let x be an s-sparse vector. If $A \in$ s-NSP then x is the unique solution to the ℓ_1 optimization problem* (10.3) *with $y = Ax$.*

The Null Space Property is a statement about certain vectors not belonging to the null space of A, thus we can again resort to Gordon's Theorem (Theorem 9.10) to establish recovery guarantees for Gaussian sensing matrices. Let us define the intersection with the unit-sphere of the cone of such vectors

$$C_s := \left\{ v \in \mathbb{S}^{p-1} : \exists_{S \subset [p],\, |S| = s} \|v_S\|_1 \geq \|v_{S^c}\|_1 \right\}. \tag{10.6}$$

Since for a matrix A, $A \in$ s-NSP is equivalent to $\ker(A) \cap C_s = \emptyset$, Gordon's Theorem, or more specifically Gordon's Escape Through a Mesh Theorem (Theorem 9.13), implies that there exists a universal $C > 0$ such that if A is drawn with iid Gaussian entries, it will satisfy the s-NSP with high probability provided that $M \geq C\omega^2(C_s)$, where $\omega(C_s)$ is the Gaussian width of C_s.

Proposition 10.3. *Let $s \leq p$ and $C_s \subset \mathbb{S}^{p-1}$ defined in* (10.6)*, there exists a universal constant C such that*

$$\omega(C_s) \leq C\sqrt{s \log\left(\frac{p}{s}\right)},$$

where $\omega(C_s)$ is the Gaussian width of C_s.

Proof. The goal is to bound upper bound

$$\omega(C_s) = \mathbb{E} \max_{v \in C_s} v^T g,$$

for $g \sim \mathcal{N}(0, I)$. Note that C_s is invariant under permutations of the indices. Thus, the maximizer $v \in C_s$ will have its largest entries (in absolute value) in the coordinates g has it largest entries (in absolute value). Let S be the set of the s coordinates with largest absolute value of g. We have

$$\mathbb{E} \max_{v \in C_s} v^T g = \mathbb{E} \max_{v : \|v_S\|_1 \geq \|v_{S^c}\|_1, \, \|v\|_2 = 1} v_S^T g_S + v_{S^c}^T g_{S^c}.$$

The key idea is to notice that the condition $\|v_S\|_1 \geq \|v_{S^c}\|_1$ imposes a strong bound on the ℓ_1 norm of v_{S^c} via $\|v_{S^c}\|_1 \leq \|v_S\|_1 \leq \sqrt{s} \|v_S\|_2 \leq \sqrt{s}$. This can be leveraged by noticing that

$$v_S^T g_S + v_{S^c}^T g_{S^c} \leq \|v_S\|_2 \|g_S\|_2 + \|v_{S^c}\|_1 \|g_{S^c}\|_\infty .$$

This gives

$$\omega(C_s) \leq \mathbb{E} \|g_S\|_2 + \sqrt{s} \|g_{S^c}\|_\infty ,$$

where S corresponds to the set of the s coordinates with largest absolute value of g.

We saw in the proof of Proposition 9.16, in the context of computing the Gaussian width of the set of sparse vectors, that $\mathbb{E} \|g_S\|_2 \lesssim \sqrt{s \log\left(\frac{p}{s}\right)}$. Since all entries of g_{S^c} are smaller, in absolute value, that any entry in g_S we have that $\|g_{S^c}\|_\infty^2 \leq \frac{1}{s} \|g_S\|_2^2$. This implies that $\mathbb{E} \|g_{S^c}\|_\infty \lesssim \sqrt{\log\left(\frac{p}{s}\right)}$, concluding the proof. □

Together with Theorem 10.2 this implies the following recovery guarantee, matching the order of number of measurements suggested by the Gaussian width of sparse vectors.

Theorem 10.4. *There exists a universal constant $C \geq 0$ such that if A is a $m \times p$ matrix with iid Gaussian entries the following holds with high probability: For any x an s-sparse vector, x is the unique solution to the ℓ_1-optimization problem* (10.3) *with $y = Ax$.*

Remark 10.5. If one is interested in understanding the probability of exact recovery of a specific sparse vector, and not a uniform guarantee on all sparse vectors simultaneously, then it is possible to do a more refined version of the arguments above that are able to predict the exact asymptotics of the number of measurements required; see [40] for an approach based on Gaussian widths and [10] for an approach based on Integral Geometry [10].

10.1.1 The Restricted Isometry Property

A classical approach to establishing exact recovery via ℓ_1-minimization is through the Restricted Isometry Property (RIP), which corresponds precisely with the property of approximately preserving the length of sparse vectors.

Definition 10.6 (Restricted Isometry Property (RIP)). *An $m \times p$ matrix A (with either real or complex valued entries) is said to satisfy the (s, δ)-Restricted Isometry Property (RIP),*

$$(1-\delta)\|x\|^2 \leq \|Ax\|^2 \leq (1+\delta)\|x\|^2,$$

for all s-sparse x.

If A satisfies the RIP for sparsity $2s$, it means that it approximately preserves distances between s-sparse vectors (hence the name *RIP*). This can be leveraged to show that A satisfies the NSP.

Theorem 10.7 ([38]). *Let $y = Ax$ where x is an s-sparse vector. Assume that A satisfies the RIP property with $\delta_{2s} < \frac{1}{3}$, then the solution x_* to the ℓ_1-minimization problem*

$$\min_z \|z\|_1, \qquad \text{subject to } Az = y = Ax$$

becomes x exactly, i.e., $x_* = x$

To prove this theorem we need the following lemma.

Lemma 10.8 ([38]). *We have*

$$|\langle Ax, Ax' \rangle| \leq \delta_{s+s'} \|x\|_2 \|x'\|_2$$

for all x, x' supported on disjoint subsets $S, S' \subseteq [1, \cdots, p]$, $x, x' \in \mathbb{R}^p$, and $|S| \leq s$, $|S'| \leq s'$

Proof. Without loss of generality, we can assume $\|x\|_2 = \|x'\|_2 = 1$, so that the right hand size of the inequality becomes just $\delta_{s+s'}$. Since A satisfies the RIP property, we have

$$(1-\delta_{s+s'})\|x \pm x'\|_2^2 \leq \|A(x \pm x')\|_2^2 \leq (1+\delta_{s+s'})\|x \pm x'\|_2^2.$$

Since x and x' have disjoint support, $\|x \pm x'\|_2^2 = \|x\|_2^2 + \|x\|_2^2 = 2$; the RIP property then becomes

$$2(1-\delta_{s+s'}) \leq \|Ax \pm Ax'\|_2^2. \leq 2(1+\delta_{s+s'})$$

The polarization identity implies:

$$\begin{aligned} |\langle Ax, Ax' \rangle| &= \frac{1}{4}\Big| \|Ax + Ax'\|_2^2 - \|Ax - Ax'\|_2^2 \Big| \\ &\leq \frac{1}{4}\Big| 2(1+\delta_{s+s'}) - 2(1-\delta_{s+s'}) \Big| \\ &= \delta_{s+s'}. \end{aligned}$$

To prove Theorem 10.7, we simply need to show that the Null Space Property holds for the given conditions.

Proof (of Theorem 10.7). Take $h \in (A) \setminus 0$. Let index set S_0 be the set of indices of s largest entries (by modulus) of h. Let index sets $S_1, S_2, \cdots$ be index sets corresponding to the next s to $2s$, $2s$ to $3s$, $\cdots$ largest entries of h.

Since A satisfies the RIP, we have

$$\|h_{S_0}\|_2^2 \le \frac{1}{1-\delta_s}\|Ah_{S_0}\|_2 \tag{10.7}$$

$$= \frac{1}{1-\delta_s}\sum_{j\ge 1}\langle Ah_{S_0}, A(-h_{S_j})\rangle \quad (\text{because } h_{S_0} = \sum_{j\ge 1}(-h_{S_j})) \tag{10.8}$$

$$\le \frac{1}{1-\delta_s}\sum_{j\ge 1}\delta_{2s}\|h_{S_0}\|_2\|h_{S_j}\|_2 \qquad (\text{by Lemma 10.8}) \tag{10.9}$$

$$\le \frac{\delta_{2s}}{1-\delta_s}\|h_{S_0}\|_2\sum_{j\ge 1}\|h_{S_j}\|_2 \tag{10.10}$$

$$\|h_{S_0}\|_2 \le \frac{\delta_{2s}}{1-\delta_s}\sum_{j\ge 1}\|h_{S_j}\|_2. \tag{10.11}$$

Note that

$$\|h_{S_j}\|_2 \le s^{\frac{1}{2}}\|h_{S_j}\|_\infty \le s^{-\frac{1}{2}}\|h_{S_{j-1}}\|_1.$$

We can rewrite (10.11) as

$$\|h_{S_0}\|_2 \le \frac{\delta_{2s}}{1-\delta_s}s^{-\frac{1}{2}}\sum_{j\ge 1}\|h_{S_{j-1}}\|_1 \tag{10.12}$$

$$= \frac{\delta_{2s}}{1-\delta_s}s^{-\frac{1}{2}}\|h\|_1. \tag{10.13}$$

Also, by the Cauchy-Schwarz inequality,

$$\|h_{S_0}\|_1 = \sum_{i\in S_0} 1\times|h_i| \le \sqrt{\sum_{i\in S_0}1^2}\sqrt{\sum_{i\in S_0}h_i^2} = \sqrt{s}\|h_{S_0}\|_2. \tag{10.14}$$

We have $\delta_{2s} < \frac{1}{3}$ as a condition, so

$$\frac{\delta_{2s}}{1-\delta_s} < \frac{\delta_{2s}}{1-\delta_{2s}} < \frac{1}{2} \quad \text{for } \delta_{2s} < \frac{1}{3}. \tag{10.15}$$

Combining (10.13), (10.14), and (10.15), we get

$$\|h_{S_0}\|_1 < \frac{1}{2}\|h\|_1. \tag{10.16}$$

Now we show that (10.16) is equivalent to $\|h_S\|_1 < \|h_{S^C}\|_1$:

$$\begin{aligned} & \|h_S\|_1 < \|h_{S^C}\|_1 \\ \Leftrightarrow \quad & 2\|h_S\|_1 < \|h_{S^C}\|_1 + \|h_S\|_1 \\ \Leftrightarrow \quad & 2\|h_S\|_1 < \|h\|_1 \\ \Leftrightarrow \quad & \|h_S\|_1 < \frac{1}{2}\|h\|_1. \end{aligned}$$

Thus, we have shown that $\|h_{S_0}\|_1 < \|h_{S^C}\|_1$, which is the Null Space Property and by virtue of Theorem 10.2 our proof is complete. □

Many results in compressive sensing (such as Theorem 10.7) can be extended will little extra effort to the case where x is not exactly s-sparse, but only approximately s-sparse, a property that is sometimes referred to as *compressible*. See [36, 57] for a detailed discussion.

Which matrices do satisfy the RIP under favorable conditions? Clearly, we want the number of measurements necessary to recover a sparse vector with ℓ_1-minimization to be as small as possible.

In Chapter 9 we computed the number of rows needed for a Gaussian matrix to approximately preserve the norm of sparse vectors, via estimates of the Gaussian width of the set of sparse vectors. In fact, using Proposition 9.16 and Theorem 9.11, one can readily show[5] that matrices with Gaussian entries satisfy the RIP with $m \approx s \log\left(\frac{p}{s}\right)$.

Theorem 10.9. *Let A be an $m \times p$ matrix with i.i.d. standard Gaussian entries, there exists a constant C such that, if*

$$m \geq Cs \log\left(\frac{p}{s}\right),$$

then $\frac{1}{\sqrt{m}}A$ satisfies the $\left(s, \frac{1}{3}\right)$-RIP with high probability.

We point out an important aspect in this context. Theorem 10.9 combined with Theorem 10.7 yields a *uniform recovery guarantee* for sparse vectors with Gaussian sensing matrices. Once a Gaussian matrix satisfies the RIP (which it will for certain parameters with high probability), then exact recovery via ℓ_1-minimization holds uniformly for *all* sufficiently sparse vectors.

While there are obvious similarities between Johnson-Lindenstrauss projections and sensing matrices that satisfy the RIP, there are also important differences. We note that for JL dimension reduction to be applicable (an upper estimate of) the number of vectors must be known a priori (and this number if finite). JL projection preserves (up to ε) pairwise distances between these vectors, but the vectors do not have to be sparse. As a consequence, JL projections P are a one-way street, as in general one cannot recover x from $y = Px$. In contrast, a matrix that satisfies the RIP works for infinitely many vectors, however with the caveat that these vectors must be sparse. Moreover, one can recover such sparse vectors x from $y = Ax$ (and can do so numerically efficiently).

As a consequence of these considerations, a matrix that satisfies the RIP does not necessarily have to satisfy the Johnson-Lindenstrauss Lemma. While a Gaussian random matrix does indeed satisfy both, RIP and the Johnson-Lindenstrauss Lemma, other matrices do not satisfy both simultaneously. For

[5] Note that the $1 \pm \delta$ term in the RIP property corresponds to $(1 \pm \varepsilon)^2$ in Gordon's Theorem. Since the RIP is a stronger property when δ is smaller, one can simply use $\varepsilon = \frac{1}{3}\delta$.

example, take a randomly subsampled Fourier matrix A of dimensions $m \times p$. In the notation of the definition of the Fast Johnson-Lindenstrauss transform, this matrix A would correspond to $A = SF$, but without the diagonal matrix D that randomizes phases (or signs) of x. This matrix A will not meet the Johnson-Lindenstrauss properties of Theorem 9.1. But the absence of the phase randomization matrix D is not a hurdle for $A = SF$ to satisfy the RIP.

It is known [34] that if $m = \Omega_\delta(s \operatorname{polylog} p)$, then the partial Fourier matrix satisfies the RIP with high probability. The exact number of logarithmic factors needed is the object of much research with the best known upper bound due to Haviv and Regev [67], giving an upper bound of $m = \Omega_\delta(s \log^2 s \log p)$. On the side of lower bounds it is know that the asymptotics established for Gaussian matrices of $m = \Theta_\delta(s \log(p/s))$ are not achievable in general [22].

Checking whether a matrix satisfies the RIP or not is in general NP-hard [19, 126]. While Theorem 10.9 suggests that RIP matrices are abundant for $s \approx \frac{m}{\log(p)}$, it appears to be very difficult to deterministically construct matrices that satisfy RIP for $s \gg \sqrt{m}$, known as the square bottleneck [122, 21, 20, 23, 30, 94]. The only known unconditional construction that is able to break this bottleneck is due to Bourgain et al. [30]; their construction achieves $s \approx m^{\frac{1}{2}+\varepsilon}$ for a small, but positive, ε. There is also a conditional construction, based on the Paley Equiangular Tight Frame [21, 23].

In Section 10.3 we will consider more practical conditions for designing sensing matrices. These conditions, which are better suited for applications, are based on the concept of the *coherence* of a matrix. Interestingly, the phase randomization of x that is notably absent in the partial Fourier matrix mentioned above, will reappear in this context in connection with *nonuniform recovery guarantees*.

10.2 Duality and exact recovery

In this section we describe yet another approach to show exact recovery of sparse vectors via (10.3). In this section we take an approach based on duality, the same strategy we took on Chapter 8 to show exact recovery in the Stochastic Block Model. The approach presented here is essentially the same as the one followed in [58] for the real case, and in [128] for the complex case.

Let us start by presenting duality in Linear Programming with a game theoretic view point, similarly to how we did for Semidefinite Programming in Chapter 8. The idea is again to to reformulate (10.4) without constraints, by adding a dual player that wants to maximize the objective and would exploit a deviation from the original constraints (by, for example, giving the dual player a variable u and adding to to the objective $u^T (y - A(\omega^+ - \omega^-))$). More precisely consider the following

$$\min_{\substack{\omega^+\\ \omega^-}} \max_{\substack{u\\ v^+\geq 0\\ v^-\geq 0}} \mathbf{1}^T\left(\omega^+ + \omega^-\right) - \left(v^+\right)^T\omega^+ - \left(v^-\right)^T\omega^- + u^T\left(y - A\left(\omega^+ - \omega^-\right)\right). \tag{10.17}$$

Indeed, if the primal player (picking ω^+ and ω^- and attempting to minimize the objective) picks variables that do not satisfy the original constraints, then the dual player (picking u, v^+, and v^- and trying to maximize the objective) will be able to make the objective value as large as possible. It is then clear that (10.4) = (10.17).

If the order with which the players choose variable values, this can only benefit the primal player, that now gets to see the value of the dual variables before picking the primal variables, meaning that (10.17) $\geq$ (10.18), where (10.18) is given by:

$$\max_{\substack{u\\ v^+\geq 0\\ v^-\geq 0}} \min_{\substack{\omega^+\\ \omega^-}} \mathbf{1}^T\left(\omega^+ + \omega^-\right) - \left(v^+\right)^T\omega^+ - \left(v^-\right)^T\omega^- + u^T\left(y - A\left(\omega^+ - \omega^-\right)\right). \tag{10.18}$$

Rewriting

$$\max_{\substack{u\\ v^+\geq 0\\ v^-\geq 0}} \min_{\substack{\omega^+\\ \omega^-}} \left(\mathbf{1} - v^+ - A^Tu\right)^T\omega^+ + \left(\mathbf{1} - v^- + A^Tu\right)^T\omega^- + u^Ty \tag{10.19}$$

With this formulation, it becomes clear that the dual players needs to set $\mathbf{1} - v^+ - A^Tu = 0$, $\mathbf{1} - v^- + A^Tu = 0$ and thus (10.19) is equivalent to

$$\max_{\substack{u\\ v^+\geq 0\\ v^-\geq 0\\ \mathbf{1}-v^+-A^Tu=0\\ \mathbf{1}-v^-+A^Tu=0}} u^Ty$$

or equivalently,

$$\begin{aligned} &\max_u\ u^Ty\\ &s.t.\ \ -\mathbf{1} \leq A^Tu \leq \mathbf{1}. \end{aligned} \tag{10.20}$$

The linear program (10.20) is known as the dual program to (10.4). The discussion above shows that (10.20) $\leq$ (10.4) which is known as weak duality. More remarkably, strong duality guarantees that the optimal values of the two programs match.

There is a considerably easier way to show weak duality (although not as enlightening as the one above). If ω^- and ω^+ are primal feasible and u is dual feasible, then

$$\begin{aligned} 0 &\leq \left(\mathbf{1}^T - u^TA\right)\omega^+ + \left(\mathbf{1}^T + u^TA\right)\omega^- \\ &= 1^T\left(\omega^+ + \omega^-\right) - u^T\left[A\left(\omega^+ - \omega^-\right)\right] = 1^T\left(\omega^+ + \omega^-\right) - u^Ty, \end{aligned} \tag{10.21}$$

showing that (10.20) $\leq$ (10.4).

10.2.1 Finding a dual certificate

In order to show that $\omega^+ - \omega^- = x$ is an optimal solution[6] to (10.4), we will find a dual feasible point u for which the dual matches the value of $\omega^+ - \omega^- = x$ in the primal, u is known as a *dual certificate* or *dual witness*.

From (10.21) it is clear that u must satisfy $\left(\mathbf{1}^T - u^T A\right)\omega^+ = 0$ and $\left(\mathbf{1}^T + u^T A\right)\omega^- = 0$, this is known as *complementary slackness*. This means that we must take the entries of $A^T u$ be $+1$ or -1 when x is non-zero (and be $+1$ when it is positive and -1 when it is negative), in other words

$$\left(A^T u\right)_S = \operatorname{sign}\left(x_S\right),$$

where $S = \operatorname{supp}(x)$, and $\left\|A^T u\right\|_\infty \leq 1$ (in order to be dual feasible).

Remark 10.10. It is not difficult to see that if we further ask that $\left\|\left(A^T u\right)_{S^c}\right\|_\infty < 1$ any optimal primal solution would have to have its support contained in the support of x. This observation gives us the following Lemma.

Lemma 10.11. *Consider the problem of sparse recovery discussed above. Let $S = \operatorname{supp}(x)$, if A_S is injective and there exists $u \in \mathbb{R}^M$ such that*

$$\left(A^T u\right)_S = \operatorname{sign}\left(x_S\right),$$

and

$$\left\|\left(A^T u\right)_{S^c}\right\|_\infty < 1,$$

then x is the unique optimal solution to the ℓ_1-minimization problem (10.3).

Since we know that $\left(A^T u\right)_S = \operatorname{sign}\left(x_S\right)$ (and that A_S is injective), we try to construct[7] u by least squares and hope that it satisfies $\left\|\left(A^T u\right)_{S^c}\right\|_\infty < 1$. More precisely, we take

$$u = \left(A_S^T\right)^\dagger \operatorname{sign}\left(x_S\right),$$

where $\left(A_S^T\right)^\dagger = A_S\left(A_S^T A_S\right)^{-1}$ is the Moore Penrose pseudo-inverse of A_S^T. This gives the following Corollary.

Corollary 10.12. *Consider the problem of sparse recovery discussed this lecture. Let $S = \operatorname{supp}(x)$, if A_S is injective and*

$$\left\|\left(A_{S^c}^T A_S\left(A_S^T A_S\right)^{-1}\operatorname{sign}\left(x_S\right)\right)_{S^c}\right\|_\infty < 1,$$

then x is the unique optimal solution to the ℓ_1-minimization problem (10.3).

[6] For now we will focus on showing that it is an optimal solution, see Remark 10.10 for a brief discussion of how to strengthen the argument to show uniqueness.

[7] Note how this differs from the situation in Chapter 8 where the linear inequalities were enough to determine a unique candidate for a dual certificate.

Theorem 10.9 establishes that if $m \gg s \log\left(\frac{p}{s}\right)$ and $A \in \mathbb{R}^{m \times p}$ is drawn with i.i.d. Gaussian entries $\mathcal{N}\left(0, \frac{1}{m}\right)$ then[8] it will, with high probability, satisfy the $(s, 1/3)$-RIP. Note that, if A satisfies the $(s, 1/3)$-RIP then, for any $|S| \leq s$ one has $\|A_S\| \leq \sqrt{1 + \frac{1}{3}}$ and $l\left(A_S^T A_S\right)^{-1} \| \leq \left(1 - \frac{1}{3}\right)^{-1} = \frac{3}{2}$, where $\|\cdot\|$ denotes the operator norm $\|B\| = \max_{\|x\|=1} \|Bx\|$.

This means that if we take A random with i.i.d. $\mathcal{N}\left(0, \frac{1}{m}\right)$ entries then, for any $|S| \leq s$ we have that

$$\|A_S \left(A_S^T A_S\right)^{-1} \operatorname{sign}(x_S) \| \leq \sqrt{1 + \frac{1}{3}} \frac{3}{2} \sqrt{s} = \sqrt{3}\sqrt{s},$$

and because of the independency among the entries of A, A_{S^c} is independent of this vector and so for each $j \in S^c$ we have

$$\mathbb{P}\left(\left|A_j^T A_S \left(A_S^T A_S\right)^{-1} \operatorname{sign}(x_S)\right| \geq \frac{1}{\sqrt{M}} \sqrt{3}\sqrt{s}t\right) \leq 2 \exp\left(-\frac{t^2}{2}\right),$$

where A_j is the j-th column of A.

An application of the union bound gives

$$\mathbb{P}\left(\left\|A_{S^c}^T A_S \left(A_S^T A_S\right)^{-1} \operatorname{sign}(x_S)\right\|_\infty \geq \frac{1}{\sqrt{M}} \sqrt{3}\sqrt{s}t\right) \leq 2N \exp\left(-\frac{t^2}{2}\right),$$

which implies

$$\begin{aligned}\mathbb{P}\left(\left\|A_{S^c}^T A_S \left(A_S^T A_S\right)^{-1} \operatorname{sign}(x_S)\right\|_\infty \geq 1\right) &\leq 2p \exp\left(-\frac{\left(\frac{\sqrt{m}}{\sqrt{3s}}\right)^2}{2}\right) \\ &= \exp\left(-\frac{1}{2}\left[\frac{m}{3s} - 2\log(2p)\right]\right),\end{aligned}$$

which means that we expect to exactly recover x via ℓ_1 minimization when $m \gg s \log(p)$. While this can be asymptotically worse then the bound of $m \gtrsim s \log\left(\frac{p}{s}\right)$, and this guarantee is not uniformly obtained for all sparse vectors, the technique in this section is generalizable to many circumstances and illustrates the flexibility of approaches based in construction of dual witnesses.

10.3 Sensing matrices and incoherence

In applications, we usually cannot completely freely choose the sensing matrix to our liking. This means that Gaussian random matrices play an important role as benchmark, but from a practical viewpoint they play a marginal role.

[8] Note that the normalization here is taken slightly differently: entries are normalized by $\frac{1}{\sqrt{m}}$, rather than $\frac{1}{a_m}$, but the difference is negligible for our purposes.

Clearly, randomness in the sensing matrix seems to be very beneficial for compressive sensing. However, in practice, there are many design constraints on the sensing matrix A, as in many applications one only has access to structured measurement systems. For example, we may still have the freedom to choose, say the positions of the antennas in radar systems that employ multiple antennas or the position of sensors in MRI. By choosing these randomly, we can still introduce randomness in our system. Or, we can transmit random waveforms in sonar and radar systems. Yet, in all these cases the overall structure of A is still dictated by the physics of wave propagation. In other applications, it will be other physical constraints or design limitations that will limit how much randomness we can introduce into sensing matrix.

While establishing the RIP for Gaussian or Bernoulli random matrices is not too difficult, it is already significantly harder to do so for the partial Fourier matrix [34, **?**, 67], and time-frequency matrices [52], and even harder for more specific sensing matrices.

A useful concept to overcome the practical limitations of the RIP is via the concept of the (in)coherence of a matrix. This concept has proven to be widely applicable in practice. While we want to avoid the constraints of the RIP, we nevertheless take it as our point of departure. Recall that the RIP (Definition 10.6) asks that any $S \subset [p]$, $|S| \leq s$ satisfies:

$$(1-\delta)\|x\|^2 \leq \|A_S x\|^2 \leq (1+\delta)\|x\|^2,$$

for all $x \in \mathbb{R}^{|S|}$. This is equivalent to

$$\max_x \frac{x^T \left(A_S^T A_S - I\right) x}{x^T x} \leq \delta,$$

or equivalently

$$\left\|A_S^T A_S - I\right\| \leq \delta.$$

If the columns of A are unit-norm vectors (in $\mathbb{R}^m$), then the diagonal of $A_S^T A_S$ is all-ones, this means that $A_S^T A_S - I$ consists only of the non-diagonal elements of $A_S^T A_S$. If, moreover, for any two columns a_i, a_j, of A we have $\left|a_i^T a_j\right| \leq \mu$ for some μ then, Gershgorin's circle theorem tells us that $\left\|A_S^T A_S - I\right\| \leq \mu(s-1)$.

More precisely, given a symmetric matrix B, the Gershgorin's Circle Theorem [68] states that all of the eigenvalues of B are contained in the so called Gershgorin discs (for each i, the Gershgorin disc corresponds to $\left\{\lambda : |\lambda - B_{ii}| \leq \sum_{j\neq i} |B_{ij}|\right\}$. If B has zero diagonal, then this reads: $\|B\| \leq \sum_{j\neq i} |B_{ij}|$.

Given a set of p unit-norm vectors $a_1, \ldots, a_p \in \mathbb{R}^m$ we define its worst-case coherence μ as

$$\mu = \max_{i\neq j} \left|a_i^T a_j\right|. \tag{10.22}$$

Given a set of unit-norm vectors $a_1, \ldots, a_p \in \mathbb{R}^m$ with worst-case coherence μ, if we form a matrix with these vectors as columns, then it will be $(s, \mu(s-1))$-RIP, meaning that it will be $\left(s, \frac{1}{3}\right)$- RIP for $s \leq \frac{1}{3}\frac{1}{\mu}$.

This motivates the problem of designing sets of vectors $a_1, \ldots, a_p \in \mathbb{R}^m$ with smallest possible worst-case coherence. This is a central problem in Frame Theory [120, 39]. The smallest coherence of a set of p unit-norm vectors in m dimensions is bounded below by the Welch bound (see for example, [120, 21] for a discussion) which reads:

$$\mu \geq \sqrt{\frac{p-m}{m(p-1)}}.$$

Due to this limitation, deterministic constructions based on coherence cannot yield matrices that satisfy the RIP for $s \gg \sqrt{m}$, known as the square-root bottleneck [21, 122].

There are constructions that achieve the Welch bound, known as Equiangular Tight Frames (ETFs), these are sets of vectors (frames) for which all inner products between pairs of vectors have the same modulus $\mu = \sqrt{\frac{p-m}{m(p-1)}}$, meaning that they are "equiangular", see [120]. It is known that for an ETF to exist an ETF in $\mathbb{C}^m$ one needs $p \leq m^2$. For which dimensions m this bound is actually saturated is an important question in Quantum Mechanics and intimately connected to the famous Zauner's Conjecture [139, 115, 12].

To overcome this square root bottleneck something has to give. One fruitful direction is to sacrifice the *uniform recovery* granted by the RIP. Namely, once a matrix satisfies the RIP, the ℓ_0-ℓ_1 equivalence is guaranteed to hold for all s-sparse vectors. In contrast we can consider scenarios in which we are guaranteed the ℓ_0-ℓ_1 equivalence "only" for most s-sparse vectors. This leads to *nonuniform recovery* results, which we will pursue below. The benefits are worth the sacrifice, since we end up with theoretical guarantees that are much more practical.

Recall that we consider a general linear system of equations $Ax = y$, where $A \in \mathbb{C}^{m \times p}$, $x \in \mathbb{C}^p$ and $m \leq p$. We introduce the following generic s-sparse model:

(i) The support $I \subset \{1, \ldots, m\}$ of the s nonzero coefficients of x is selected uniformly at random.
(ii) The non-zero entries of $\operatorname{sign}(x)$ form a Steinhaus sequence, i.e., $\operatorname{sign}(x_k) := x_k/|x_k|, k \in I_s$, is a complex random variable that is uniformly distributed on the unit circle.

To make our result even more practical, we will consider noisy measurements. A standard approach to find a sparse (and under appropriate conditions *the sparsest*) solution to a noisy system $y = Ax + w$ is via

$$\min_x \frac{1}{2}\|Ax - y\|_2^2 + \lambda \|x\|_1, \tag{10.23}$$

which is also known as lasso [125]. Here $\lambda > 0$ is a regularization parameter.

We will consider the following two-step version of lasso as it often gives improved performance. In the first step we compute an estimate $\tilde{I}$ for the support of x by solving (10.23). In the second step we estimate the amplitudes of x by solving the reduced-size least squares problem $\min \|A_{\tilde{I}} x_{\tilde{I}} - y\|_2$, where $A_{\tilde{I}}$ is the submatrix of A consisting of the columns corresponding to the index set $\tilde{I}$, and similarly for $x_{\tilde{I}}$. This is a standard way to "debias" the solution, and we thus will call this approach *debiased lasso*.

As an example for a theoretical performance guarantee of this debiased lasso we state (without proof) the following theorem, which is a slightly extended version of Theorem 1.3 in [36].

Theorem 10.13. *Given $y = Ax + w$, where $A \in \mathbb{C}^{m \times p}$ has all unit-ℓ_2-norm columns, $x \in \mathbb{C}^p$ is drawn from the generic s-sparse model and $w_i \sim \mathcal{CN}(0, \sigma^2)$. Assume that*

$$\mu(A) \leq \frac{C_0}{\log p}, \tag{10.24}$$

where $C_0 > 0$ is a constant independent of m, p. Furthermore, suppose

$$s \leq \frac{c_0 p}{\|A\|_{op}^2 \log p} \tag{10.25}$$

for some constant $c_0 > 0$ and that

$$\min_{k \in I_s} |x_k| > 8\sigma\sqrt{2 \log p}. \tag{10.26}$$

Then the solution $\hat{x}$ to the debiased lasso computed with $\lambda = 2\sigma\sqrt{2 \log p}$ obeys

$$\operatorname{supp}(\hat{x}) = \operatorname{supp}(x), \tag{10.27}$$

and

$$\frac{\|\hat{x} - x\|_2}{\|x\|_2} \leq \frac{\sigma\sqrt{3n}}{\|y\|_2} \tag{10.28}$$

with probability at least

$$1 - 2p^{-1}(2\pi \log p + sp^{-1}) - \mathcal{O}(p^{-2 \log 2}). \tag{10.29}$$

Various other versions of *nonuniform recovery* results can be found e.g., in [129, 36, 57]. See [119, 70] for some theoretical results geared towards applications.

How does Theorem 10.13 compare to RIP based conditions in terms of required number of measurements? Assume that the columns of A form a unit-norm tight frame. In this case it is easy to see that $\|A\|_{\text{op}}^2 = \frac{p}{m}$ and condition (10.25) becomes $m \gtrsim s \log p$. We emphasize that the condition on the

coherence (10.24) is rather mild. For example an $m \times p$ Gaussian random matrix would satisfy it as long as the number of its columns is not exponentially larger than the number of its rows. But the point of the coherence condition is of course not to apply to Gaussian random matrices, but to structured random sensing matrices, see also [109].

There are various other efficient and rigorous methods to recover sparse vectors from underdetermined systems besides ℓ_1-minimization. For example, homotopy methods, greedy algorithms or methods based on approximate message passing. We refer to [57] for a comprehensive discussion of these techniques. Moreover, practice has shown that some adaptation of the random sampling pattern is highly desirable to improve performance, see e.g. [89, 5]. Furthermore, we refer to [5] for a thorough discussion of some subtle potential numerical stability issues one should be aware of.

References

1. E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):417–487, 2016. (Cited on pp. 112, 120.)
2. E. Abbe, J. Fan, and Y. Wang, K.and Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Available online at arXiv:1709.09565*, 2017. (Cited on p. 121.)
3. E. Abbe and C. Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *Available online at arXiv:1512.09080 [math.PR]*, 2015. (Cited on p. 112.)
4. Emmanuel Abbe. Community detection and stochastic block models. *Foundations and Trends® in Communications and Information Theory*, 14(1-2):1–162, 2018. (Cited on p. 111.)
5. Ben Adcock and Anders Hansen. *Compressive Imaging*. 2020. (Cited on p. 155.)
6. Nir Ailon and Bernard Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput*, pages 302–322, 2009. (Cited on p. 126.)
7. N. Alon. Eigenvalues and expanders. *Combinatorica*, 6:83–96, 1986. (Cited on p. 59.)
8. N. Alon. Problems and results in extremal combinatorics i. *Discrete Mathematics*, 273(1–3):31–53, 2003. (Cited on p. 126.)
9. N. Alon and V. Milman. Isoperimetric inequalities for graphs, and superconcentrators. *Journal of Combinatorial Theory*, 38:73–88, 1985. (Cited on p. 59.)
10. D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference, available online*, 2014. (Cited on p. 144.)
11. G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*. Cambridge studies in advanced mathematics. Cambridge University Press, Cambridge, New York, Melbourne, 2010. (Cited on p. 38.)
12. Marcus Appleby, Ingemar Bengtsson, Steven Flammia, and Dardo Goyeneche. Tight frames, Hadamard matrices and Zauner's conjecture. *Journal of Physics A: Mathematical and Theoretical*, 52(29):295301, 2019. (Cited on p. 153.)

13. S. Arora, B. Barak, and D. Steurer. Subexponential algorithms for unique games related problems. 2010. (Cited on p. 102.)
14. Z. D. Bai. Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistics Sinica*, 9:611–677, 1999. (Cited on p. 38.)
15. J. Baik, G. Ben-Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005. (Cited on p. 40.)
16. J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. 2005. (Cited on p. 40.)
17. A. S. Bandeira. Random Laplacian matrices and convex relaxations. *Available online at arXiv:1504.03987 [math.PR]*, 2015. (Cited on pp. 120, 121.)
18. A. S. Bandeira. A note on probably certifiably correct algorithms. *Comptes Rendus Mathematique, to appear*, 2016. (Cited on pp. 116, 121.)
19. A. S. Bandeira, E. Dobriban, D.G. Mixon, and W.F. Sawin. Certifying the restricted isometry property is hard. *IEEE Trans. Inform. Theory*, 59(6):3448–3450, 2013. (Cited on p. 148.)
20. A. S. Bandeira, M. Fickus, D. G. Mixon, and J. Moreira. Derandomizing restricted isometries via the Legendre symbol. *Available online at arXiv:1406.4089 [math.CO]*, 2014. (Cited on p. 148.)
21. A. S. Bandeira, M. Fickus, D. G. Mixon, and P. Wong. The road to deterministic matrices with the restricted isometry property. *Journal of Fourier Analysis and Applications*, 19(6):1123–1149, 2013. (Cited on pp. 148, 153, 153.)
22. A. S. Bandeira, M. E. Lewis, and D. G. Mixon. Discrete uncertainty principles and sparse signal processing. *Available online at arXiv:1504.01014 [cs.IT]*, 2015. (Cited on p. 148.)
23. A. S. Bandeira, D. G. Mixon, and J. Moreira. A conditional construction of restricted isometries. *Available online at arXiv:1410.6457 [math.FA]*, 2014. (Cited on p. 148.)
24. A. S. Bandeira and R. v. Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Annals of Probability, to appear*, 2015. (Cited on pp. 90, 120.)
25. B. Barak. Sum of squares upper bounds, lower bounds, and open questions. *Available online at* `http://www.boazbarak.org/sos/files/all-notes.pdf`, 2014. (Cited on p. 104.)
26. B. Barak and D. Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. *Survey, ICM 2014*, 2014. (Cited on p. 104.)
27. Richard Bellman. *Dynamic programming*. Princeton University Press, Princeton, 1957. (Cited on p. 3.)
28. F. Benaych-Georges and R. R. Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 2011. (Cited on p. 40.)
29. F. Benaych-Georges and R. R. Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 2012. (Cited on p. 40.)
30. J. Bourgain et al. Explicit constructions of RIP matrices and related problems. *Duke Mathematical Journal*, 159(1), 2011. (Cited on p. 148.)
31. E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52:489–509, 2006. (Cited on pp. 135, 139.)

32. E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59:1207–1223, 2006. (Cited on p. 139.)
33. E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51:4203–4215, 2005. (Cited on p. 139.)
34. E. J. Candès and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52:5406–5425, 2006. (Cited on pp. 139, 148, 152.)
35. E. J. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, May 2010. (Cited on p. 136.)
36. E.J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37(5A):2145–2177, 2009. (Cited on pp. 147, 154, 154.)
37. E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. (Cited on p. 136.)
38. Emmanuel J Candès. The restricted isometry property and its implications for compressed sensing. 346(9):589–592. (Cited on p. 145.)
39. P. G. Casazza and G. Kutyniok. *Finite Frames: Theory and Applications*. 2012. (Cited on p. 153.)
40. V. Chandrasekaran, B. Recht, P.A. Parrilo, and A.S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012. (Cited on p. 144.)
41. J. Cheeger. A lower bound for the smallest eigenvalue of the Laplacian. *Problems in analysis (Papers dedicated to Salomon Bochner, 1969), pp. 195–199. Princeton Univ. Press*, 1970. (Cited on p. 59.)
42. F. Chung. Four proofs for the cheeger inequality and graph partition algorithms. *Fourth International Congress of Chinese Mathematicians, pp. 331–349*, 2010. (Cited on p. 59.)
43. Michael B Cohen, Jelani Nelson, and David P Woodruff. Optimal approximate matrix product in terms of stable rank. *ICALP 2016*, pages 11:1–11:14, 2016. (Cited on p. 126.)
44. S. Dasgupta and A. Gupta. An elementary proof of the johnson-lindenstrauss lemma. Technical report, 2002. (Cited on p. 124.)
45. A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84, December 2011. (Cited on pp. 111, 112.)
46. Ling Shuyang Deng, Shaofeng and Thomas Strohmer. Strong consistency, graph Laplacians, and the Stochastic Block Model. *arXiv preprint arXiv:2004.09780*, 2020. (Cited on p. 121.)
47. E Dobriban. Permutation methods for factor analysis and pca. *arXiv preprint arXiv:1710.00479*, 2017. (Cited on p. 44.)
48. Edgar Dobriban and Art B Owen. Deterministic parallel analysis: an improved method for selecting factors and principal components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018. (Cited on p. 44.)
49. D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52:1289–1306, 2006. (Cited on pp. 135, 139.)
50. David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(32):375, 2000. (Cited on p. 16.)

51. David L Donoho, Matan Gavish, and Iain M Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of statistics*, 46(4):1742, 2018. (Cited on p. 44.)
52. Dominik Dorsch and Holger Rauhut. Refined analysis of sparse mimo radar. *Journal of Fourier Analysis and Applications*, 23(3):485–529, 2017. (Cited on p. 152.)
53. R. Durrett. *Random Graph Dynamics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, New York, NY, USA, 2006. (Cited on p. 97.)
54. Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019. (Cited on pp. 10, 16.)
55. Li Feng, Thomas Benkert, Kai Tobias Block, Daniel K Sodickson, Ricardo Otazo, and Hersh Chandarana. Compressed sensing for body MRI. *Journal of Magnetic Resonance Imaging*, 45(4):966–987, 2017. (Cited on p. 139.)
56. D. Féral and S. Péché. The largest eigenvalue of rank one deformation of large wigner matrices. *Communications in Mathematical Physics*, 272(1):185–228, 2006. (Cited on p. 43.)
57. S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhauser, 2013. (Cited on pp. 139, 140, 140, 147, 154, 155.)
58. J. J. Fuchs. On sparse representations in arbitrary redundant bases. *Information Theory, IEEE Transactions on*, 50(6):1341–1344, 2004. (Cited on p. 148.)
59. Amir Ghasemian, Pan Zhang, Aaron Clauset, Cristopher Moore, and Leto Peel. Detectability thresholds and optimal algorithms for community structure in dynamic networks. *Available online at arXiv:1506.06179 [stat.ML]*, 2015. (Cited on p. 112.)
60. M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefine programming. *Journal of the Association for Computing Machinery*, 42:1115–1145, 1995. (Cited on pp. 99, 101.)
61. G. H. Golub. *Matrix Computations*. Johns Hopkins University Press, third edition, 1996. (Cited on pp. 27, 35.)
62. Y. Gordon. Some inequalities for gaussian processes and applications. *Israel J. Math*, 50:109–110, 1985. (Cited on p. 133.)
63. Y. Gordon. On milnan's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. 1988. (Cited on pp. 131, 132, 132, 133, 133.)
64. B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *Available online at arXiv:1412.6156*, 2014. (Cited on p. 120.)
65. N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *Available online at arXiv:0909.4061v2 [math.NA]*, 2009. (Cited on p. 36.)
66. J. Hastad. Some optimal inapproximability results. 2002. (Cited on p. 103.)
67. I. Haviv and O. Regev. The restricted isometry property of subsampled fourier matrices. *SODA 2016*. (Cited on pp. 148, 152.)
68. R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985. (Cited on pp. 27, 35, 152.)
69. R.A. Horn and C.R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1990. Corrected reprint of the 1985 original. (Cited on pp. 49, 50.)

70. Max Hügel, Holger Rauhut, and Thomas Strohmer. Remote sensing via ℓ_1-minimization. *Foundations of Computational Mathematics*, 14(1):115–150, 2014. (Cited on p. 154.)
71. W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984. (Cited on pp. 123, 124.)
72. I. M. Johnston. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001. (Cited on p. 40.)
73. N. E. Karoui. Recent results about the largest eigenvalue of random covariance matrices and statistical application. *Acta Physica Polonica B*, 36(9), 2005. (Cited on p. 40.)
74. S. Khot. On the power of unique 2-prover 1-round games. *Thiry-fourth annual ACM symposium on Theory of computing*, 2002. (Cited on p. 102.)
75. S. Khot, G. Kindler, E. Mossel, and R. O'Donnell. Optimal inapproximability results for max-cut and other 2-variable csps? 2005. (Cited on p. 102.)
76. S. A. Khot and N. K. Vishnoi. The unique games conjecture, integrality gap for cut problems and embeddability of negative type metrics into l1. *Available online at arXiv:1305.4581 [cs.CC]*, 2013. (Cited on p. 105.)
77. Shira Kritchman and Boaz Nadler. Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*, 94(1):19–32, 2008. (Cited on p. 44.)
78. Shira Kritchman and Boaz Nadler. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Transactions on Signal Processing*, 57(10):3930–3941, 2009. (Cited on p. 44.)
79. K. G. Larsen and J. Nelson. Optimality of the johnson-lindenstrauss lemma. *Available online at arXiv:1609.02094*, 2016. (Cited on p. 126.)
80. J. B. Lassere. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001. (Cited on p. 104.)
81. R. Latała. Some estimates of norms of random matrices. *Proc. Amer. Math. Soc.*, 133(5):1273–1282 (electronic), 2005. (Cited on p. 90.)
82. B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 2000. (Cited on p. 98.)
83. Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001. (Cited on p. 16.)
84. James R Lee and Assaf Naor. Embedding the diamond graph in l_p and dimension reduction in l_1. *Geometric & Functional Analysis GAFA*, 14(4):745–747, 2004. (Cited on p. 126.)
85. J.R. Lee, S.O. Gharan, and L. Trevisan. Multi-way spectral partitioning and higher–order cheeger inequalities. *STOC '12 Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, 2012. (Cited on pp. 62, 63.)
86. William Leeb and Elad Romanov. Optimal singular value shrinkage with noise homogenization. *arXiv preprint arXiv:1811.02201*, 2018. (Cited on p. 44.)
87. Lydia T Liu, Edgar Dobriban, and Amit Singer. *e* pca: High dimensional exponential family pca. *The Annals of Applied Statistics*, 12(4):2121–2150, 2018. (Cited on p. 44.)
88. S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, 1982. (Cited on p. 52.)

89. Michael Lustig, David Donoho, and John M Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. (Cited on pp. 139, 155.)
90. V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72(114):507–536, 1967. (Cited on p. 38.)
91. P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2), 2000. (Cited on p. 87.)
92. L. Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 694–703, New York, NY, USA, 2014. ACM. (Cited on p. 111.)
93. Vitali D Milman and Gideon Schechtman. Asymptotic theory of finite-dimensional normed spaces, lecture notes in mathematics 1200, 1986. (Cited on p. 16.)
94. D. G. Mixon. Explicit matrices with the restricted isometry property: Breaking the square-root bottleneck. *available online at arXiv:1403.3427 [math.FA]*, 2014. (Cited on p. 148.)
95. D. G. Mixon. Short, Fat matrices BLOG: Gordon's escape through a mesh theorem. 2014. (Cited on p. 132.)
96. M. S. Moslehian. Ky Fan inequalities. *Available online at arXiv:1108.1467 [math.FA]*, 2011. (Cited on p. 31.)
97. E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *Available online at arXiv:1311.4115 [math.PR]*, January 2014. (Cited on p. 111.)
98. E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. *Probability Theory and Related Fields (to appear)*, 2014. (Cited on p. 111.)
99. C. Musco and C. Musco. Stronger and faster approximate singular value decomposition via the block lanczos method. *Available at arXiv:1504.05477 [cs.DS]*, 2015. (Cited on p. 36.)
100. B. Nadler, N. Srebro, and X. Zhou. Semi-supervised learning with the graph laplacian: The limit of infinite unlabelled data. 2009. (Cited on p. 80.)
101. Y. Nesterov. Squared functional systems and optimization problems. *High performance optimization*, 13(405-440), 2000. (Cited on p. 104.)
102. P. A. Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, 2000. (Cited on p. 104.)
103. D. Paul. Asymptotics of the leading sample eigenvalues for a spiked covariance model. *Available online at* `http://anson.ucdavis.edu/~debashis/techrep/eigenlimit.pdf`. (Cited on p. 40.)
104. D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistics Sinica*, 17:1617–1642, 2007. (Cited on p. 40.)
105. K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6*, 2(11):559–572, 1901. (Cited on p. 32.)
106. A. Perry, A. S. Wein, A. S. Bandeira, and A. Moitra. Optimality and sub-optimality of pca for spiked random matrices and synchronization. *Available online at arXiv:1609.05573 [math.ST]*, 2016. (Cited on p. 43.)
107. G. Pisier. *Introduction to operator space theory*, volume 294 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 2003. (Cited on p. 87.)

108. P. Raghavendra. Optimal algorithms and inapproximability results for every CSP? In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, pages 245–254. ACM, 2008. (Cited on p. 103.)
109. Holger Rauhut. Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery*, 9:1–92, 2010. (Cited on p. 155.)
110. B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011. (Cited on p. 136.)
111. S. Riemer and C. Schütt. On the expectation of the norm of random matrices with non-identically distributed entries. *Electron. J. Probab.*, 18, 2013. (Cited on p. 90.)
112. V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *Available at arXiv:0809.2274 [stat.CO]*, 2009. (Cited on p. 36.)
113. Sheldon M Ross. *Introduction to probability models*. Academic press, 2014. (Cited on p. 10.)
114. K. Schmudgen. Around hilbert's 17th problem. *Documenta Mathematica - Extra Volume ISMP*, pages 433–438, 2012. (Cited on p. 104.)
115. A. J. Scott and M. Grassl. Sic-povms: A new computer study. *J. Math. Phys.*, 2010. (Cited on p. 153.)
116. Y. Seginer. The expected norm of random matrices. *Combin. Probab. Comput.*, 9(2):149–166, 2000. (Cited on p. 90.)
117. N. Shor. An approach to obtaining global extremums in polynomial mathematical programming problems. *Cybernetics and Systems Analysis*, 23(5):695–700, 1987. (Cited on p. 104.)
118. G. Stengle. A nullstellensatz and a positivstellensatz in semialgebraic geometry. *Math. Ann. 207*, 207:87–97, 1974. (Cited on p. 104.)
119. T. Strohmer and B. Friedlander. Analysis of sparse MIMO radar. *Applied and Computational Harmonic Analysis*, 37:361–388, 2014. (Cited on p. 154.)
120. T. Strohmer and R. Heath. Grassmannian frames with applications to coding and communications. *Appl. Comput. Harmon. Anal.*, 14(3):257–275, 2003. (Cited on p. 153.)
121. M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Etudes Sci. Publ. Math.*, (81):73–205, 1995. (Cited on p. 87.)
122. T. Tao. What's new blog: Open question: deterministic UUP matrices. 2007. (Cited on pp. 148, 153.)
123. T. Tao. *Topics in Random Matrix Theory*. Graduate studies in mathematics. American Mathematical Soc., 2012. (Cited on pp. 38, 84, 87.)
124. J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. (Cited on pp. 70, 71, 72, 73, 73.)
125. R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996. (Cited on p. 154.)
126. A. M. Tillmann and M. E. Pfefsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. 2013. (Cited on p. 148.)
127. L. Trevisan. in theory BLOG: CS369G Llecture 4: Spectral Partitionaing. 2011. (Cited on p. 59.)

128. J. A. Tropp. Recovery of short, complex linear combinations via ℓ_1 minimization. *IEEE Transactions on Information Theory*, 4:1568–1570, 2005. (Cited on p. 148.)
129. J. A. Tropp. On the conditioning of random subdictionaries. *Appl. Comput. Harmon. Anal.*, 25:1–24, 2008. (Cited on p. 154.)
130. J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. (Cited on pp. 83, 87.)
131. J. A. Tropp. The expected norm of a sum of independent random matrices: An elementary approach. *Available at arXiv:1506.04711 [math.PR]*, 2015. (Cited on pp. 91, 93, 93, 94, 94, 96.)
132. J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 2015. (Cited on pp. 84, 88.)
133. J. A. Tropp. Second-order matrix concentration inequalities. *In preparation*, 2015. (Cited on pp. 91, 97, 97.)
134. R. van Handel. Probability in high dimensions. *ORF 570 Lecture Notes, Princeton University*, 2014. (Cited on p. 84.)
135. L. Vanderberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38:49–95, 1996. (Cited on pp. 100, 103, 114, 116.)
136. L. Vanderberghe and S. Boyd. *Convex Optimization*. Cambridge University Press, 2004. (Cited on pp. 103, 142.)
137. Shreyas S Vasanawala, Marcus T Alley, Brian A Hargreaves, Richard A Barth, John M Pauly, and Michael Lustig. Improved pediatric MR imaging with compressed sensing. *Radiology*, 256(2):607–616, 2010. (Cited on p. 140.)
138. Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018. (Cited on pp. 14, 22, 25, 25.)
139. G. Zauner. *Quantendesigns–Grundzüge einer nichtkommutativen Designtheorie*. PhD thesis, PhD Thesis Universität Wien, 1999. (Cited on p. 153.)
140. Pan Zhang, Cristopher Moore, and Lenka Zdeborova. Phase transitions in semisupervised clustering of sparse networks. *Phys. Rev. E*, 90, 2014. (Cited on p. 112.)