Ten Lectures and Forty-Two Open Problems in the Mathematics of Data Science

Afonso S. Bandeira
bandeira@cims.nyu.edu
http://www.cims.nyu.edu/~bandeira/

October 10, 2016

Preface

These are notes from a course I am giving at NYU this Fall (2016), and one I gave at MIT on the Fall of 2015. **These notes are not in final form and will be continuously edited and/or corrected (as I am sure they contain many typos)**. Please use at your own risk and do let me know if you find any typo/mistake.

Part of the content of this course is greatly inspired by a course I took from Amit Singer while a graduate student at Princeton. Amit's course was inspiring and influential on my research interests. I can only hope that these notes may one day inspire someone's research in the same way that Amit's course inspired mine.

These notes also include a total of forty-two open problems

This list of problems does not necessarily contain the most important problems in the field (although some will be rather important). I have tried to select a mix of important, perhaps approachable, and fun problems. Hopefully you will enjoy thinking about these problems as much as I do!

I would like to thank all the students who took my course at MIT and the ones who are taking it at NYU, they form a great interactive audience! I would like to thank Nicolas Boumal, Dustin G. Mixon, Bernat Guillen Pegueroles, Philippe Rigollet, and Francisco Unda for suggesting open problems.

Contents

0.1 List of open problems

- 0.1.A: Komlos Conjecture
- 0.2.A: Matrix AM-GM Inequality
- 1.1.A: Mallat and Zeitouni's problem
- 1.2.A: Monotonicity of eigenvalues
- 1.3.B.: Cut rank constrained SDP Spike Model problem

- 2.1.A: Ramsey numbers
- 2.2.A: Erdos-Hajnal Conjecture
- 2.3.A: Planted Clique Problems
- 3.1.A: Optimality of Cheeger's inequality
- 3.2.A: Certifying positive-semidefiniteness
- 3.3.A: Multy-way Cheeger's inequality
- 4.1.A: Non-commutative Khintchine improvement
- 4.2.A: Latala-Riemer-Schutt Problem
- 4.3.A: Matrix Six deviations Suffice
- 4.4.A: OSNAP problem
- 4.5.A: Random k-lifts of graphs
- 4.6.A: Feige's Conjecture
- 5.1.A: Deterministic Restricted Isometry Property matrices
- 5.2.A: Certifying the Restricted Isometry Property
- 6.1.A: Random Partial Discrete Fourier Transform
- 6.2.A: Mutually Unbiased Bases
- 6.3.A: Zauner's Conjecture (SIC-POVM)
- 6.4.A: The Paley ETF Conjecture
- 6.5.A: Constructive Kadison-Singer
- 7.1.A: Gilbert-Varshamov bound
- 7.2.A: Boolean Classification and Annulus Conjecture
- 7.3.A: Shannon Capacity of 7 cycle
- 7.4.A: The Deletion Channel
- 8.1.A: The Unique Games Conjecture
- 8.2.A: Sum of Squares approximation ratio for Max-Cut
- 8.3.A: The Grothendieck Constant
- 8.4.A: The Paley Clique Problem

- 8.5.A: Maximum and minimum bisections on random regular graphs
- 9.1.A: Detection Threshold for SBM for three of more communities
- 9.2.A: Recovery Threshold for SBM for logarithmic many communities
- 9.3.A: Tightness of k-median LP
- 9.4.A: Stability conditions for tightness of k-median LP and k-means SDP
- 9.5.A: Positive PCA tightness
- 10.1.A: Angular Synchronization via Projected Power Method
- 10.2.A: Sharp tightness of the Angular Synchronization SDP
- 10.3.A: Tightness of the Multireference Alignment SDP
- 10.4.A: Consistency and sample complexity of Multireference Alignment

0.2 A couple of Open Problems

We start with a couple of open problems:

0.2.1 Komlós Conjecture

We start with a fascinating problem in Discrepancy Theory.

Open Problem 0.1 (Komlós Conjecture) *Given n, let $K(n)$ denote the infimum over all real numbers such that: for all set of n vectors $u_1, \dots, u_n \in \mathbb{R}^n$ satisfying $\|u_i\|_2 \leq 1$, there exist signs $\epsilon_i = \pm 1$ such that*

$$\|\epsilon_1 u_1 + \epsilon_2 u_2 + \cdots + \epsilon_n u_n\|_\infty \leq K(n).$$

There exists a universal constant K such that $K(n) \leq K$ for all n.

An early reference for this conjecture is a book by Joel Spencer [Spe94]. This conjecture is tightly connected to Spencer's famous *Six Standard Deviations Suffice* Theorem [Spe85]. Later in the course we will study semidefinite programming relaxations, recently it was shown that a certain semidefinite relaxation of this conjecture holds [Nik13], the same paper also has a good accounting of partial progress on the conjecture.

- It is not so difficult to show that $K(n) \leq \sqrt{n}$, **try it!**

0.2.2 Matrix AM-GM inequality

We move now to an interesting generalization of arithmetic-geometric means inequality, which has applications on understanding the difference in performance of with- versus without-replacement sampling in certain randomized algorithms (see [RR12]).

Open Problem 0.2 *For any collection of $d \times d$ positive semidefinite matrices $A_1, \cdots, A_n$, the following is true:*

(a)

$$\left\| \frac{1}{n!} \sum_{\sigma \in \mathrm{Sym}(n)} \prod_{j=1}^{n} A_{\sigma(j)} \right\| \leq \left\| \frac{1}{n^n} \sum_{k_1,\ldots,k_n=1}^{n} \prod_{j=1}^{n} A_{k_j} \right\|,$$

and

(b)

$$\frac{1}{n!} \sum_{\sigma \in \mathrm{Sym}(n)} \left\| \prod_{j=1}^{n} A_{\sigma(j)} \right\| \leq \frac{1}{n^n} \sum_{k_1,\ldots,k_n=1}^{n} \left\| \prod_{j=1}^{n} A_{k_j} \right\|,$$

where $\mathrm{Sym}(n)$ *denotes the group of permutations of n elements, and $\| \cdot \|$ the spectral norm.*

Morally, these conjectures state that products of matrices with repetitions are larger than without. For more details on the motivations of these conjecture (and their formulations) see [RR12] for conjecture **(a)** and [Duc12] for conjecture **(b)**.

Recently these conjectures have been solved for the particular case of $n = 3$, in [Zha14] for **(a)** and in [IKW14] for **(b)**.

0.3 Brief Review of some linear algebra tools

In this Section we'll briefly review a few linear algebra tools that will be important during the course. If you need a refresh on any of these concepts, I recommend taking a look at [HJ85] and/or [Gol96].

0.3.1 Singular Value Decomposition

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

$$M = U \Sigma V^T, \tag{1}$$

where $U \in O(m)$, $V \in O(n)$ are orthogonal matrices (meaning that $U^T U = U U^T = \mathrm{I}$ and $V^T V = V V^T = \mathrm{I}$) and $\Sigma \in \mathbb{R}^{m \times n}$ is a matrix with non-negative entries in 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.

Remark 0.1 *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 $U U^T = \mathrm{I}$, $\Sigma \in \mathbb{R}^{n \times n}$ a diagonal matrix with non-negative entries and $V \in O(n)$.*

0.3.2 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.

The spectral norm of M is defined as

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

0.3.3 Trace and norm

Given a matrix $M \in \mathbb{R}^{n \times n}$, its trace is given by

$$\mathrm{Tr}(M) = \sum_{k=1}^{n} M_{kk} = \sum_{k=1}^{n} \lambda_k \left(M \right).$$

Its Frobeniues norm is given by

$$\|M\|_F = \sqrt{\sum_{ij} M_{ij}^2} = \mathrm{Tr}(M^T M)$$

A particularly important property of the trace is that:

$$\mathrm{Tr}(AB) = \sum_{i,j=1}^{n} A_{ij} B_{ji} = \mathrm{Tr}(BA).$$

Note that this implies that, e.g., $\mathrm{Tr}(ABC) = \mathrm{Tr}(CAB)$, it does not imply that, e.g., $\mathrm{Tr}(ABC) = \mathrm{Tr}(ACB)$ which is not true in general!

0.4 Quadratic Forms

During the course we will be interested in solving problems of the type

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

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

Note that this is equivalent to

$$\max_{\substack{v_1,\ldots,v_d \in \mathbb{R}^n \\ v_i^T v_j = \delta_{ij}}} \sum_{k=1}^{d} v_k^T M v_k, \tag{2}$$

where δ_{ij} is the Kronecker delta (is 1 is $i = j$ and 0 otherwise).

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

$$\max_{\substack{v \in \mathbb{R}^n \\ \|v\|_2 = 1}} v^T M v. \tag{3}$$

It is easy to see (for example, using the spectral decomposition of M) that (3) is maximized by the leading eigenvector of M and

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

It is also not very difficult to see (it follows for example from a Theorem of Fan (see, for example, page 3 of [Mos11]) that (2) is maximized by taking $v_1, \ldots, v_d$ to be the k leading eigenvectors of M and that its value is simply the sum of the k largest eigenvalues of M. The nice consequence of this is that the solution to (2) can be computed sequentially: we can first solve for $d = 1$, computing v_1, then v_2, and so on.

Remark 0.2 *All of the tools and results above have natural analogues when the matrices have complex entries (and are Hermitian instead of symmetric).*

1 Principal Component Analysis in High Dimensions and the Spike Model

1.1 Dimension Reduction and PCA

When faced with a high dimensional dataset, a natural approach is to try to reduce its dimension, either by projecting it to a lower dimension space or by finding a better representation for the data. During this course we will see a few different ways of doing dimension reduction.

We will start with Principal Component Analysis (PCA). In fact, PCA continues to be one of the best (and simplest) tools for exploratory data analysis. Remarkably, it dates back to a 1901 paper by Karl Pearson [Pea01]!

Let's say 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. There are a couple of different ways we can try 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 preserved 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

$$\mu_n = \frac{1}{n}\sum_{k=1}^{n} x_k, \tag{4}$$

and its sample covariance as

$$\Sigma_n = \frac{1}{n-1}\sum_{k=1}^{n} \left(x_k - \mu_n\right)\left(x_k - \mu_n\right)^T. \tag{5}$$

Remark 1.1 *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.*

We will start with the first interpretation of PCA and then show that it is equivalent to the second.

1.1.1 PCA as best d-dimensional affine fit

We are trying to approximate each x_k by

$$x_k \approx \mu + \sum_{i=1}^{d} \left(\beta_k\right)_i v_i, \tag{6}$$

where $v_1, \ldots, v_d$ is an orthonormal basis for the d-dimensional subspace, $\mu \in \mathbb{R}^p$ represents the translation, and β_k corresponds to the coefficients of x_k. If we represent the subspace by $V = [v_1 \cdots v_d] \in \mathbb{R}^{p \times d}$ then we can rewrite (7) as

$$x_k \approx \mu + V\beta_k, \tag{7}$$

where $V^T V = \mathrm{I}_{d\times d}$ as the vectors v_i are orthonormal.

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

$$\min_{\substack{\mu,\ V,\ \beta_k \\ V^T V = \mathrm{I}}} \sum_{k=1}^{n} \left\| x_k - (\mu + V\beta_k) \right\|_2^2 \tag{8}$$

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

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

Thus, the optimal value μ^* of μ satisfies

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

Because we can assume, without loss of generality, 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 on finding the solution for (9) by solving

$$\min_{\substack{V,\ \beta_k \\ V^T V = \mathrm{I}}} \sum_{k=1}^{n} \left\| x_k - \mu_n - V\beta_k \right\|_2^2. \tag{9}$$

Let us proceed by optimizing for β_k. Since the problem decouples for each k, we can focus on, for each k,

$$\min_{\beta_k} \left\| x_k - \mu_n - V\beta_k \right\|_2^2 = \min_{\beta_k} \left\| x_k - \mu_n - \sum_{i=1}^{d} (\beta_k)_i \, v_i \right\|_2^2. \tag{10}$$

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 $\beta_k = V^T (x_k - \mu_n)$. Thus, (9) is equivalent to

$$\min_{V^T V = \mathrm{I}} \sum_{k=1}^{n} \left\| (x_k - \mu_n) - VV^T (x_k - \mu_n) \right\|_2^2. \tag{11}$$

Note that

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

Since $(x_k-\mu_n)^T(x_k-\mu_n)$ does not depend on V, minimizing (9) is equivalent to

$$\max_{V^TV=\mathrm{I}}\sum_{k=1}^{n}(x_k-\mu_n)^T VV^T(x_k-\mu_n). \qquad (12)$$

A few more simple algebraic manipulations using properties of the trace:

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

This means that the solution to (13) is given by

$$\max_{V^TV=\mathrm{I}}\mathrm{Tr}\left[V^T\Sigma_n V\right]. \qquad (13)$$

As we saw above (recall (2)) the solution is given by $V=[v_1,\cdots,v_d]$ where $v_1,\dots,v_d$ correspond to the d leading eigenvectors of Σ_n.

Let us first show that interpretation (2) of finding the d-dimensional projection of $x_1,\dots,x_n$ that preserves the most variance also arrives to the optimization problem (13).

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

We aim to find an orthonormal basis $v_1,\dots,v_d$ (organized as $V=[v_1,\dots,v_d]$ with $V^TV=\mathrm{I}_{d\times d}$) of a d-dimensional space such that the projection of $x_1,\dots,x_n$ projected on this subspace has the most variance. Equivalently we can ask for the points

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

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

$$\max_{V^TV=\mathrm{I}} \sum_{k=1}^{n} \left\| V^T x_k - \frac{1}{n}\sum_{r=1}^{n} V^T x_r \right\|^2 . \tag{14}$$

Note that

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

showing that (14) is equivalent to (13) and that the two interpretations of PCA are indeed equivalent.

1.1.3 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 find the leading eigenvectors of

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

A naive way of doing this would be 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 [HJ85] and/or [Gol96]).

An alternative is to use the Singular Value Decomposition (1). Let $X = [x_1 \cdots x_n]$ recall that,

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

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,

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

meaning that U_L correspond to the eigenvectors of Σ_n. Computing the SVD of $X - \mu_n \mathbf{1}^T$ takes $\mathcal{O}(\min n^2 p, p^2 n)$ 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 (see for example [HMT09, RST09, MM15]).[1]

1.1.4 Which d should we pick?

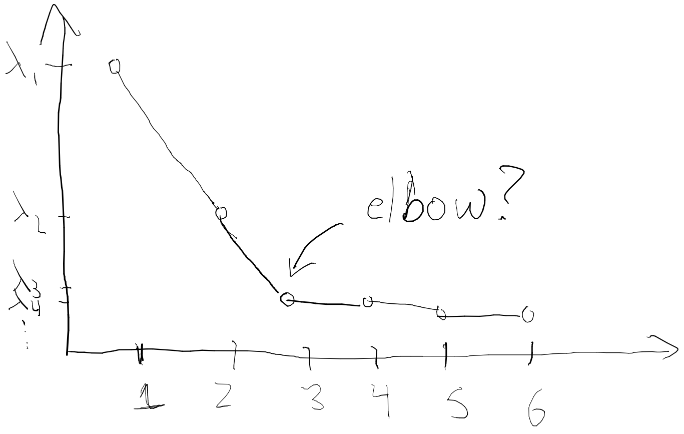
Given a dataset, if the objective is to visualize it then picking $d = 2$ or $d = 3$ might make the most sense. However, PCA is useful for many other purposes, for example: (1) often times the data belongs to a lower dimensional space but is corrupted by high dimensional noise. When using PCA it is oftentimess possible to reduce the noise while keeping the signal. (2) One may be interested in running an algorithm that would be too computationally expensive to run in high dimensions,

[1] If there is time, we might discuss some of these methods later in the course.

dimension reduction may help there, etc. In these applications (and many others) it is not clear how to pick d.

If we denote the k-th largest eigenvalue of Σ_n as $\lambda_k^{(+)}(\Sigma_n)$, then the k-th principal component has a $\frac{\lambda_k^{(+)}(\Sigma_n)}{\mathrm{Tr}(\Sigma_n)}$ proportion of the variance. [2]

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. This is usually visualized by a scree plot: a plot of the values of the ordered eigenvalues. 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 try to understand better the behavior of the eigenvalues of Σ_n and it will help us understand when to cut-off.

1.1.5 A related open problem

We now show an interesting open problem posed by Mallat and Zeitouni at [MZ11]

Open Problem 1.1 (A. Mallat and Zeitouni [MZ11]) *Let $g \sim \mathcal{N}(0, \Sigma)$ be a gaussian random vector in $\mathbb{R}^p$ with a known covariance matrix Σ and $d < p$. Now, for any orthonormal basis $V = [v_1, \dots, v_p]$ of $\mathbb{R}^p$, consider the following random variable Γ_V: Given a draw of the random vector g, Γ_V is the squared ℓ_2 norm of the largest projection of g on a subspace generated by d elements of the basis V. The question is:*

What is the basis V for which $\mathbb{E}[\Gamma_V]$ is maximized?

[2] Note that $\mathrm{Tr}(\Sigma_n) = \sum_{k=1}^{p} \lambda_k(\Sigma_n)$.

The conjecture in [MZ11] is that the optimal basis is the eigendecomposition of Σ. It is known that this is the case for $d = 1$ (see [MZ11]) but the question remains open for $d > 1$. It is not very difficult to see that one can assume, without loss of generality, that Σ is diagonal.

A particularly intuitive way of stating the problem is:

1. Given $\Sigma \in \mathbb{R}^{p\times p}$ and d

2. Pick an orthonormal basis $v_1, \dots, v_p$

3. Given $g \sim \mathcal{N}(0, \Sigma)$

4. Pick d elements $\tilde{v}_1, \dots, \tilde{v}_d$ of the basis

5. `Score`: $\sum_{i=1}^{d} \left(\tilde{v}_i^T g\right)^2$

The objective is to pick the basis in order to maximize the expected value of the `Score`.

Notice that if the steps of the procedure were taken in a slightly different order on which step 4 would take place before having access to the draw of g (step 3) then the best basis is indeed the eigenbasis of Σ and the best subset of the basis is simply the leading eigenvectors (notice the resemblance with PCA, as described above).

More formally, we can write the problem as finding

$$\underset{\substack{V\in\mathbb{R}^{p\times p}\\ V^TV=\mathrm{I}}}{\operatorname{argmax}} \left(\mathbb{E}\left[\max_{\substack{S\subset[p]\\ |S|=d}} \sum_{i\in S} \left(v_i^T g\right)^2 \right] \right), \tag{15}$$

where $g \sim \mathcal{N}(0, \Sigma)$. The observation regarding the different ordering of the steps amounts to saying that the eigenbasis of Σ is the optimal solution for

$$\underset{\substack{V\in\mathbb{R}^{p\times p}\\ V^TV=\mathrm{I}}}{\operatorname{argmax}} \left(\max_{\substack{S\subset[p]\\ |S|=d}} \mathbb{E}\left[\sum_{i\in S} \left(v_i^T g\right)^2 \right] \right). \tag{16}$$

Recently, it was shown that the conjecture is true up to a multiplicative constant [LT16]. In other words, it was shown that (15) and (16) differ by at most a multiplicative universal constant[LT16].

1.2 PCA in high dimensions and Marcenko-Pastur

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

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

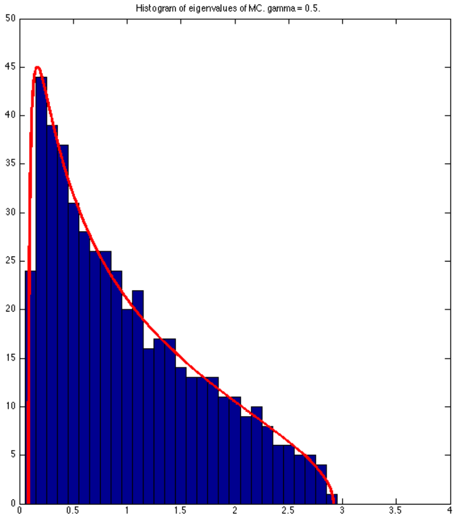
p the number of pixels per image; it is conceivable that the number of pixels be on the order of the number of images in a set. Unfortunately, in that case, it is no longer clear that $\Sigma_n \to \Sigma$. Dealing with this type of difficulties is the realm of high dimensional statistics.

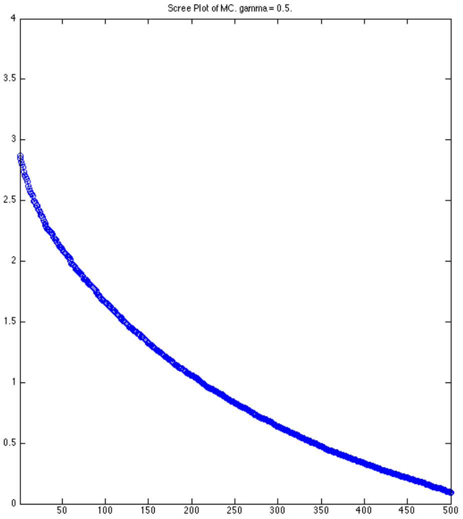
For simplicity we will instead try to understand the spectral properties of

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

where $x_1, \dots, x_n$ are the columns of X. Since $x \sim \mathcal{N}(0, \Sigma)$ we know that $\mu_n \to 0$ (and, clearly, $\frac{n}{n-1} \to 1$) the spectral properties of S_n will be essentially the same as Σ_n.[3]

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 Marchenko-Pastur distribution (17), that we will discuss below.





As one can see in the image, there are many eigenvalues considerably larger than 1 (and some considerably larger than others). 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. [Tao12] and [AGZ10]). This particular limiting distribution was first established in 1967 by Marchenko and Pastur [MP67] and is now referred to as the Marchenko-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

[3] In this case, S_n is actually the Maximum likelihood estimator for Σ, we'll talk about Maximum likelihood estimation later in the course.

$$dF_\gamma(\lambda) = \frac{1}{2\pi}\frac{\sqrt{(\gamma_+ - \lambda)(\lambda - \gamma_-)}}{\gamma\lambda}1_{[\gamma_-,\gamma_+]}(\lambda)d\lambda, \tag{17}$$

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

Remark 1.2 *We will not show the proof of the Marchenko-Pastur Theorem here (you can see, for example, [Bai99] for several different proofs of it), but an approach to a proof is using the so-called moment method. The core of the 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,*

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

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

1.2.1 A related open problem

Open Problem 1.2 (Monotonicity of singular values [BKS13a]) *Consider the setting above but with $p = n$, then $X \in \mathbb{R}^{n\times n}$ is a matrix with iid $\mathcal{N}(0,1)$ entries. Let*

$$\sigma_i\left(\frac{1}{\sqrt{n}}X\right),$$

denote the i-th singular value[4] of $\frac{1}{\sqrt{n}}X$, and define

$$\alpha_{\mathbb{R}}(n) := \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i\left(\frac{1}{\sqrt{n}}X\right)\right],$$

as the expected value of the average singular value of $\frac{1}{\sqrt{n}}X$.

The conjecture is that, for every $n \geq 1$,

$$\alpha_{\mathbb{R}}(n+1) \geq \alpha_{\mathbb{R}}(n).$$

Moreover, for the analogous quantity $\alpha_{\mathbb{C}}(n)$ defined over the complex numbers, meaning simply that each entry of X is an iid complex valued standard gaussian $\mathbb{C}\mathcal{N}(0,1)$ the reverse inequality is conjectured for all $n \geq 1$:

$$\alpha_{\mathbb{C}}(n+1) \leq \alpha_{\mathbb{C}}(n).$$

[4] The i-th diagonal element of Σ in the SVD $\frac{1}{\sqrt{n}}X = U\Sigma V$.

Notice that the singular values of $\frac{1}{\sqrt{n}}X$ are simply the square roots of the eigenvalues of S_n,

$$\sigma_i\left(\frac{1}{\sqrt{n}}X\right) = \sqrt{\lambda_i\left(S_n\right)}.$$

This means that we can compute $\alpha_{\mathbb{R}}$ in the limit (since we know the limiting distribution of $\lambda_i\left(S_n\right)$) and get (since $p = n$ we have $\gamma = 1$, $\gamma_- = 0$, and $\gamma_+ = 2$)

$$\lim_{n\to\infty} \alpha_{\mathbb{R}}(n) = \int_0^2 \lambda^{\frac{1}{2}} dF_1(\lambda) = \frac{1}{2\pi}\int_0^2 \lambda^{\frac{1}{2}}\frac{\sqrt{(2-\lambda)\lambda}}{\lambda} = \frac{8}{3\pi} \approx 0.8488.$$

Also, $\alpha_{\mathbb{R}}(1)$ simply corresponds to the expected value of the absolute value of a standard gaussian g

$$\alpha_{\mathbb{R}}(1) = \mathbb{E}|g| = \sqrt{\frac{2}{\pi}} \approx 0.7990,$$

which is compatible with the conjecture.

On the complex valued side, the Marchenko-Pastur distribution also holds for the complex valued case and so $\lim_{n\to\infty} \alpha_{\mathbb{C}}(n) = \lim_{n\to\infty} \alpha_{\mathbb{R}}(n)$ and $\alpha_{\mathbb{C}}(1)$ can also be easily calculated and seen to be larger than the limit.

L. D. Abreu recently resolved the complex version of Open Problem 1.2, the solution is available here [Abr16]. The real part of the Conjecture remains, to be best of our knowledge, open.

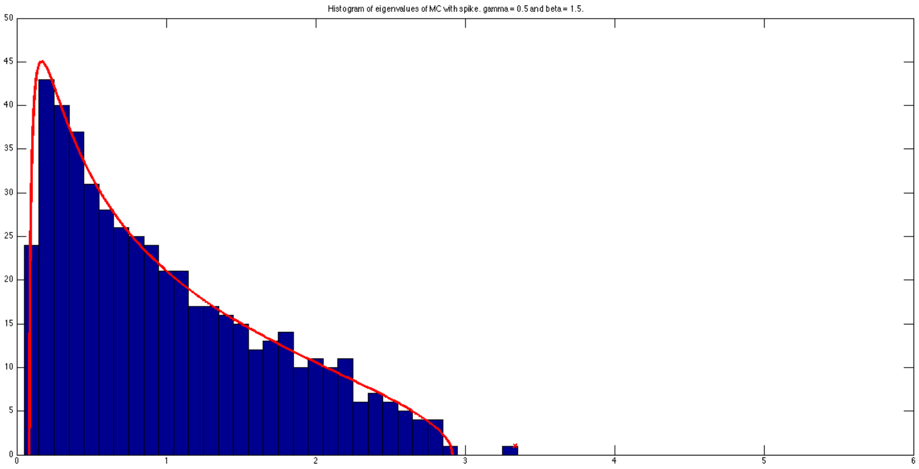
1.3 Spike Models and BBP transition

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

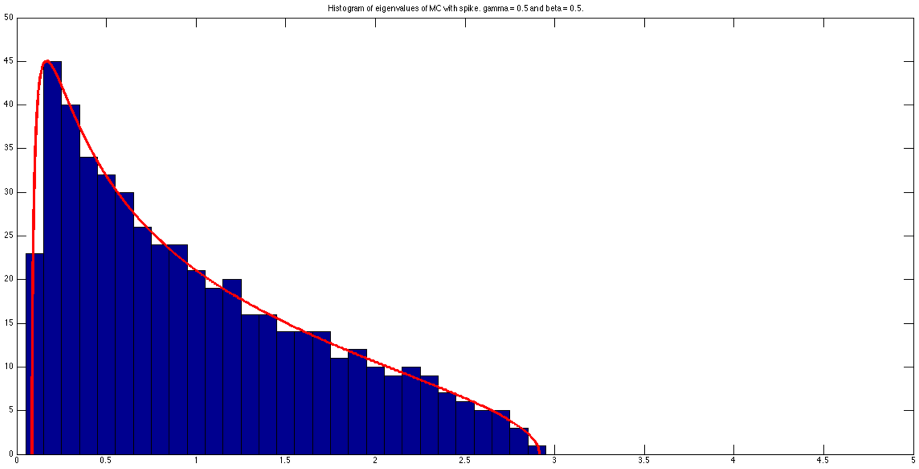
One way to think about this instance is as each data point x consisting of a signal part $\sqrt{\beta}g_0v$ where g_0 is a one-dimensional standard gaussian (a gaussian multiple of a fixed vector $\sqrt{\beta}v$ and a noise part $g \sim \mathcal{N}(0, I)$ (independent of g_0. Then $x = g + \sqrt{\beta}g_0v$ is a gaussian random variable

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

A natural question is whether this rank 1 perturbation can be seen in S_n. Let us build some intuition with an example, the following is the histogram of the eigenvalues of a sample of S_n for $p = 500$, $n = 1000$, v is the first element of the canonical basis $v = e_1$, and $\beta = 1.5$:



The images suggests that there is an eigenvalue of S_n that "pops out" of the support of the Marchenko-Pastur distribution (below we will estimate the location of this eigenvalue, and that estimate corresponds to the red "x"). It is worth noticing 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 $\beta = 0.5$:



and it appears that, for $\beta = 0.5$, the distribution of the eigenvalues appears to be undistinguishable from when $\Sigma = \mathrm{I}$.

This motivates the following question:

Question 1.3 *For which values of γ and β do we expect to see an eigenvalue of S_n popping out of the support of the Marchenko-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 below which we don't expect to see a change in the distribution of eigenvalues and above which we expect one of the eigenvalues to pop out of the support, this is known as BBP transition (after Baik, Ben Arous, and Péché [BBAP05]). There are many very nice papers about this and similar phenomena, including [Pau, Joh01, BBAP05, Pau07, BS05, Kar05, BGN11, BGN12].[5]

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

First of all, it is not difficult to see that we can assume that $v = e_1$ (since everything else is rotation invariant). We want to understand the behavior of the leading eigenvalue of

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

where

$$X = [x_1, \dots, x_n] \in \mathbb{R}^{p\times n}.$$

We can write X as

$$X = \left[\begin{array}{c} \sqrt{1+\beta} Z_1^T \\ Z_2^T \end{array} \right],$$

where $Z_1 \in \mathbb{R}^{n\times 1}$ and $Z_2 \in \mathbb{R}^{n\times(p-1)}$, both populated with i.i.d. standard gaussian entries ($\mathcal{N}(0,1)$). Then,

$$S_n = \frac{1}{n} X X^T = \frac{1}{n}\left[\begin{array}{cc} (1+\beta) Z_1^T Z_1 & \sqrt{1+\beta} Z_1^T Z_2 \\ \sqrt{1+\beta} Z_2^T Z_1 & Z_2^T Z_2 \end{array} \right].$$

Now, let $\hat{\lambda}$ and $v = \left[\begin{array}{c} v_1 \\ v_2 \end{array} \right]$ where $v_2 \in \mathbb{R}^{p-1}$ and $v_1 \in \mathbb{R}$, denote, respectively, an eigenvalue and associated eigenvector for S_n. By the definition of eigenvalue and eigenvector we have

$$\frac{1}{n}\left[\begin{array}{cc} (1+\beta) Z_1^T Z_1 & \sqrt{1+\beta} Z_1^T Z_2 \\ \sqrt{1+\beta} Z_2^T Z_1 & Z_2^T Z_2 \end{array} \right] \left[\begin{array}{c} v_1 \\ v_2 \end{array} \right] = \hat{\lambda} \left[\begin{array}{c} v_1 \\ v_2 \end{array} \right],$$

which can be rewritten as

$$\frac{1}{n}(1+\beta) Z_1^T Z_1 v_1 + \frac{1}{n}\sqrt{1+\beta} Z_1^T Z_2 v_2 = \hat{\lambda} v_1 \tag{18}$$

$$\frac{1}{n}\sqrt{1+\beta} Z_2^T Z_1 v_1 + \frac{1}{n} Z_2^T Z_2 v_2 = \hat{\lambda} v_2. \tag{19}$$

(19) is equivalent to

$$\frac{1}{n}\sqrt{1+\beta} Z_2^T Z_1 v_1 = \left(\hat{\lambda}\, \mathrm{I} - \frac{1}{n} Z_2^T Z_2 \right) v_2.$$

[5] Notice that the Marchenko-Pastur theorem does not imply that all eigenvalues are actually in the support of the Marchenk-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, [Pau]).

If $\hat{\lambda}\,\mathrm{I} - \frac{1}{n} Z_2^T Z_2$ is invertible (this won't be justified here, but it is in [Pau]) then we can rewrite it as

$$v_2 = \left(\hat{\lambda}\,\mathrm{I} - \frac{1}{n} Z_2^T Z_2 \right)^{-1} \frac{1}{n} \sqrt{1+\beta} Z_2^T Z_1 v_1,$$

which we can then plug in (18) to get

$$\frac{1}{n}(1+\beta) Z_1^T Z_1 v_1 + \frac{1}{n}\sqrt{1+\beta} Z_1^T Z_2 \left(\hat{\lambda}\,\mathrm{I} - \frac{1}{n} Z_2^T Z_2 \right)^{-1} \frac{1}{n}\sqrt{1+\beta} Z_2^T Z_1 v_1 = \hat{\lambda} v_1$$

If $v_1 \neq 0$ (again, not properly justified here, see [Pau]) then this means that

$$\hat{\lambda} = \frac{1}{n}(1+\beta) Z_1^T Z_1 + \frac{1}{n}\sqrt{1+\beta} Z_1^T Z_2 \left(\hat{\lambda}\,\mathrm{I} - \frac{1}{n} Z_2^T Z_2 \right)^{-1} \frac{1}{n}\sqrt{1+\beta} Z_2^T Z_1 \tag{20}$$

First observation is that because $Z_1 \in \mathbb{R}^n$ has standard gaussian entries then $\frac{1}{n} Z_1^T Z_1 \to 1$, meaning that

$$\hat{\lambda} = (1+\beta)\left[1 + \frac{1}{n} Z_1^T Z_2 \left(\hat{\lambda}\,\mathrm{I} - \frac{1}{n} Z_2^T Z_2 \right)^{-1} \frac{1}{n} Z_2^T Z_1 \right]. \tag{21}$$

Consider the SVD of $Z_2 = U\Sigma V^T$ where $U \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{R}^{p\times p}$ have orthonormal columns (meaning that $U^T U = \mathrm{I}_{p\times p}$ and $V^T V = \mathrm{I}_{p\times p}$), and Σ is a diagonal matrix. Take $D = \frac{1}{n}\Sigma^2$ then

$$\frac{1}{n} Z_2^T Z_2 = \frac{1}{n} V \Sigma^2 V^T = V D V^T,$$

meaning that the diagonal entries of D correspond to the eigenvalues of $\frac{1}{n} Z_2^T Z_2$ which we expect to be distributed (in the limit) according to the Marchenko-Pastur distribution for $\frac{p-1}{n} \approx \gamma$. Replacing back in (21)

$$\begin{aligned}
\hat{\lambda} &= (1+\beta)\left[1 + \frac{1}{n} Z_1^T \left(\sqrt{n} U D^{1/2} V^T \right) \left(\hat{\lambda}\,\mathrm{I} - V D V^T \right)^{-1} \frac{1}{n} \left(\sqrt{n} U D^{1/2} V^T \right)^T Z_1 \right] \\
&= (1+\beta)\left[1 + \frac{1}{n} \left(U^T Z_1 \right)^T D^{1/2} V^T \left(\hat{\lambda}\,\mathrm{I} - V D V^T \right)^{-1} V D^{1/2} \left(U^T Z_1 \right) \right] \\
&= (1+\beta)\left[1 + \frac{1}{n} \left(U^T Z_1 \right)^T D^{1/2} V^T \left(V \left[\hat{\lambda}\,\mathrm{I} - D \right] V^T \right)^{-1} V D^{1/2} \left(U^T Z_1 \right) \right] \\
&= (1+\beta)\left[1 + \frac{1}{n} \left(U^T Z_1 \right)^T D^{1/2} \left(\left[\hat{\lambda}\,\mathrm{I} - D \right] \right)^{-1} D^{1/2} \left(U^T Z_1 \right) \right].
\end{aligned}$$

Since the columns of U are orthonormal, $g := U^T Z_1 \in \mathbb{R}^{p-1}$ is an isotropic gaussian ($g \sim \mathcal{N}(0,1)$), in fact,

$$\mathbb{E} g g^T = \mathbb{E} U^T Z_1 \left(U^T Z_1 \right)^T = \mathbb{E} U^T Z_1 Z_1^T U = U^T \mathbb{E}\left[Z_1 Z_1^T \right] U = U^T U = \mathrm{I}_{(p-1)\times(p-1)}.$$

We proceed

$$\begin{aligned}
\hat{\lambda} &= (1+\beta)\left[1 + \frac{1}{n} g^T D^{1/2} \left(\left[\hat{\lambda}\,\mathrm{I} - D \right] \right)^{-1} D^{1/2} g \right] \\
&= (1+\beta)\left[1 + \frac{1}{n} \sum_{j=1}^{p-1} g_j^2 \frac{D_{jj}}{\hat{\lambda} - D_{jj}} \right]
\end{aligned}$$

Because we expect the diagonal entries of D to be distributed according to the Marchenko-Pastur distribution and g to be independent to it we expect that (again, not properly justified here, see [Pau])

$$\frac{1}{p-1}\sum_{j=1}^{p-1} g_j^2 \frac{D_{jj}}{\hat{\lambda} - D_{jj}} \to \int_{\gamma_-}^{\gamma_+} \frac{x}{\hat{\lambda}-x} dF_\gamma(x).$$

We thus get an equation for $\hat{\lambda}$:

$$\hat{\lambda} = (1+\beta)\left[1+\gamma\int_{\gamma_-}^{\gamma_+} \frac{x}{\hat{\lambda}-x} dF_\gamma(x)\right],$$

which can be easily solved with the help of a program that computes integrals symbolically (such as Mathematica) to give (you can also see [Pau] for a derivation):

$$\hat{\lambda} = (1+\beta)\left(1+\frac{\gamma}{\beta}\right), \tag{22}$$

which is particularly elegant (specially considering the size of some the equations used in the derivation).

An important thing to notice is that for $\beta = \sqrt{\gamma}$ we have

$$\hat{\lambda} = (1+\sqrt{\gamma})\left(1+\frac{\gamma}{\sqrt{\gamma}}\right) = (1+\sqrt{\gamma})^2 = \gamma_+,$$

suggesting that $\beta = \sqrt{\gamma}$ is the critical point.

Indeed this is the case and it is possible to make the above argument rigorous[6] and show that in the model described above,

- If $\beta \le \sqrt{\gamma}$ then
$$\lambda_{\max}(S_n) \to \gamma_+,$$
- and if $\beta > \sqrt{\gamma}$ then
$$\lambda_{\max}(S_n) \to (1+\beta)\left(1+\frac{\gamma}{\beta}\right) > \gamma_+.$$

Another important question is wether the leading eigenvector actually correlates with the planted perturbation (in this case e_1). Turns out that very similar techniques can answer this question as well [Pau] and show that the leading eigenvector $v_{\max}$ of S_n will be non-trivially correlated with e_1 if and only if $\beta > \sqrt{\gamma}$, more precisely:

- If $\beta \le \sqrt{\gamma}$ then
$$|\langle v_{\max}, e_1\rangle|^2 \to 0,$$
- and if $\beta > \sqrt{\gamma}$ then
$$|\langle v_{\max}, e_1\rangle|^2 \to \frac{1-\frac{\gamma}{\beta^2}}{1-\frac{\gamma}{\beta}}.$$

[6] Note that in the argument above it wasn't even completely clear where it was used that the eigenvalue was actually the leading one. In the actual proof one first needs to make sure that there is an eigenvalue outside of the support and the proof only holds for that one, you can see [Pau]

1.3.1 A brief mention of Wigner matrices

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

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

and there is also a BBP like transition for this matrix ensemble [FP06]. 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
$$\lambda_{\max}\left(\frac{1}{\sqrt{n}}W + \xi vv^T\right) \to \xi + \frac{1}{\xi}. \tag{23}$$

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

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

Recent work addresses the problem of when is it that it is possible to statistically detect a spike in a random matrix, for different distributions on the spike and the underlying matrix [PWBM16]

1.3.2 An open problem about spike models

Open Problem 1.3 (Spike Model for rank constrainted SDP) *Let W denote a symmetric Wigner matrix with i.i.d. entries $W_{ij} \sim \mathcal{N}(0,1)$. Also, given $B \in \mathbb{R}^{n\times n}$ symmetric, define:*

$$Q_r(B) = \max\left\{\operatorname{Tr}(BX) : X \succeq 0,\ X_{ii} = 1,\ \operatorname{rank}(X) \leq r\right\}.$$

Define $q_r(\xi)$ as

$$q_r(\xi) = \lim_{n\to\infty}\frac{1}{n}\mathbb{E}Q\left(\frac{\xi}{n}\mathbf{1}\mathbf{1}^T + \frac{1}{\sqrt{n}}W\right).$$

What is the value of ξ_r^*, defined as

$$\xi_r^* = \inf\{\xi \geq 0 : q_r(\xi) > 2\}.$$

Remark 1.4 *Optimization problems of the type of* $\max\left\{\operatorname{Tr}(BX) : X \succeq 0,\ X_{ii} = 1\right\}$ *are semidefinite programs, they will be a major player later in the course!*

The case $r = n$ (which was Open Problem (1.3).A. was recently solved in [MS15], showing that $\xi_n^* = 2$. This implies that a certain semidefinite programming based algorithm for clustering under the Stochastic Block Model on 2 clusters (we will discuss these things later in the course) is optimal for detection (see [MS15]).[7]

Since $\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\left[11^T\left(\frac{\xi}{n}\mathbf{1}\mathbf{1}^T + \frac{1}{\sqrt{n}}W\right)\right] \approx \xi$, by taking $X = 11^T$ we expect that $q_r(\xi) \geq \xi$ for all r. For $r = 1$ the value of $q_1(0)$ relates to the celebrated Sherrington-Kirkpatrick Model [Pan13], and it is known that $q_0(0) = 2P_*$, where P_* is the so called Parisi constant [Pan13].

The value of ξ_1^* is conjectured to be 2 in [JMRT15] (inspired in replica calculations) but a proof of this fact remains an interesting open problem.

[7] Later in the course we will discuss clustering under the Stochastic Block Model quite thoroughly, and will see how this same SDP is known to be optimal for exact recovery [ABH14, HWX14, Ban15b].

2 Graphs, Diffusion Maps, and Semi-supervised Learning

2.1 Graphs

Graphs will be one of the main objects of study through these lectures, it is time to introduce them. 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[8]:

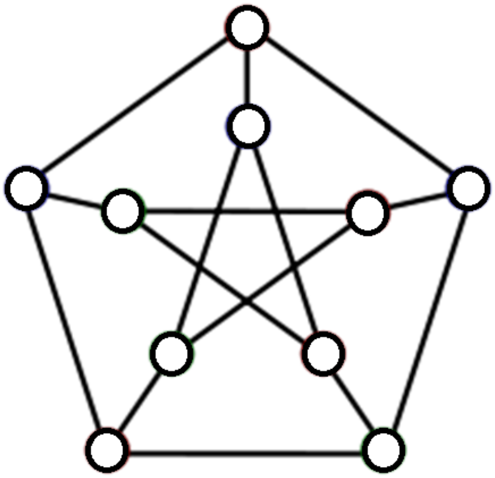


Figure 1: The Petersen graph

Graphs are crucial tools in many fields, the intuitive reason being that many phenomena, while complex, can often be thought about through pairwise interactions between objects (or data points), which can be nicely modeled with the help of a graph.

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.

- 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.

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

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 $w_{ij} = w_{ji}$.

2.1.1 Cliques and Ramsey numbers

Cliques are important structures in graphs and may have important application-specific applications. For example, in a social network graph (e.g., where people correspond to vertices and two vertices are connected if the respective people are friends) cliques have a clear interpretation.

A natural question is whether it is possible to have arbitrarily large graphs without cliques (and without its complement having cliques), Ramsey answer this question in the negative in 1928 [Ram28]. Let us start with some definitions: given a graph G we define $r(G)$ as the size of the largest clique of independence set, i.e.

$$r(G) := \max\left\{c(G), c\left(G^c\right)\right\}.$$

Given r, let $R(r)$ denote the smallest integer n such that every graph G on n nodes must have $r(G) \geq r$. Ramsey [Ram28] showed that $R(r)$ is finite, for every r.

Remark 2.1 *It is easy to show that $R(3) = 6$, try it!*

We will need a simple estimate for what follows (it is a very useful consequence of Stirling's approximation, e.g.).

Proposition 2.2 *For every $k \leq n$ positive integers,*

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{ne}{k}\right)^k.$$

We will show a simple lower bound on $R(r)$. But first we introduce a random graph construction, an Erdős-Renyí graph.

Definition 2.3 *Given n and p, the random Erdős-Renyí graph $G(n,p)$ is a random graph on n vertices where each possible edge appears, independently, with probability p.*

The proof of the lower bound on $R(r)$ is based on the probabilistic method, a beautiful non-constructive method pioneered by Paul Erdős to establish the existence of certain objects. The core idea is the simple observation that if a random variable has a certain expectation then there must exist a draw of it whose value is at least that of the expectation. It is best understood with an example.

Theorem 2.4 *For every $r \geq 2$,*

$$R(r) \geq 2^{\frac{r-1}{2}}.$$

Proof. Let G be drawn from the $G\left(n, \frac{1}{2}\right)$ distribution, $G \sim G\left(n, \frac{1}{2}\right)$. For every set S of r nodes, let $X(S)$ denote the random variable

$$X(S) = \begin{cases} 1 & \text{if } S \text{ is a clique or independent set,} \\ 0 & \text{otherwise.} \end{cases}$$

Also, let X denote the random variable

$$X = \sum_{S \in \binom{V}{r}} X(S).$$

We will proceed by estimating $\mathbb{E}[X]$. Note that, by linearity of expectation,

$$\mathbb{E}[X] = \sum_{S \in \binom{V}{r}} \mathbb{E}[X(S)],$$

and $\mathbb{E}[X(S)] = \text{Prob}\{S \text{ is a clique or independent set}\} = \frac{2}{2^{\binom{|S|}{2}}}$. This means that

$$\mathbb{E}[X] = \sum_{S \in \binom{V}{r}} \frac{2}{2^{\binom{|S|}{2}}} = \binom{n}{r} \frac{2}{2^{\binom{r}{2}}} = \binom{n}{r} \frac{2}{2^{\frac{r(r-1)}{2}}}.$$

By Proposition 2.2 we have,

$$\mathbb{E}[X] \leq \left(\frac{ne}{r}\right)^r \frac{2}{2^{\frac{r(r-1)}{2}}} = 2\left(\frac{n}{2^{\frac{r-1}{2}}} \frac{e}{r}\right)^r.$$

That means that if $n \leq 2^{\frac{r-1}{2}}$ and $r \geq 3$ then $\mathbb{E}[X] < 1$. This means that $\text{Prob}\{X < 1\} > 0$ and since X is a non-negative integer we must have $\text{Prob}\{X = 0\} = \text{Prob}\{X < 1\} > 0$ (another way of saying that is that if $\mathbb{E}[X] < 1$ then there must be an instance for which $X < 1$ and since X is a non-negative integer, we must have $X = 0$). This means that there exists a graph with $2^{\frac{r-1}{2}}$ nodes that does not have cliques or independent sets of size r which implies the theorem. □

Remarkably, this lower bound is not very different from the best known. In fact, the best known lower and upper bounds known [Spe75, Con09] for $R(r)$ are

$$(1 + o(1))\frac{\sqrt{2}r}{e}\left(\sqrt{2}\right)^r \leq R(r) \leq r^{-\frac{c \log r}{\log \log r}} 4^r. \tag{24}$$

Open Problem 2.1 *Recall the definition of $R(r)$ above, the following questions are open:*

- *What is the value of $R(5)$?*
- *What are the asymptotics of $R(s)$? In particular, improve on the base of the exponent on either the lower bound ($\sqrt{2}$) or the upper bound (4).*

- *Construct a family of graphs $G = (V, E)$ with increasing number of vertices for which there exists $\varepsilon > 0$ such that*[9]
$$|V| \lesssim (1+\varepsilon)^r.$$

It is known that $43 \leq R(5) \leq 49$. There is a famous quote in Joel Spencer's book [Spe94] that conveys the difficulty of computing Ramsey numbers:

"*Erdős asks us to imagine an alien force, vastly more powerful than us, landing on Earth and demanding the value of $R(5)$ or they will destroy our planet. In that case, he claims, we should marshal all our computers and all our mathematicians and attempt to find the value. But suppose, instead, that they ask for $R(6)$. In that case, he believes, we should attempt to destroy the aliens.*"

There is an alternative useful way to think about (24), by taking $\log_2$ of each bound and rearranging, we get that
$$\left(\frac{1}{2} + o(1)\right)\log_2 n \leq \min_{G=(V,E),\,|V|=n} r(G) \leq (2+o(1))\log_2 n$$
The current "world record" (see [CZ15, Coh15]) for deterministic construction of families of graphs with small $r(G)$ achieves $r(G) \lesssim 2^{(\log\log|V|)^c}$, for some constant $c > 0$. Note that this is still considerably larger than polylog$|V|$. In contrast, it is very easy for randomized constructions to satisfy $r(G) \leq 2\log_2 n$, as made precise by the folloing theorem.

Theorem 2.5 *Let $G \sim G\left(n, \frac{1}{2}\right)$ be and Erdős-Renyí graph with edge probability $\frac{1}{2}$. Then, with high probability,*[10]
$$R(G) \leq 2\log_2(n).$$

Proof. Given n, we are interested in upper bounding $\operatorname{Prob}\{R(G) \geq \lceil 2\log_2 n\rceil\}$. and we proceed by union bounding (and making use of Proposition 2.2):
$$\begin{aligned}
\operatorname{Prob}\{R(G) \geq \lceil 2\log_2 n\rceil\} &= \operatorname{Prob}\left\{\exists_{S\subset V,\,|S|=\lceil 2\log_2 n\rceil}\ S \text{ is a clique or independent set}\right\} \\
&= \operatorname{Prob}\left\{\bigcup_{S\in\binom{V}{\lceil 2\log_2 n\rceil}} \{S \text{ is a clique or independent set}\}\right\} \\
&\leq \sum_{S\in\binom{V}{\lceil 2\log_2 n\rceil}} \operatorname{Prob}\{S \text{ is a clique or independent set}\} \\
&= \binom{n}{\lceil 2\log_2 n\rceil}\frac{2}{2^{\binom{\lceil 2\log_2 n\rceil}{2}}} \\
&\leq 2\left(\frac{n}{2^{\frac{\lceil 2\log_2 n\rceil - 1}{2}}}\frac{e}{\lceil 2\log_2 n\rceil}\right)^{\lceil 2\log_2 n\rceil} \\
&\leq 2\left(\frac{e\sqrt{2}}{2\log_2 n}\right)^{\lceil 2\log_2 n\rceil} \\
&\lesssim n^{-\Omega(1)}.
\end{aligned}$$

[9] By $a_k \lesssim b_k$ we mean that there exists a constant c such that $a_k \leq c\, b_k$.

[10] We say an event happens with high probability if its probability is $\geq 1 - n^{-\Omega(1)}$.

□

The following is one of the most fascinating conjectures in Graph Theory

Open Problem 2.2 (Erdős-Hajnal Conjecture [EH89]) *Prove or disprove the following:*

For any finite graph H, there exists a constant $\delta_H > 0$ such that any graph on n nodes that does not contain H as a subgraph (is a H-free graph) must have

$$r(G) \gtrsim n^{\delta_H}.$$

It is known that $r(G) \gtrsim \exp\left(c_H\sqrt{\log n}\right)$, for some constant $c_H > 0$ (see [Chu13] for a survey on this conjecture). Note that this lower bound already shows that H-free graphs need to have considerably larger $r(G)$. This is an amazing local to global effect, where imposing a constraint on small groups of vertices are connected (being H-free is a local property) creates extremely large cliques or independence sets (much larger than polylog(n) as in random Erdős-Renyí graphs).

Since we do not know how to deterministically construct graphs with $r(G) \leq \text{polylog} n$, one approach could be to take $G \sim G\left(n, \frac{1}{2}\right)$ and check that indeed it has small clique and independence number. However, finding the largest clique on a graph is known to be NP-hard (meaning that there is no polynomial time algorithm to solve it, provided that the widely believed conjecture $NP \neq P$ holds). That is a worst-case statement and thus it doesn't necessarily mean that it is difficult to find the clique number of random graphs. That being said, the next open problem suggests that this is indeed still difficult.

First let us describe a useful construct. Given n and ω, let us consider a random graph G that consists of taking a graph drawn from $G\left(n, \frac{1}{2}\right)$, picking ω of its nodes (say at random) and adding an edge between every pair of those ω nodes, thus "planting" a clique of size ω. This will create a clique of size ω in G. If $\omega > 2\log_2 n$ this clique is larger than any other clique that was in the graph before planting. This means that, if $\omega > 2\log_2 n$, there is enough information in the graph to find the planted clique. In fact, one can simply look at all subsets of size $2\log_2 n + 1$ and check wether it is clique: if it is a clique then it very likely these vertices belong to the planted clique. However, checking all such subgraphs takes super-polynomial time $\sim n^{\mathcal{O}(\log n)}$. This motivates the natural question of whether this can be done in polynomial time.

Since the degrees of the nodes of a $G\left(n, \frac{1}{2}\right)$ have expected value $\frac{n-1}{2}$ and standard deviation $\sim \sqrt{n}$, if $\omega > c\sqrt{n}$ (for sufficiently large constant c) then the degrees of the nodes involved in the planted clique will have larger degrees and it is easy to detect (and find) the planted clique. Remarkably, there is no known method to work for ω significant smaller than this. There is a quasi-linear time algorithm [DM13] that finds the largest clique, with high probability, as long as $\omega \geq \sqrt{\frac{n}{e}} + o(\sqrt{n})$.[11]

Open Problem 2.3 (The planted clique problem) *Let G be a random graph constructed by taking a $G\left(n, \frac{1}{2}\right)$ and planting a clique of size ω.*

1. *Is there a polynomial time algorithm that is able to find the largest clique of G (with high probability) for $\omega \ll \sqrt{n}$. For example, for $\omega \approx \frac{\sqrt{n}}{\log n}$.*

[11] There is an amplification technique that allows one to find the largest clique for $\omega \approx c\sqrt{n}$ for arbitrarily small c in polynomial time, where the exponent in the runtime depends on c. The rough idea is to consider all subsets of a certain finite size and checking whether the planted clique contains them.

2. *Is there a polynomial time algorithm that is able to distinguish, with high probability, G from a draw of $G\left(n, \frac{1}{2}\right)$ for $\omega \ll \sqrt{n}$. For example, for $\omega \approx \frac{\sqrt{n}}{\log n}$.*
3. *Is there a quasi-linear time algorithm able to find the largest clique of G (with high probability) for $\omega \leq \left(\frac{1}{\sqrt{e}} - \varepsilon\right)\sqrt{n}$, for some $\varepsilon > 0$.*

This open problem is particularly important. In fact, the hypothesis that finding planted cliques for small values of ω is behind several cryptographic protocols, and hardness results in average case complexity (hardness for Sparse PCA being a great example [BR13]).

2.2 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 $\mathbb{R}^d$. As we will see below, oftentimes when we have a set of data points $x_1, \dots, x_n \in \mathbb{R}^p$ it will be beneficial to first associate to each a graph and then use Diffusion Maps to represent the points in d-dimensions, rather than using something like Principal Component Analysis.

Before presenting Diffusion Maps, we'll introduce a few important notions. Given $G = (V, E, W)$ we consider a random walk (with independent steps) on the vertices of V with transition probabilities:

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

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

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

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

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

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

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

Remark 2.6 *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, $S = D^{\frac{1}{2}}MD^{-\frac{1}{2}}$ is, indeed $S = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$. We consider the spectral decomposition of S

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

where $V = [v_1, \dots, v_n]$ satisfies $V^TV = I_{n\times n}$ 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,

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

We define $\Phi = D^{-\frac{1}{2}}V$ with columns $\Phi = [\varphi_1, \dots, \varphi_n]$ and $\Psi = D^{\frac{1}{2}}V$ with columns $\Psi = [\psi_1, \dots, \psi_n]$. 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 $1 \leq k \leq n$:

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

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

$$M^t = \sum_{k=1}^{n} \lambda_k^t\varphi_k\psi_k^T. \tag{25}$$

Let's revisit the embedding suggested on Remark 2.6. 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

$$v_i \to M^t[i,:] = \begin{bmatrix} \lambda_1^t\varphi_1(i) \\ \lambda_2^t\varphi_2(i) \\ \vdots \\ \lambda_n^t\varphi_n(i) \end{bmatrix}, \tag{26}$$

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 2.7 *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 $\varphi_k\left(i_{\max}\right)$. Then,

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

This means, by triangular inequality that, that

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

□

Remark 2.8 *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 (26) does not help differentiate points on the graph. This suggests removing that coordinate:

Definition 2.9 (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 $\phi_t : V \to \mathbb{R}^{n-1}$ given by*

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

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 2.10 (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 $\phi_t : V \to \mathbb{R}^d$ given by*

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

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 2.11 *For any pair of nodes v_{i_1}, v_{i_2} we have*

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

Proof.

Note that $\sum_{j=1}^{n} \frac{1}{\deg(j)} \left[\text{Prob}\{X(t)=j|X(0)=i_1\} - \text{Prob}\{X(t)=j|X(0)=i_2\}\right]^2$ 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

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

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

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

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

2.2.1 A couple of examples

The ring graph is a graph on n nodes $\{1, \dots, n\}$ such that node k is connected to $k-1$ and $k+1$ and 1 is connected to n. Figure 2 has the Diffusion Map of it truncated to two dimensions

Another simple graph is K_n, the complete graph on n nodes (where every pair of nodes share an edge), see Figure 3.

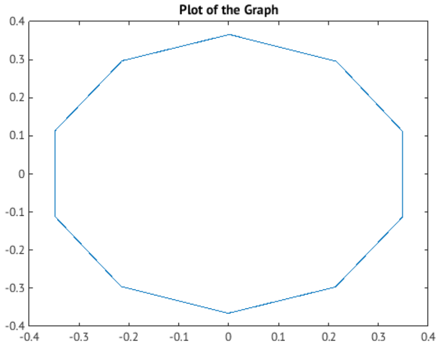


Figure 2: The Diffusion Map of the ring graph gives a very natural way of displaying (indeed, if one is asked to draw the ring graph, this is probably the drawing that most people would do). It is actually not difficult to analytically compute the Diffusion Map of this graph and confirm that it displays the points in a circle.

2.2.2 Diffusion Maps of point clouds

Very often we are interested in embedding in $\mathbb{R}^d$ a point cloud of points $x_1, \dots, x_n \in \mathbb{R}^p$ and necessarily a graph. One option (as discussed before in the course) is 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's say 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's say that we are given a point cloud that is sampled from a two dimensional swiss roll embedded in three dimension (see Figure 4). In order to learn the two dimensional structure of this object we need to differentiate points that are near eachother 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 4 for an example). We will achieve this by creating a graph from the data 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.

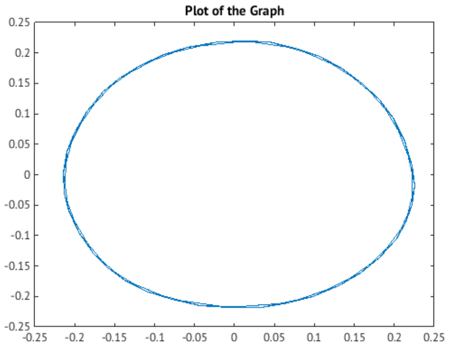


Figure 3: The Diffusion Map of the complete graph on 4 nodes in 3 dimensions appears to be a regular tetrahedron suggesting that there is no low dimensional structure in this graph. This is not surprising, since every pair of nodes is connected we don't expect this graph to have a natural representation in low dimensions.

2.2.3 A 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!

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 the arcade game Asteroids, 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.

For the two dimensional example, we expect the structure of the underlying manifold to be a two-dimensional torus. Indeed, Figure 6 shows that the three-dimensional diffusion map captures the toroidal structure of the data.

2.2.4 Similar non-linear dimensional reduction techniques

There are several other similar non-linear dimensional reduction methods. A particularly popular one is ISOMAP [**?**]. 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 pairs of nodes v_i, v_j finding their geodesic distance and then using, for example, Multidimensional

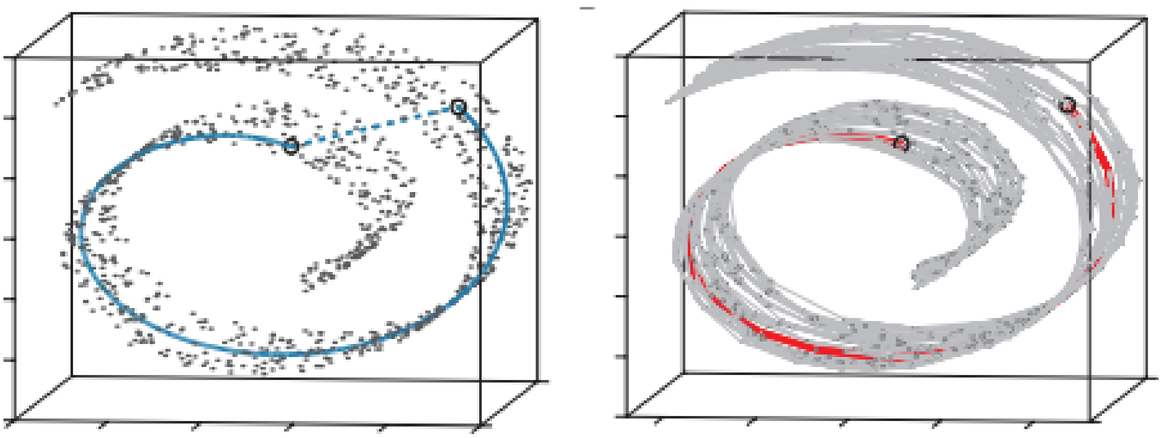


Figure 4: A swiss roll point cloud (see, for example, [TdSL00]). The points are sampled from a two dimensional manifold curved in $\mathbb{R}^3$ and then a graph is constructed where nodes correspond to points.

Scaling to find points $y_i \in \mathbb{R}^d$ that minimize (say)

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

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

2.3 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's say that there are two labels, $\{-1, +1\}$.

Let's say we are given the task of labeling point "?" in Figure 10 given the labeled points. The natural label to give to the unlabeled point would be 1.

However, let's say that we are given not just one unlabeled point, but many, as in Figure 11; then it starts being apparent that -1 is a more reasonable guess.

Intuitively, the unlabeled data points allowed us to better learn the geometry of the dataset. That's the idea behind Semi-supervised learning, to make use of the fact that often one has access to many unlabeled data points in order to improve classification.

The approach we'll take is to 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 $f_1, \dots, f_l$, and u the number of unlabeled points (with $n = l + u$), the first l nodes $v_1, \dots, v_l$ correspond to labeled points and the rest $v_{l+1}, \dots, v_n$ are unlabaled. We want to find a function $f : V \to \{-1, 1\}$ that agrees on labeled points: $f(i) = f_i$ for $i = 1, \dots, l$ and that is "as smooth as possible" the graph. A way to pose this is the following

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

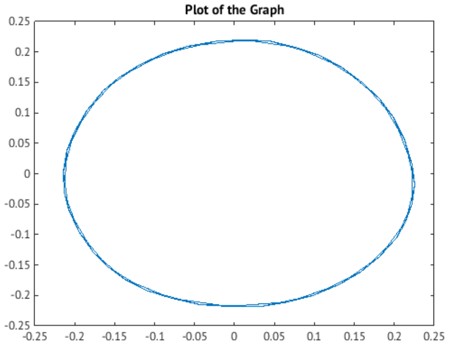


Figure 5: The two-dimensional diffusion map of the dataset of the datase where each data point is an image with the same vertical strip in different positions in the x-axis, the circular structure is apparent.

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 we are can rewrite the problem as

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

The matrix $\sum_{i<j} w_{ij}\left(e_i-e_j\right)\left(e_i-e_j\right)^T$ will play a central role throughout this course, it is called the graph Laplacian [Chu97].

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

Note that the entries of L_G are given by

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

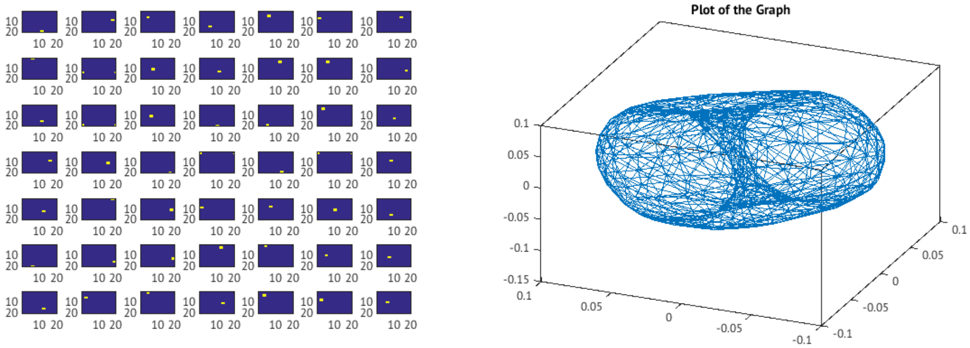


Figure 6: On the left the data set considered and on the right its three dimensional diffusion map, the fact that the manifold is a torus is remarkably captured by the embedding.

meaning that

$$L_G = D - W,$$

where D is the diagonal matrix with entries $D_{ii} = \deg(i)$.

Remark 2.12 *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.

Let us consider our problem

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

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 $W_{ul} = W_{lu}^T$)

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

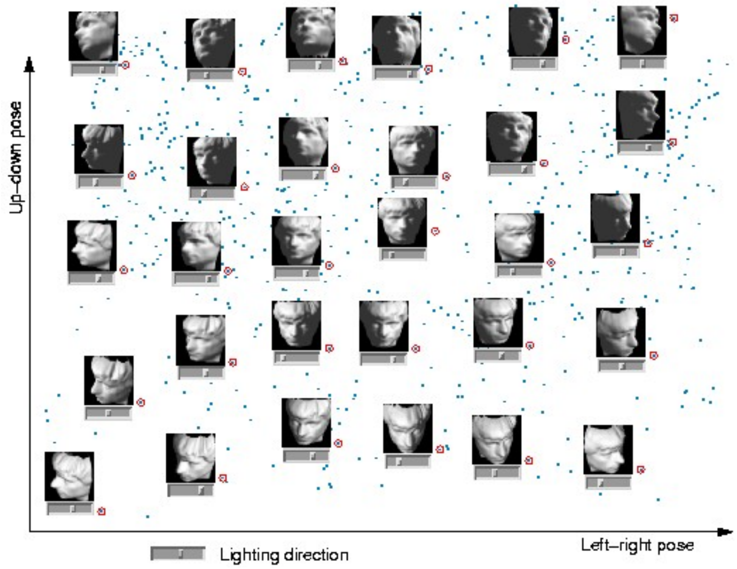


Figure 7: The two dimensional represention of a data set of images of faces as obtained in [TdSL00] using ISOMAP. Remarkably, the two dimensionals are interpretable

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

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

If $D_u - W_{uu}$ is invertible[12] then

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

Remark 2.13 *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.

2.3.1 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

[12] 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.

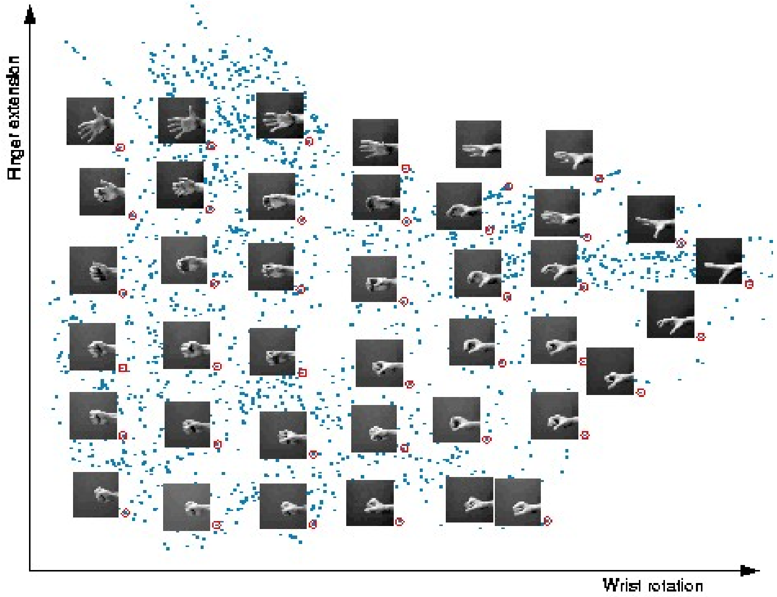


Figure 8: The two dimensional represention of a data set of images of human hand as obtained in [TdSL00] using ISOMAP. Remarkably, the two dimensionals are interpretable

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 12, $d = 2$ in Figure 13, and $d = 3$ in Figure 14. While for $d \leq 2$ it appears to be smoothly interpolating between 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 14.

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

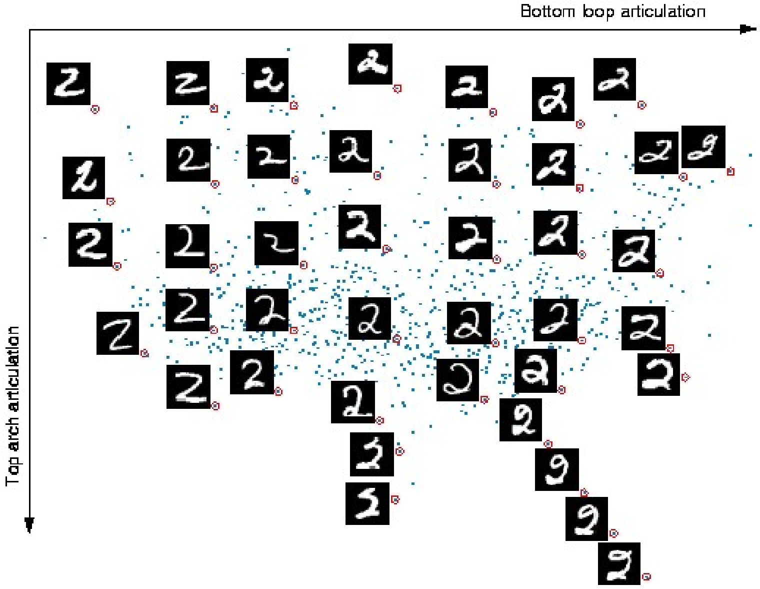


Figure 9: The two dimensional represention of a data set of handwritten digits as obtained in [TdSL00] using ISOMAP. Remarkably, the two dimensionals are interpretable

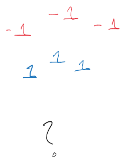


Figure 10: Given a few labeled points, the task is to label an unlabeled point.

out the behavior observed in Figure 14. 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 15 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., [NSZ09] for more on this phenomenon).

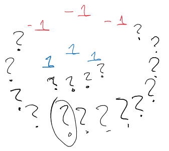


Figure 11: In this example we are given many unlabeled points, the unlabeled points help us learn the geometry of the data.

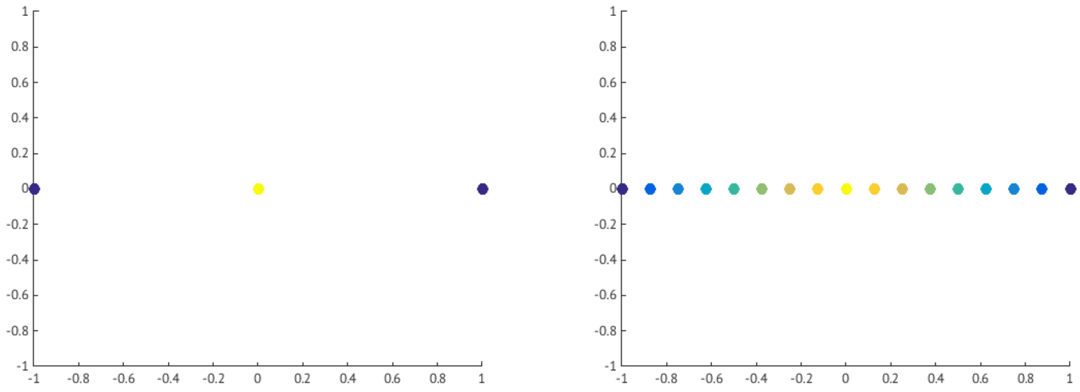


Figure 12: The $d = 1$ example of the use of this method to the example described above, the value of the nodes is given by color coding. For $d = 1$ it appears to smoothly interpolate between the labeled points.

3 Spectral Clustering and Cheeger's Inequality

3.1 Clustering

Clustering is one of the central tasks in machine learning. Given a set of data points, 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).

3.1.1 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:

$$\min_{\substack{\text{partition } S_1,\ldots,S_k \\ \mu_1,\ldots,\mu_k}} \sum_{l=1}^{k} \sum_{i \in S_i} \left\| x_i - \mu_l \right\|^2 . \tag{27}$$

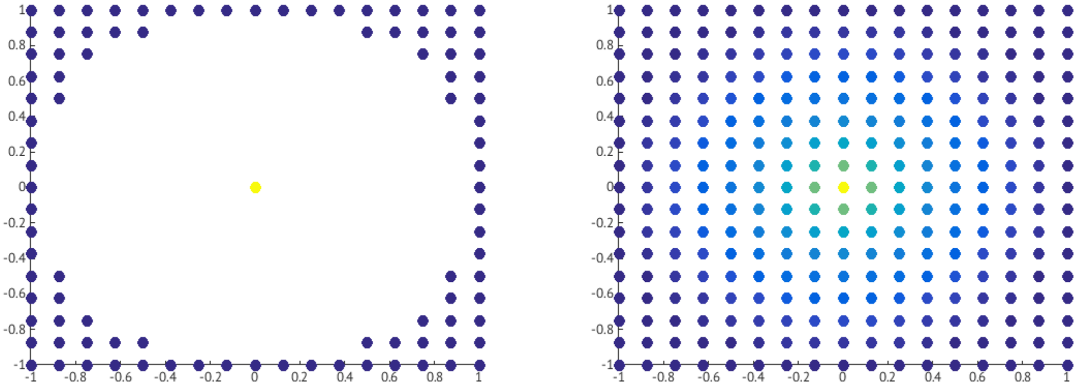


Figure 13: The $d = 2$ example of the use of this method to the example described above, the value of the nodes is given by color coding. For $d = 2$ it appears to smoothly interpolate between the labeled points.

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 [Llo82] (also known as the k-means algorithm), is an iterative algorithm that alternates between

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

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

- 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 (27). Indeed, Lloyd's algorithm oftentimes gets stuck in local optima of (27). A few lectures from now we'll discuss convex relaxations for clustering, which can be used as an alternative algorithmic approach to Lloyd's algorithm, but since optimizing (27) is NP-hard there is not polynomial time algorithm that works in the worst-case (assuming the widely believed conjecture $P \neq NP$)

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 (27) 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 (27) in terms of distances only).

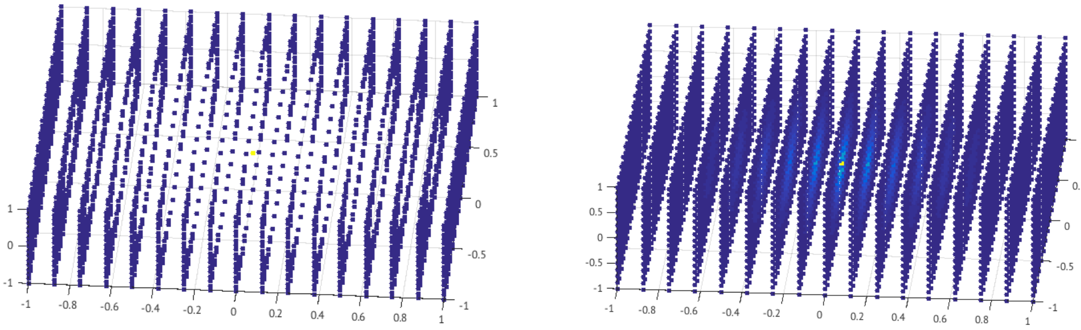


Figure 14: The $d = 3$ example of the use of this method to the example described above, the value of the nodes is given by color coding. For $d = 3$ the solution appears to only learn the label -1.

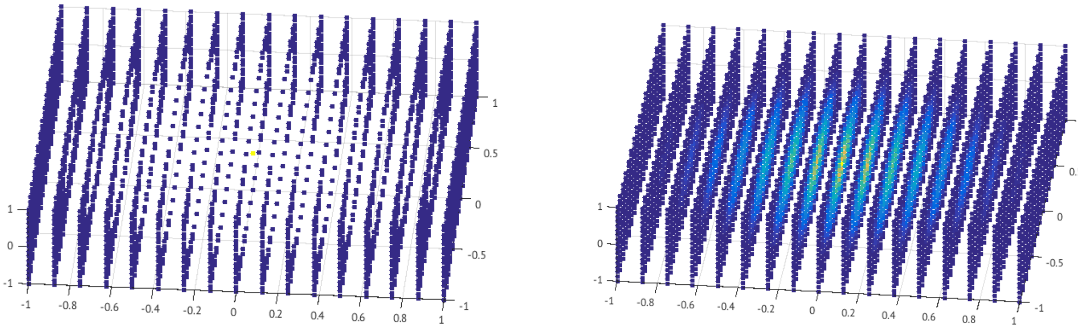


Figure 15: The $d = 3$ example of the use of this method with the extra regularization $f^T L^2 f$ to the example described above, the value of the nodes is given by color coding. The extra regularization seems to fix the issue of discontinuities.

- The formulation is computationally hard, so algorithms may produce suboptimal 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 17.

3.2 Spectral Clustering

A natural way to try to overcome the issues of k-means depicted in Figure 17 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 $K_\epsilon(u) = \exp\left(\frac{1}{2\epsilon}u^2\right)$, by associating each point to a vertex and, for which pair of nodes, set the edge weight as

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

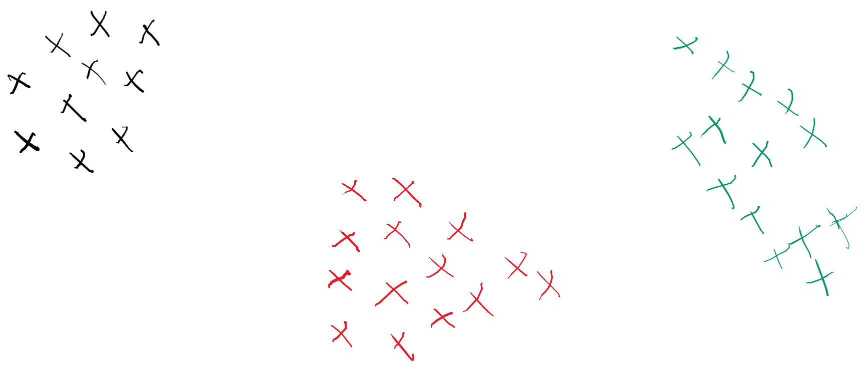


Figure 16: Examples of points separated in clusters.

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

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

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

$$\phi_t^{(d)}(i) = \left[\begin{array}{c} \lambda_2^t \varphi_2(i) \\ \vdots \\ \lambda_{d+1}^t \varphi_{d+1}(i) \end{array} \right],$$

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 3.1 (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*

$$\phi_t^{(k-1)}(i) = \left[\begin{array}{c} \lambda_2^t \varphi_2(i) \\ \vdots \\ \lambda_k^t \varphi_k(i) \end{array} \right]$$

and clustering the points $\phi_t^{(k-1)}(1), \phi_t^{(k-1)}(2), \ldots, \phi_t^{(k-1)}(n) \in \mathbb{R}^{k-1}$ using, for example, k-means clustering.

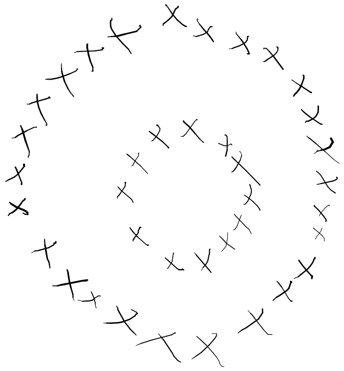


Figure 17: Because the solutions of k-means are always convex clusters, it is not able to handle some cluster structures.

3.3 Two clusters

We will mostly focus in the case of two cluster ($k = 2$). For $k = 2$, Algorithm 3.1 consists in assigning to each vertex i a real number $\varphi_2(i)$ and then clustering the points in the real line. Note in $\mathbb{R}$, clustering reduces to setting a threshold τ and taking $S = \{i \in V : \varphi_2(i) \leq \tau\}$. Also, it is computationally tractable to try all possible thresholds (there are $\leq n$ different possibilities).

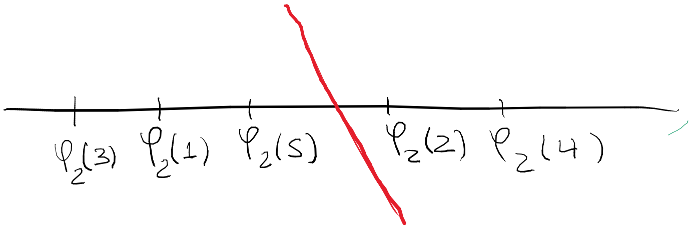


Figure 18: For two clusters, spectral clustering consists in assigning to each vertex i a real number $\varphi_2(i)$, then setting a threshold τ and taking $S = \{i \in V : \varphi_2(i) \leq \tau\}$.

Algorithm 3.2 (Spectral Clustering for two clusters) *Given a graph $G = (V, E, W)$, consider the two-dimensional Diffusion Map*

$$i \to \varphi_2(i).$$

set a threshold τ (one can try all different possibilities) and set

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

In what follows we'll give a different motivation for Algorithm 3.2.

3.3.1 Normalized Cut

Given a graph $G = (V, E, W)$, a natural measure to measure a vertex partition (S, S^c) is

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

Note however that the minimum cut is achieved for $S = \emptyset$ (since $\operatorname{cut}(\emptyset) = 0$) which is a rather meaningless choice of partition.

Remark 3.3 *One way to circumvent this issue is to ask that $|S| = |S^c|$ (let's say that the number of vertices $n = |V|$ is even), corresponding to a balanced partition. 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*

$$\min_{\substack{S \subset V \\ |S| = |S^c|}} \operatorname{cut}(S) = \min_{\substack{y \in \{-1,1\}^n \\ \mathbf{1}^T y = 0}} \frac{1}{4} \sum_{i \leq j} w_{ij} \left(y_i - y_j\right)^2 = \frac{1}{4} \min_{\substack{y \in \{-1,1\}^n \\ \mathbf{1}^T y = 0}} y^T L_G y,$$

where $L_G = D - W$ is the graph Laplacian.[13].

Since asking for the partition to be balanced is too restrictive in many cases, there are several ways to evaluate a partition that are variations of $\operatorname{cut}(S)$ 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 3.4 (Cheeger's cut) *Given a graph and a vertex partition (S, S^c), the cheeger cut (also known as conductance, and sometimes expansion) of S is given by*

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

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

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

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

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

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

[13] W is the matrix of weights and D the degree matrix, a diagonal matrix with diagonal entries $D_{ii} = \deg(i)$.

Both $h(S)$ and $\mathrm{Ncut}(S)$ favor nearly balanced partitions, Proposition 3.5 below will give an interpretation of Ncut via random walks.

Let us recall the construction form previous lectures of a random walk on $G = (V, E, W)$:

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

where $M = D^{-1}W$. Recall that $M = \Phi\Lambda\Psi^T$ where Λ is the diagonal matrix with the eigenvalues λ_k of M and Φ and Ψ form a biorthogonal system $\Phi^T\Psi = I$ and correspond to, respectively, the right and left eigenvectors of M. Moreover they are given by $\Phi = D^{-\frac{1}{2}}V$ and $\Psi = D^{\frac{1}{2}}V$ where $V^TV = I$ and $D^{-\frac{1}{2}}WD^{-\frac{1}{2}} = V\Lambda V^T$ is the spectral decomposition of $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$.

Recall also that $M\mathbf{1} = \mathbf{1}$, corresponding to $M\varphi_1 = \varphi_1$, which means that $\psi_1^T M = \psi_1^T$, where

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

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 3.5 *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:*

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

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

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:

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

Analogously,

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

which concludes the proof. □

3.3.2 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.

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 (say $y_i = a$ 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

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

which corresponds to $\mathbf{1}^T D y = 0$.

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

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

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 3.6 *For a and b to satisfy $a \operatorname{vol}(S) + b \operatorname{vol}(S^c) = 0$ and $a^2 \operatorname{vol}(S) + b^2 \operatorname{vol}(S^c) = 1$ it must be that*

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

corresponding to

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

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

Proposition 3.7

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

where y is given by

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

Proof.

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

□

This means that finding the minimum Ncut corresponds to solving

$$
\begin{array}{ll}
\min & y^T L_G y \\
\text{s. t.} & y \in \{a,b\}^n \text{ for some } a \text{ and } b \\
& y^T D y = 1 \\
& y^T D\mathbf{1} = 0.
\end{array}
\tag{28}
$$

Since solving (28) is, in general, NP-hard, we consider a similar problem where the constraint that y can only take two values is removed:

$$
\begin{array}{ll}
\min & y^T L_G y \\
\text{s. t.} & y \in \mathbb{R}^n \\
& y^T D y = 1 \\
& y^T D\mathbf{1} = 0.
\end{array}
\tag{29}
$$

Given a solution of (29) 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 (29) is an eigenvector problem (for this reason we call (29) a spectral relaxation) and, moreover, that the solution corresponds to y a multiple of φ_2 meaning that this approach corresponds exactly to Algorithm 3.2.

In order to better see that (29) is an eigenvector problem (and thus computationally tractable), set $z = D^{\frac{1}{2}}y$ and $\mathcal{L}_G = D^{-\frac{1}{2}}L_G D^{-\frac{1}{2}}$, then (29) is equivalent

$$\begin{array}{rl} \min & z^T \mathcal{L}_G z \\ \text{s. t.} & z \in \mathbb{R}^n \\ & \|z\|^2 = 1 \\ & \left(D^{\frac{1}{2}}\mathbf{1}\right)^T z = 0. \end{array} \tag{30}$$

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 (30) 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 is the second largest eigenvector of $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ which we know is v_2. This means that the optimal y in (29) is given by $\varphi_2 = D^{-\frac{1}{2}} v_2$. This confirms that this approach is equivalent to Algorithm 3.2.

Because the relaxation (29) is obtained from (28) by removing a constraint we immediately have that

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

This means that

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

In what follows we will show a guarantee for Algorithm 3.2.

Lemma 3.8 *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 3.2 is suboptimal at most by a square root factor.

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

Theorem 3.9 (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 [Che70], the graph version is due to Noga Alon and Vitaly Milman [Alo86, AM85] in the mid 80s.

The upper bound in Cheeger's inequality (corresponding to Lemma 3.8) 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 [Chu10] for four different proofs!). The proof that follows is an adaptation of the proof in this blog post [Tre11] for the case of weighted graphs.

Proof. [of Lemma 3.8]

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 $\delta = \lambda_2\left(\mathcal{L}_G\right)$ 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 $y^D\mathbf{1} = 0$ 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

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

and

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

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

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

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

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

It is not difficult to check that

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

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

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

Note that $\operatorname{Prob}\{(S, S^c) \text{ cuts the edge } (i,j)\}$ 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 $|x_i - x_j|\left(|x_i| + |x_j|\right)$. This means that

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

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,

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

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,

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

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

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

However, since both random variables are positive,

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

or equivalently

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

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

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

which is equivalent to

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

which concludes the proof of the Lemma. □

3.4 Small Clusters and the Small Set Expansion Hypothesis

We now restrict to unweighted regular graphs $G = (V, E)$.

Cheeger's inequality allows to efficiently approximate its Cheeger number up to a square root factor. It means in particular that, given $G = (V, E)$ and ϕ we can efficiently between the cases where: $h_G \leq \phi$ or $h_G \geq 2\sqrt{\phi}$. Can this be improved?

Open Problem 3.1 *Does there exists a constant $c > 0$ such that it is NP-hard to, given ϕ, and G distinguis between the cases*

1. *$h_G \leq \phi$, and*
2. *$h_G \geq c\sqrt{\phi}$?*

It turns out that this is a consequence [RST12] of an important conjecture in Theoretical Computer Science (see [BS14] for a nice description of it). This conjecture is known [RS10] to imply the Unique-Games Conjecture [Kho10], that we will discuss in future lectures.

Conjecture 3.10 (Small-Set Expansion Hypothesis [RS10]) *For every $\epsilon > 0$ there exists $\delta > 0$ such that it is NP-hard to distinguish between the cases*

1. *There exists a subset $S \subset V$ with $\operatorname{vol}(S) = \delta \operatorname{vol}(V)$ such that $\frac{\operatorname{cut}(S)}{\operatorname{vol}(S)} \leq \epsilon$,*
2. *$\frac{\operatorname{cut}(S)}{\operatorname{vol}(S)} \geq 1 - \epsilon$, for every $S \subset V$ satisfying $\operatorname{vol}(S) \leq \delta \operatorname{vol}(V)$.*

3.5 Computing Eigenvectors

Spectral clustering requires us to compute the second smallest eigenvalue of $\mathcal{L}_G$. One of the most efficient ways of computing eigenvectors is through the power method. For simplicity we'll consider the case on which we are computing the leading eigenvector of a matrix $A \in \mathbb{R}^{n \times n}$ with m non-zero entries, for which $|\lambda_{\max}(A)| \geq |\lambda_{\min}(A)|$ (the idea is easily adaptable). The power method proceeds by starting with a guess y^0 and taking iterates $y^{t+1} = \frac{Ay^t}{\|Ay^t\|}$. One can show [KW92] that the variantes of the power method can find a vector x in randomized time $\mathcal{O}\left(\delta^{-1}(m+n)\log n\right)$ satisfying $x^T Ax \geq \lambda_{\max}(A)(1-\delta)x^T x$. Meaning that an approximate solution can be found in quasi-linear time.[14]

One drawback of the power method is that when using it, one cannot be sure, a posteriori, that there is no eigenvalue of A much larger than what we have found, since it could happen that all our guesses were orthogonal to the corresponding eigenvector. It simply guarantees us that if such an eigenvalue existed, it would have been extremely likely that the power method would have found it. This issue is addressed in the open problem below.

Open Problem 3.2 *Given a symmetric matrix M with small condition number, is there a quasi-linear time (on n and the number of non-zero entries of M) procedure that certifies that $M \succeq 0$. More specifically, the procedure can be randomized in the sense that it may, with some probably not certify that $M \succeq 0$ even if that is the case, what is important is that it never produces erroneous certificates (and that it has a bounded-away-from-zero probably of succeeding, provided that $M \succeq 0$).*

[14] Note that, in spectral clustering, an error on the calculation of φ_2 propagates gracefully to the guarantee given by Cheeger's inequality.

The Cholesky decomposition produces such certificates, but we do not know how to compute it in quasi-linear time. Note also that the power method can be used in $\alpha I - M$ to produce certificates that have arbitrarily small probability of being false certificates. Later in these lecture we will discuss the practical relevance of such a method as a tool to quickly certify solution produced by heuristics [Ban16].

3.6 Multiple Clusters

Given a graph $G = (V, E, W)$, a natural way of evaluating k-way clusterign is via the k-way expansion constant (see [LGT12]):

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

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

Another natural definition is

$$\varphi_G(k) = \min_{S:\operatorname{vol} S \le \frac{1}{k}\operatorname{vol}(G)} \frac{\operatorname{cut}(S)}{\operatorname{vol}(S)}.$$

It is easy to see that

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

The following is known.

Theorem 3.11 ([LGT12]) *Let $G = (V, E, W)$ be a graph and k a positive integer*

$$\rho_G(k) \le \mathcal{O}\left(k^2\right)\sqrt{\lambda_k}, \tag{31}$$

Also,

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

Open Problem 3.3 *Let $G = (V, E, W)$ be a graph and k a positive integer, is the following true?*

$$\rho_G(k) \le \operatorname{polylog}(k)\sqrt{\lambda_k}. \tag{32}$$

We note that (32) is known not to hold if we ask that the subsets form a partition (meaning that every vertex belongs to at least one of the sets) [LRTV12]. Note also that no dependency on k would contradict the Small-Set Expansion Hypothesis above.

4 Concentration Inequalities, Scalar and Matrix Versions

4.1 Large Deviation Inequalities

Concentration and large deviations inequalities are among the most useful tools when understanding the performance of some algorithms. In a nutshell they control the probability of a random variable being very far from its expectation.

The simplest such inequality is Markov's inequality:

Theorem 4.1 (Markov's Inequality) *Let $X \geq 0$ be a non-negative random variable with $\mathbb{E}[X] < \infty$. Then,*

$$\operatorname{Prob}\{X > t\} \leq \frac{\mathbb{E}[X]}{t}. \tag{33}$$

Proof. Let $t > 0$. Define a random variable Y_t as

$$Y_t = \begin{cases} 0 & \text{if } X \leq t \\ t & \text{if } X > t \end{cases}$$

Clearly, $Y_t \leq X$, hence $\mathbb{E}[Y_t] \leq \mathbb{E}[X]$, and

$$t \operatorname{Prob}\{X > t\} = \mathbb{E}[Y_t] \leq \mathbb{E}[X],$$

concluding the proof. □

Markov's inequality can be used to obtain many more concentration inequalities. Chebyshev's inequality is a simple inequality that control fluctuations from the mean.

Theorem 4.2 (Chebyshev's inequality) *Let X be a random variable with $\mathbb{E}[X^2] < \infty$. Then,*

$$\operatorname{Prob}\{|X - \mathbb{E}X| > t\} \leq \frac{\operatorname{Var}(X)}{t^2}.$$

Proof. Apply Markov's inequality to the random variable $(X - \mathbb{E}[X])^2$ to get:

$$\operatorname{Prob}\{|X - \mathbb{E}X| > t\} = \operatorname{Prob}\{(X - \mathbb{E}X)^2 > t^2\} \leq \frac{\mathbb{E}\left[(X - \mathbb{E}X)^2\right]}{t^2} = \frac{\operatorname{Var}(X)}{t^2}.$$

□

4.1.1 Sums of independent random variables

In what follows we'll show two useful inequalities involving sums of independent random variables. The intuitive idea is that if we have a sum of independent random variables

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

where X_i are iid 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 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 quadratic, which will be crucial in many applications to come.

Theorem 4.3 (Hoeffding's Inequality) *Let $X_1, X_2, \ldots, X_n$ be independent bounded random variables, i.e., $|X_i| \leq a$ and $\mathbb{E}[X_i] = 0$. Then,*

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

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

Proof. 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.

$$\begin{aligned}
\operatorname{Prob}\left\{\sum_{i=1}^n X_i > t\right\} &= \operatorname{Prob}\left\{\sum_{i=1}^n X_i > t\right\} && (34)\\
&= \operatorname{Prob}\left\{e^{\lambda\sum_{i=1}^n X_i} > e^{\lambda t}\right\}\\
&\leq \frac{\mathbb{E}[e^{\lambda\sum_{i=1}^n X_i}]}{e^{t\lambda}}\\
&= e^{-t\lambda}\prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}], && (35)
\end{aligned}$$

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[15]

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

Hence,

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

Together with (34), this gives

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

[15] 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 $(2n)! \geq 2^n n!$.

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

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

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,

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

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

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

□

Remark 4.4 *Let's say that we have random variables $r_1, \dots, r_n$ 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:

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

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

A natural way to quantify this intuition is by noting that the variance of $\sum_{i=1}^{n} r_i$ depends on p as $\text{Var}(r_i) = p$. 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 (35) 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 $k \geq 2$, $\mathbb{E}X_i^k \leq \sigma^2 a^{k-2} = \left(\frac{\sigma^2}{a^2}\right)a^k$.

Theorem 4.5 (Bernstein's Inequality) *Let $X_1, X_2, \ldots, X_n$ be independent centered bounded random variables, i.e., $|X_i| \le a$ and $\mathbb{E}[X_i] = 0$, with variance $\mathbb{E}[X_i^2] = \sigma^2$. Then,*

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

Remark 4.6 *Before proving Bernstein's Inequality, note that on the example of Remark 4.4 we get*

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

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

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

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

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

Now comes the source of the improvement over Hoeffding's,

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

Therefore,

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

We will use a few simple inequalities (that can be easily proved with calculus) such as[16] $1+x \leq e^x$, for all $x \in \mathbb{R}$.

This means that,

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

which readily implies

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

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

$$u = \frac{at}{n\sigma^2}, \tag{36}$$

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

Now, the value of the minimum is given by

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

Which means that,

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

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

$$(1+u)\log(1+u) - u \geq \frac{u}{\frac{2}{u} + \frac{2}{3}}, \tag{37}$$

which implies:

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

□

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

4.2 Gaussian Concentration

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 these lectures. 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 4.7 (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 $t \geq 0$*

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

For the sake of simplicity we will show the proof for a slightly weaker bound (in terms of the constant inside the exponent): $\operatorname{Prob}\left\{|F(X) - \mathbb{E}F(X)| \geq t\right\} \leq 2\exp\left(-\frac{2}{\pi^2}\frac{t^2}{\sigma^2}\right)$. This exposition follows closely the proof of Theorem 2.1.12 in [Tao12] and the original argument is due to Maurey and Pisier. For a proof with the optimal constants see, for example, Theorem 3.25 in these notes [vH14]. We will also assume 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 to F, we can assume that $\mathbb{E}F(X) = 0$. Also, it is enough to show a one-sided bound

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

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

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

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. Also, its derivative with respect to θ, $X'_\theta = -Y\sin\theta + X\cos\theta$ also is. 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 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 in 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

$$\text{Prob}\left\{F(X)\geq t\right\} \leq \exp\left[\frac{1}{2}\left(\lambda\frac{\pi}{2}\sigma\right)^2 - \lambda t\right],$$

Optimizing for λ gives $\lambda^* = \left(\frac{2}{\pi}\right)^2\frac{t}{\sigma^2}$, which gives

$$\text{Prob}\left\{F(X)\geq t\right\} \leq \exp\left[-\frac{2}{\pi^2}\frac{t^2}{\sigma^2}\right].$$

□

4.2.1 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 4.7) we immediately get

$$\mathrm{Prob}\left\{ \|W\| \geq \mathbb{E}\|W\| + t \right\} \leq \exp\left(-\frac{t^2}{4} \right).$$

Since[17] $\mathbb{E}\|W\| \leq 2\sqrt{n}$ we get

Proposition 4.8 *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,*

$$\mathrm{Prob}\left\{ \|W\| \geq 2\sqrt{n} + t \right\} \leq \exp\left(-\frac{t^2}{4} \right).$$

Note that this gives an extremely precise control of the fluctuations of $\|W\|$. In fact, for $t = 2\sqrt{\log n}$ this gives

$$\mathrm{Prob}\left\{ \|W\| \geq 2\sqrt{n} + 2\sqrt{\log n} \right\} \leq \exp\left(-\frac{4\log n}{4} \right) = \frac{1}{n}.$$

4.2.2 Talagrand's concentration inequality

A remarkable result by Talagrand [Tal95], Talangrad's concentration inequality, provides an analogue of Gaussian concentration to bounded random variables.

Theorem 4.9 (Talangrand concentration inequality, Theorem 2.1.13 [Tao12]) *Let $K > 0$, and let $X_1, \ldots, X_n$ be independent bounded random variables, $|X_i| \leq K$ 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 $t \geq 0$,*

$$\mathrm{Prob}\left\{ |F(X) - \mathbb{E}\left[F(X)\right]| \geq tK \right\} \leq c_1 \exp\left(-c_2 \frac{t^2}{\sigma^2} \right),$$

for positive constants c_1, and c_2.

Other useful similar inequalities (with explicit constants) are available in [Mas00].

[17] It is an excellent exercise to prove $\mathbb{E}\|W\| \leq 2\sqrt{n}$ using Slepian's inequality.

4.3 Other useful large deviation inequalities

This Section contains, without proof, some scalar large deviation inequalities that I have found useful.

4.3.1 Additive Chernoff Bound

The additive Chernoff bound, also known as Chernoff-Hoeffding theorem concerns Bernoulli random variables.

Theorem 4.10 *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 $\varepsilon > 0$:*

- $\operatorname{Prob}\left\{\frac{1}{n}\sum_{i=1}^n X_i \geq p+\varepsilon\right\} \leq \left[\left(\frac{p}{p+\varepsilon}\right)^{p+\varepsilon}\left(\frac{1-p}{1-p-\varepsilon}\right)^{1-p-\varepsilon}\right]^n$
- $\operatorname{Prob}\left\{\frac{1}{n}\sum_{i=1}^n X_i \leq p-\varepsilon\right\} \leq \left[\left(\frac{p}{p-\varepsilon}\right)^{p-\varepsilon}\left(\frac{1-p}{1-p+\varepsilon}\right)^{1-p+\varepsilon}\right]^n$

4.3.2 Multiplicative Chernoff Bound

There is also a multiplicative version (see, for example Lemma 2.3.3. in [Dur06]), which is particularly useful.

Theorem 4.11 *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 $\delta > 0$:*

- $\operatorname{Prob}\{X > (1+\delta)\mu\} < \left[\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right]^\mu$
- $\operatorname{Prob}\{X < (1-\delta)\mu\} < \left[\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right]^\mu$

4.3.3 Deviation bounds on χ_2 variables

A particularly useful deviation inequality is Lemma 1 in Laurent and Massart [LM00]:

Theorem 4.12 (Lemma 1 in Laurent and Massart [LM00]) *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 $t > 0$:

- $\operatorname{Prob}\{Z \geq 2\|a\|_2\sqrt{x} + 2\|a\|_\infty x\} \leq \exp(-x)$,

- Prob $\{Z \leq -2\|a\|_2\sqrt{x}\} \leq \exp(-x)$,

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 with n degrees of freedom, so this theorem immediately gives a deviation inequality for χ_2 random variables.

4.4 Matrix Concentration

In many important applications, some of which we will see in the proceeding lectures, one needs to use a matrix version of the inequalities above.

Given $\{X_k\}_{k=1}^n$ independent random symmetric $d \times d$ matrices one is interested in deviation inequalities for

$$\lambda_{\max}\left(\sum_{k=1}^n X_k\right).$$

For example, a very useful adaptation of Bernstein's inequality exists for this setting.

Theorem 4.13 (Theorem 1.4 in [Tro12]) *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 $t \geq 0$,

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

Note that $\|A\|$ denotes the spectral norm of A.

In what follows we will state and prove various matrix concentration results, somewhat similar to Theorem 4.13. Motivated by the derivation of Proposition 4.8, that allowed us to easily transform bounds on the expected spectral norm of a random matrix into tail bounds, we will mostly focus on bounding the expected spectral norm. Tropp's monograph [Tro15b] is a nice introduction to matrix concentration and includes a proof of Theorem 4.13 as well as many other useful inequalities.

A particularly important inequality of this type is for gaussian series, it is intimately related to the non-commutative Khintchine inequality [Pis03], and for that reason we will often refer to it as Non-commutative Khintchine (see, for example, (4.9) in [Tro12]).

Theorem 4.14 (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:*

$$\mathbb{E}\left\|\sum_{k=1}^n g_k A_k\right\| \leq \Big(2 + 2\log(2d)\Big)^{\frac{1}{2}}\sigma,$$

where

$$\sigma^2 = \left\|\sum_{k=1}^n A_k^2\right\|^2. \tag{38}$$

Note that, akin to Proposition 4.8, 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 the first inequality made use of the triangular inequality and the last one of the Cauchy-Schwarz inequality.

This motivates us to define a new parameter, the weak variance σ_*.

Definition 4.15 (Weak Variance (see, for example, [Tro15b])) *Given $A_1, \ldots, A_n \in \mathbb{R}^{d \times d}$ 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

$$\text{Prob}\left\{ \left\| \sum_{k=1}^{n} g_k A_k \right\| \geq \left(2 + 2\log(2d) \right)^{\frac{1}{2}} \sigma + u\sigma_* \right\} \leq \exp\left(-\frac{1}{2} u^2 \right). \tag{39}$$

This means that 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 4.16 *Given $A_1, \ldots, A_n \in \mathbb{R}^{d\times d}$ symmetric matrices, recall that*

$$\sigma = \sqrt{\left\|\sum_{k=1}^n A_k^2\right\|^2} \text{ and } \sigma_* = \sqrt{\max_{v\colon \|v\|=1} \sum_{k=1}^n \left(v^T A_k v\right)^2}.$$

We have

$$\sigma_* \le \sigma.$$

Proof. Using the Cauchy-Schwarz inequality,

$$\begin{aligned}
\sigma_*^2 &= \max_{v\colon \|v\|=1} \sum_{k=1}^n \left(v^T A_k v\right)^2 \\
&= \max_{v\colon \|v\|=1} \sum_{k=1}^n \left(v^T \left[A_k v\right]\right)^2 \\
&\le \max_{v\colon \|v\|=1} \sum_{k=1}^n \left(\|v\|\|A_k v\|\right)^2 \\
&= \max_{v\colon \|v\|=1} \sum_{k=1}^n \|A_k v\|^2 \\
&= \max_{v\colon \|v\|=1} \sum_{k=1}^n v^T A_k^2 v \\
&= \left\|\sum_{k=1}^n A_k^2\right\| \\
&= \sigma^2.
\end{aligned}$$

□

4.5 Optimality of matrix concentration result for gaussian series

The following simple calculation is suggestive that the parameter σ in Theorem 4.14 is indeed the correct parameter to understand $\mathbb{E}\left\|\sum_{k=1}^n g_k A_k\right\|$.

$$\begin{aligned}
\mathbb{E}\left\|\sum_{k=1}^n g_k A_k\right\|^2 &= \mathbb{E}\left\|\left(\sum_{k=1}^n g_k A_k\right)^2\right\| = \mathbb{E} \max_{v\colon\ \|v\|=1} v^T \left(\sum_{k=1}^n g_k A_k\right)^2 v \\
&\ge \max_{v\colon\ \|v\|=1} \mathbb{E} v^T \left(\sum_{k=1}^n g_k A_k\right)^2 v = \max_{v\colon\ \|v\|=1} v^T \left(\sum_{k=1}^n A_k^2\right) v = \sigma^2 \qquad (40)
\end{aligned}$$

But a natural question is whether the logarithmic term is needed. Motivated by this question we'll explore a couple of examples.

Example 4.17 *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 means that Theorem 4.14 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 4.14) is, in this case, suboptimal by a logarithmic factor.[18]

The next example will show that the logarithmic factor is in fact needed in some examples

Example 4.18 *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 it's 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 [RS13, Seg00, Lat05, BvH15] is now somewhat understood:

Theorem 4.19 ([BvH15]) *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*

$$\mathbb{E}\|X\| \lesssim \sqrt{\max_{1 \leq i \leq d} \sum_{j=1}^d b_{ij}^2} + \max_{ij} |b_{ij}| \sqrt{\log d}.$$

Remark 4.20 *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_*:

$$\sigma^2 = \max_{1 \leq i \leq d} \sum_{j=1}^d b_{ij}^2 \text{ and } \sigma_*^2 \asymp b_{ij}^2.$$

This means that, when the random matrix in NCK (Theorem 4.14) has negative entries (modulo symmetry) then

$$\mathbb{E}\|X\| \lesssim \sigma + \sqrt{\log d}\, \sigma_*. \tag{41}$$

[18] By $a \asymp b$ we mean $a \lesssim b$ and $a \gtrsim b$.

Theorem 4.19 together with a recent improvement of Theorem 4.14 by Tropp [Tro15c][19] motivate the bold possibility of (41) holding in more generality.

Conjecture 4.21 *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:*

$$\mathbb{E}\left\|\sum_{k=1}^{n} g_k A_k\right\| \lesssim \sigma + (\log d)^{\frac{1}{2}}\,\sigma_*,$$

While it may very will be that this Conjecture 4.21 is false, no counter example is known, up to date.

Open Problem 4.1 (Improvement on Non-Commutative Khintchine Inequality) *Prove or disprove Conjecture 4.21.*

I would also be pretty excited to see interesting examples that satisfy the bound in Conjecture 4.21 while such a bound would not trivially follow from Theorems 4.14 or 4.19.

4.5.1 An interesting observation regarding random matrices with independent matrices

For the independent entries setting, Theorem 4.19 is tight (up to constants) for a wide range of variance profiles $\left\{b_{ij}^2\right\}_{i\leq j}$ – the details are available as Corollary 3.15 in [BvH15]; the basic idea is that if the largest variance is comparable to the variance of a sufficient number of entries, then the bound in Theorem 4.19 is tight up to constants.

However, the situation is not as well understood when the variance profiles $\left\{b_{ij}^2\right\}_{i\leq j}$ are arbitrary. Since the spectral norm of a matrix is always at least the ℓ_2 norm of a row, the following lower bound holds (for X a symmetric random matrix with independent gaussian entries):

$$\mathbb{E}\|X\| \geq \mathbb{E}\max_k \|Xe_k\|_2.$$

Observations in papers of Latała [Lat05] and Riemer and Schutt [RS13], together with the results in [BvH15], motivate the conjecture that this lower bound is always tight (up to constants).

Open Problem 4.2 (Latała-Riemer-Schutt) *Given X a symmetric random matrix with independent gaussian entries, is the following true?*

$$\mathbb{E}\|X\| \lesssim \mathbb{E}\max_k \|Xe_k\|_2.$$

The results in [BvH15] answer this in the positive for a large range of variance profiles, but not in full generality. Recently, van Handel [vH15] proved this conjecture in the positive with an extra factor of $\sqrt{\log\log d}$. More precisely, that

$$\mathbb{E}\|X\| \lesssim \sqrt{\log\log d}\mathbb{E}\max_k \|Xe_k\|_2,$$

where d is the number of rows (and columns) of X.

[19] We briefly discuss this improvement in Remark 4.32

4.6 A matrix concentration inequality for Rademacher Series

In what follows, we closely follow [Tro15a] 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 4.14)

Theorem 4.22 *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:*

$$\mathbb{E}\left\|\sum_{k=1}^n \varepsilon_k H_k\right\| \leq \Big(1 + 2\lceil\log(d)\rceil\Big)^{\frac{1}{2}}\sigma,$$

where

$$\sigma^2 = \left\|\sum_{k=1}^n H_k^2\right\|^2. \tag{42}$$

Before proving this theorem, we take first a small detour in discrepancy theory followed by derivations, using this theorem, of a couple of useful matrix concentration inequalities.

4.6.1 A small detour on discrepancy theory

The following conjecture appears in a nice blog post of Raghu Meka [Mek14].

Conjecture 4.23 *[Matrix Six-Deviations Suffice] There exists a universal constant C such that, for any choice of n symmetric matrices $H_1, \ldots, H_n \in \mathbb{R}^{n\times n}$ satisfying $\|H_k\| \leq 1$ (for all $k = 1, \ldots, n$), there exists $\varepsilon_1, \ldots, \varepsilon_n \in \{\pm 1\}$ such that*

$$\left\|\sum_{k=1}^n \varepsilon_k H_k\right\| \leq C\sqrt{n}.$$

Open Problem 4.3 *Prove or disprove Conjecture 4.23.*

Note that, when the matrices H_k are diagonal, this problem corresponds to Spencer's Six Standard Deviations Suffice Theorem [Spe85].

Remark 4.24 *Also, using Theorem 4.22, it is easy to show that if one picks ε_i as i.i.d. Rademacher random variables, then with positive probability (via the probabilistic method) the inequality will be satisfied with an extra $\sqrt{\log n}$ term. In fact one has*

$$\mathbb{E}\left\|\sum_{k=1}^n \varepsilon_k H_k\right\| \lesssim \sqrt{\log n}\sqrt{\left\|\sum_{k=1}^n H_k^2\right\|} \leq \sqrt{\log n}\sqrt{\sum_{k=1}^n \|H_k\|^2} \leq \sqrt{\log n}\sqrt{n}.$$

Remark 4.25 *Remark 4.24 motivates asking whether Conjecture 4.23 can be strengthened to ask for $\varepsilon_1, \ldots, \varepsilon_n$ such that*

$$\left\|\sum_{k=1}^n \varepsilon_k H_k\right\| \lesssim \left\|\sum_{k=1}^n H_k^2\right\|^{\frac{1}{2}}. \tag{43}$$

4.6.2 Back to matrix concentration

Using Theorem 4.22, we'll prove the following Theorem.

Theorem 4.26 *Let $T_1, \ldots, T_n \in \mathbb{R}^{d\times d}$ be random independent 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

$$C(d) := 4 + 8\lceil \log d\rceil. \tag{44}$$

A key step in the proof of Theorem 4.26 is an idea that is extremely useful in Probability, the trick of symmetrization. For this reason we isolate it in a lemma.

Lemma 4.27 (Symmetrization) *Let $T_1, \ldots, T_n$ be independent random matrices (note that they don't 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. 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

$$\begin{aligned}\mathbb{E}\left\|\sum_{i=1}^n (T_i - \mathbb{E}T_i)\right\| &= \mathbb{E}_T\left\|\sum_{i=1}^n \left(T_i - \mathbb{E}T_i - \mathbb{E}_{T_i'}\left[T_i' - \mathbb{E}_{T_i'}T_i'\right]\right)\right\| \\ &= \mathbb{E}_T\left\|\mathbb{E}_{T'}\sum_{i=1}^n \left(T_i - T_i'\right)\right\| \leq \mathbb{E}\left\|\sum_{i=1}^n \left(T_i - T_i'\right)\right\|,\end{aligned}$$

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, it is identically distributed to $\varepsilon_i\left(T_i - T_i'\right)$ which gives

$$\mathbb{E}\left\|\sum_{i=1}^n \left(T_i - T_i'\right)\right\| = \mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i\left(T_i - T_i'\right)\right\| \leq \mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i T_i\right\| + \mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i T_i'\right\| = 2\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i T_i\right\|,$$

concluding the proof. □

Proof. [of Theorem 4.26]

Using Lemma 4.27 and Theorem 4.22 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 [Tro15a] 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

$$\mathbb{E}\left\|\sum_{i=1}^{n} T_i\right\| \leq \left\|\sum_{i=1}^{n} \mathbb{E}T_i\right\| + \sqrt{C(d)}\mathbb{E}\left[\left(\max_i \|T_i\|\right)^{\frac{1}{2}} \left\|\sum_{i=1}^{n} T_i\right\|^{\frac{1}{2}}\right].$$

Further 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 [Tro15a] for details).

□

We now show an inequality for general symmetric matrices

Theorem 4.28 *Let $Y_1, \ldots, Y_n \in \mathbb{R}^{d\times d}$ be random independent positive semidefinite matrices, then*

$$\mathbb{E}\left\|\sum_{i=1}^{n} Y_i\right\| \leq \sqrt{C(d)}\sigma + C(d)L,$$

where,

$$\sigma^2 = \left\|\sum_{i=1}^{n} \mathbb{E}Y_i^2\right\| \text{ and } L^2 = \mathbb{E}\max_i \|Y_i\|^2 \tag{45}$$

and, as in (44),

$$C(d) := 4 + 8\lceil \log d\rceil.$$

Proof.

Using Symmetrization (Lemma 4.27) and Theorem 4.22, 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 4.26. □

Remark 4.29 (The rectangular case) *One can extend Theorem 4.28 to general rectangular matrices $S_1, \ldots, S_n \in \mathbb{R}^{d_1 \times d_2}$ by setting*

$$Y_i = \left[\begin{array}{cc} 0 & S_i \\ S_i^T & 0 \end{array}\right],$$

and noting that

$$\left\|Y_i^2\right\| = \left\|\left[\begin{array}{cc} 0 & S_i \\ S_i^T & 0 \end{array}\right]^2\right\| = \left\|\left[\begin{array}{cc} S_i S_i^T & 0 \\ 0 & S_i^T S_i \end{array}\right]\right\| = \max\left\{\left\|S_i^T S_i\right\|, \left\|S_i S_i^T\right\|\right\}.$$

We defer the details to [Tro15a]

In order to prove Theorem 4.22, we will use an AM-GM like inequality for matrices for which, unlike the one on Open Problem 0.2. in [Ban15c], an elementary proof is known.

Lemma 4.30 *Given symmetric matrices $H, W, Y \in \mathbb{R}^{d \times d}$ and non-negative integers r, q satisfying $q \leq 2r$,*

$$\operatorname{Tr}\left[H W^q H Y^{2r-q}\right] + \operatorname{Tr}\left[H W^{2r-q} H Y^q\right] \leq \operatorname{Tr}\left[H^2\left(W^{2r} + Y^{2r}\right)\right],$$

and summing over q gives

$$\sum_{q=0}^{2r} \operatorname{Tr}\left[H W^q H Y^{2r-q}\right] \leq \left(\frac{2r+1}{2}\right) \operatorname{Tr}\left[H^2\left(W^{2r} + Y^{2r}\right)\right]$$

We refer to Fact 2.4 in [Tro15a] for an elementary proof but note that it is a matrix analogue to the inequality,

$$\mu^\theta \lambda^{1-\theta} + \mu^{1-\theta}\lambda^\theta \leq \lambda + \theta$$

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 4.22]

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

$$\mathbb{E}\operatorname{Tr} X^{2p} = \sum_{i=1}^{n}\sum_{q=0}^{2p-2} \mathbb{E}\operatorname{Tr}\left[H_i X_{+i}^{q} H_i X_{-i}^{2p-2-q}\right]. \tag{46}$$

We now make use of Lemma 4.30 to get[20] to get

$$\mathbb{E}\operatorname{Tr} X^{2p} \leq \sum_{i=1}^{n} \frac{2p-1}{2}\mathbb{E}\operatorname{Tr}\left[H_i^2\left(X_{+i}^{2p-2} + X_{-i}^{2p-2}\right)\right]. \tag{47}$$

[20] See Remark 4.32 regarding the suboptimality of this step.

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

$$\mathrm{Tr}\left[\left(\sum_{i=1}^{n}H_i^2\right)X^{2p-2}\right]\le\left\|\sum_{i=1}^{n}H_i^2\right\|\mathrm{Tr}\,X^{2p-2}=\sigma^2\,\mathrm{Tr}\,X^{2p-2}, \tag{48}$$

which gives

$$\mathbb{E}\,\mathrm{Tr}\,X^{2p}\le\sigma^2(2p-1)\mathbb{E}\,\mathrm{Tr}\,X^{2p-2}. \tag{49}$$

Applying this inequality, recursively, we get

$$\mathbb{E}\,\mathrm{Tr}\,X^{2p}\le\left[(2p-1)(2p-3)\cdots(3)(1)\right]\sigma^{2p}\mathbb{E}\,\mathrm{Tr}\,X^{0}=(2p-1)!!\sigma^{2p}d$$

Hence,

$$\mathbb{E}\|X\|\le\left(\mathbb{E}\,\mathrm{Tr}\,X^{2p}\right)^{\frac{1}{2p}}\le\left[(2p-1)!!\right]^{\frac{1}{2p}}\sigma d^{\frac{1}{2p}}.$$

Taking $p=\lceil\log d\rceil$ and using the fact that $(2p-1)!!\le\left(\frac{2p+1}{e}\right)^p$ (see [Tro15a] for an elementary proof consisting essentially of taking logarithms and comparing the sum with an integral) we get

$$\mathbb{E}\|X\|\le\left(\frac{2\lceil\log d\rceil+1}{e}\right)^{\frac{1}{2}}\sigma d^{\frac{1}{2\lceil\log d\rceil}}\le\left(2\lceil\log d\rceil+1\right)^{\frac{1}{2}}\sigma.$$

□

Remark 4.31 *A similar argument can be used to prove Theorem 4.14 (the gaussian series case) based on gaussian integration by parts, see Section 7.2. in [Tro15c].*

Remark 4.32 *Note that, up until the step from* (46) *to* (47) *all steps are equalities suggesting that this step may be the lossy step responsible by the suboptimal dimensional factor in several cases (although* (48) *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 [Tro15c] recently proved an improvement over the NCK inequality that, essentially, consists in replacing inequality (47) *with a tighter argument. In a nutshell, the idea is that, if the* H_i*'s are non-commutative, most summands in* (46) *are actually expected to be smaller than the ones corresponding to* $q=0$ *and* $q=2p-2$*, which are the ones that appear in* (47)*.*

4.7 Other Open Problems

4.7.1 Oblivious Sparse Norm-Approximating Projections

There is an interesting random matrix problem related to Oblivious Sparse Norm-Approximating Projections [NN], a form of dimension reduction useful for fast linear algebra. In a nutshell, The idea is to try to find random matrices Π that achieve dimension reduction, meaning $\Pi \in \mathbb{R}^{m \times n}$ with $m \ll n$, and that preserve the norm of every point in a certain subspace [NN], moreover, for the sake of computational efficiency, these matrices should be sparse (to allow for faster matrix-vector multiplication). In some sense, this is a generalization of the ideas of the Johnson-Lindenstrauss Lemma and Gordon's Escape through the Mesh Theorem that we will discuss next Section.

Open Problem 4.4 (OSNAP [NN]) *Let $s \leq d \leq m \leq n$.*

1. *Let $\Pi \in \mathbb{R}^{m \times n}$ be a random matrix with i.i.d. entries*

$$\Pi_{ri} = \frac{\delta_{ri}\sigma_{ri}}{\sqrt{s}},$$

where σ_{ri} is a Rademacher random variable and

$$\delta_{ri} = \begin{cases} \frac{1}{\sqrt{s}} & \text{with probability} \quad \frac{s}{m} \\ 0 & \text{with probability} \quad 1 - \frac{s}{m} \end{cases}$$

Prove or disprove: there exist positive universal constants c_1 and c_2 such that

For any $U \in \mathbb{R}^{n \times d}$ for which $U^T U = I_{d \times d}$

$$\text{Prob}\left\{ \left\| (\Pi U)^T (\Pi U) - I \right\| \geq \varepsilon \right\} < \delta,$$

for $m \geq c_1 \frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$ and $s \geq c_2 \frac{\log\left(\frac{d}{\delta}\right)}{\varepsilon^2}$.

2. *Same setting as in (1) but conditioning on*

$$\sum_{r=1}^{m} \delta_{ri} = s, \quad \text{for all } i,$$

meaning that each column of Π has exactly s non-zero elements, rather than on average. The conjecture is then slightly different:

Prove or disprove: there exist positive universal constants c_1 and c_2 such that

For any $U \in \mathbb{R}^{n \times d}$ for which $U^T U = I_{d \times d}$

$$\text{Prob}\left\{ \left\| (\Pi U)^T (\Pi U) - I \right\| \geq \varepsilon \right\} < \delta,$$

for $m \geq c_1 \frac{d + \log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$ and $s \geq c_2 \frac{\log\left(\frac{d}{\delta}\right)}{\varepsilon}$.

3. *The conjecture in (1) but for the specific choice of U:*

$$U = \begin{bmatrix} I_{d\times d} \\ 0_{(n-d)\times d} \end{bmatrix}.$$

In this case, the object in question is a sum of rank 1 independent matrices. More precisely, $z_1, \dots, z_m \in \mathbb{R}^d$ (corresponding to the first d coordinates of each of the m rows of Π) are i.i.d. random vectors with i.i.d. entries

$$(z_k)_j = \begin{cases} -\frac{1}{\sqrt{s}} & \text{with probability} \quad \frac{s}{2m} \\ 0 & \text{with probability} \quad 1-\frac{s}{m} \\ \frac{1}{\sqrt{s}} & \text{with probability} \quad \frac{s}{2m} \end{cases}$$

Note that $\mathbb{E} z_k z_k^T = \frac{1}{m} I_{d\times d}$. The conjecture is then that, there exists c_1 and c_2 positive universal constants such that

$$\text{Prob}\left\{ \left\| \sum_{k=1}^m \left[z_k z_k^T - \mathbb{E} z_k z_k^T \right] \right\| \geq \varepsilon \right\} < \delta,$$

for $m \geq c_1 \frac{d+\log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$ and $s \geq c_2 \frac{\log\left(\frac{d}{\delta}\right)}{\varepsilon^2}$.

I think this would is an interesting question even for fixed δ, for say $\delta = 0.1$, or even simply understand the value of

$$\mathbb{E} \left\| \sum_{k=1}^m \left[z_k z_k^T - \mathbb{E} z_k z_k^T \right] \right\|.$$

4.7.2 k-lifts of graphs

Given a graph G, on n nodes and with max-degree Δ, and an integer $k \geq 2$ a random k lift $G^{\otimes k}$ of G is a graph on kn nodes obtained by replacing each edge of G by a random $k \times k$ bipartite matching. More precisely, the adjacency matrix $A^{\otimes k}$ of $G^{\otimes k}$ is a $nk \times nk$ matrix with $k \times k$ blocks given by

$$A_{ij}^{\otimes k} = A_{ij} \Pi_{ij},$$

where Π_{ij} is uniformly randomly drawn from the set of permutations on k elements, and all the edges are independent, except for the fact that $\Pi_{ij} = \Pi_{ji}$. In other words,

$$A^{\otimes k} = \sum_{i<j} A_{ij} \left(e_i e_j^T \otimes \Pi_{ij} + e_j e_i^T \otimes \Pi_{ij}^T \right),$$

where $\otimes$ corresponds to the Kronecker product. Note that

$$\mathbb{E} A^{\otimes k} = A \otimes \left(\frac{1}{k} J \right),$$

where $J = \mathbf{1}\mathbf{1}^T$ is the all-ones matrix.

Open Problem 4.5 (Random k-lifts of graphs) *Give a tight upperbound to*

$$\mathbb{E}\left\|A^{\otimes k} - \mathbb{E}A^{\otimes k}\right\|.$$

Oliveira [Oli10] gives a bound that is essentially of the form $\sqrt{\Delta \log(nk)}$, while the results in [ABG12] suggest that one may expect more concentration for large k. It is worth noting that the case of $k = 2$ can essentially be reduced to a problem where the entries of the random matrix are independent and the results in [BvH15] can be applied to, in some case, remove the logarithmic factor.

4.8 Another open problem

Feige [Fei05] posed the following remarkable conjecture (see also [Sam66, Sam69, Sam68])

Conjecture 4.33 *Given n independent random variables $X_1, \dots, X_n$ s.t., for all i, $X_i \geq 0$ and $\mathbb{E}X_i = 1$ we have*

$$\text{Prob}\left(\sum_{i=1}^{n} X_i \geq n + 1\right) \leq 1 - e^{-1}$$

Note that, if X_i are i.i.d. and $X_i = n + 1$ with probability $1/(n + 1)$ and $X_i = 0$ otherwise, then $\text{Prob}\left(\sum_{i=1}^{n} X_i \geq n + 1\right) = 1 - \left(\frac{n}{n+1}\right)^n \approx 1 - e^{-1}$.

Open Problem 4.6 *Prove or disprove Conjecture 4.33.*[21]

[21] We thank Francisco Unda and Philippe Rigollet for suggesting this problem.

5 Johnson-Lindenstrauss Lemma and Gordons Theorem

5.1 The Johnson-Lindenstrauss Lemma

Suppose one has n points, $X = \{x_1, \dots, x_n\}$, in $\mathbb{R}^d$ (with d large). If $d > n$, since the points have to lie in a subspace of dimension n it is clear that one can consider the projection $f : \mathbb{R}^d \to \mathbb{R}^n$ of the points to that subspace 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 now we allow a bit of distortion, and look for $f : \mathbb{R}^d \to \mathbb{R}^k$ that is an $\epsilon-$isometry, meaning that

$$(1-\epsilon)\|x_i - x_j\|^2 \le \|f(x_i) - f(x_j)\|^2 \le (1+\epsilon)\|x_i - x_j\|^2. \tag{50}$$

Can we do better than $k = n$?

In 1984, Johnson and Lindenstrauss [JL84] showed a remarkable Lemma (below) that answers this question positively.

Theorem 5.1 (Johnson-Lindenstrauss Lemma [JL84]) *For any $0 < \epsilon < 1$ and for any integer n, let k be such that*

$$k \ge 4 \frac{1}{\epsilon^2/2 - \epsilon^3/3} \log n.$$

Then, for any set X of n points in $\mathbb{R}^d$, there is a linear map $f : \mathbb{R}^d \to \mathbb{R}^k$ that is an $\epsilon-$isometry for X (see (50)). This map can be found in randomized polynomial time.

We borrow, from [DG02], 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 [DG02] and are essentially the same ideas as those used to show Hoeffding's inequality.

Lemma 5.2 (see [DG02]) *Let $y_1, \dots, y_d$ be i.i.d standard Gaussian random variables and $Y = (y_1, \dots, y_d)$. Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be the projection into the first k coordinates and $Z = g\left(\frac{Y}{\|Y\|}\right) = \frac{1}{\|Y\|}(y_1, \dots, y_k)$ and $L = \|Z\|^2$. It is clear that $\mathbb{E}L = \frac{k}{d}$. In fact, L is very concentrated around its mean*

- *If $\beta < 1$,*

$$\Pr\left[L \le \beta \frac{k}{d}\right] \le \exp\left(\frac{k}{2}(1 - \beta + \log \beta)\right).$$

- *If $\beta > 1$,*

$$\Pr\left[L \ge \beta \frac{k}{d}\right] \le \exp\left(\frac{k}{2}(1 - \beta + \log \beta)\right).$$

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 (50) with high probability. WLOG, 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 k is the same as understanding the norm of the projection of a (uniformly) random point on S^{d-1} the unit sphere in $\mathbb{R}^d$ on a specific $k-$dimensional subspace, let's say the one generated by the first k 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^{d-1} – this distribution is the same as taking a standard Gaussian vector in $\mathbb{R}^d$ 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 5.2. Using Lemma 5.2, we have, given a pair x_i, x_j,

$$\Pr\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \leq (1-\epsilon)\right] \leq \exp\left(\frac{k}{2}(1-(1-\epsilon)+\log(1-\epsilon))\right),$$

since, for $\epsilon \geq 0$, $\log(1-\epsilon) \leq -\epsilon - \epsilon^2/2$ we have

$$\begin{aligned}\Pr\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \leq (1-\epsilon)\right] &\leq \exp\left(-\frac{k\epsilon^2}{4}\right)\\ &\leq \exp\left(-2\log n\right) = \frac{1}{n^2}.\end{aligned}$$

On the other hand,

$$\Pr\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \geq (1+\epsilon)\right] \leq \exp\left(\frac{k}{2}(1-(1+\epsilon)+\log(1+\epsilon))\right).$$

since, for $\epsilon \geq 0$, $\log(1+\epsilon) \leq \epsilon - \epsilon^2/2 + \epsilon^3/3$ we have

$$\begin{aligned}\text{Prob}\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \leq (1-\epsilon)\right] &\leq \exp\left(-\frac{k\left(\epsilon^2-2\epsilon^3/3\right)}{4}\right)\\ &\leq \exp\left(-2\log n\right) = \frac{1}{n^2}.\end{aligned}$$

By union bound it follows that

$$\Pr\left[\frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \notin [1-\epsilon, 1+\epsilon]\right] \leq \frac{2}{n^2}.$$

Since there exist $\binom{n}{2}$ such pairs, again, a simple union bound gives

$$\Pr\left[\exists_{i,j} : \frac{\|f(x_i)-f(x_j)\|^2}{\|x_i-x_j\|^2} \notin [1-\epsilon, 1+\epsilon]\right] \leq \frac{2}{n^2}\frac{n(n-1)}{2} = 1 - \frac{1}{n}.$$

Therefore, choosing f as a properly scaled projection onto a random $k-$dimensional subspace is an $\epsilon-$ isometry on X (see (50)) 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 $\epsilon-$isometry in randomized polynomial time.

□

Note that by considering k slightly larger one can get a good projection on the first random attempt with very good confidence. In fact, it's trivial to adapt the proof above to obtain the following Lemma:

Lemma 5.3 *For any $0 < \epsilon < 1$, $\tau > 0$, and for any integer n, let k be such that*

$$k \geq (2+\tau)\frac{2}{\epsilon^2/2 - \epsilon^3/3}\log n.$$

Then, for any set X of n points in $\mathbb{R}^d$, take $f : \mathbb{R}^d \to \mathbb{R}^k$ to be a suitably scaled projection on a random subspace of dimension k, then f is an $\epsilon-$isometry for X (see (50)) with probability at least $1 - \frac{1}{n^\tau}$.

Lemma 5.3 is quite remarkable. Think about 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 5.3 you can just choose a projection at random in the beginning of the process (all ones needs to know is an estimate of the log of the size of the data set) and just map each point using this projection matrix which can be done online – we don't need to see the next point to compute the projection of the current data point. Lemma 5.3 ensures that this (seemingly naïve) procedure will, with high probably, not distort the data by more than ϵ.

5.1.1 Optimality of the Johnson-Lindenstrauss Lemma

It is natural to ask whether the dependency on ϵ and n in Lemma 5.3 can be improved. Noga Alon [Alo03] showed that there are n points for which the smallest dimension k on which they can be embedded with a distortion as in Lemma 5.3, satisfies $k = \Omega\left(\frac{1}{\log(1/\epsilon)}\epsilon^{-2}\log n\right)$, this was recently improved by Larsen and Nelson [**?**], for linear maps, to $\Omega\left(\epsilon^{-2}\log n\right)$, closing the gap.[22]

5.1.2 Fast Johnson-Lindenstrauss

(Disclaimer: the purpose of this section is just to provide a bit of intuition, there is a lot of hand-waving!!)

Let's continue thinking about the high-dimensional streaming data. After we draw the random projection matrix, say M, for each data point x, we still have to compute Mx which, since M has $\mathcal{O}(\epsilon^{-2}\log(n)d)$ entries, has a computational cost of $\mathcal{O}(\epsilon^{-2}\log(n)d)$. In some applications this might be too expensive, can one do better? There is no hope of (significantly) reducing the number of rows (Recall Open Problem **??** and the lower bound by Alon [Alo03]). The only hope is to speed up the matrix-vector multiplication. If we were able to construct a sparse matrix M then we would definitely speed up the computation of Mx but sparse matrices tend to distort sparse vectors, and the data set may contain. Another option would be to exploit the Fast Fourier Transform and compute the Fourier Transform of x (which takes $\mathcal{O}(d\log d)$ time) and then multiply the Fourier Transform of x by a sparse matrix. However, this again may not work because x might have a sparse Fourier Transform.

[22] An earlier version of these notes marked closing the gap as an open problem, this has been corrected.

The solution comes from leveraging an uncertainty principle — it is impossible for both x and the FT of x to be sparse simultaneously. The idea is that if, before one takes the Fourier Transform of x, one flips (randomly) the signs of x, then the probably of obtaining a sparse vector is very small so a sparse matrix can be used for projection. In a nutshell the algorithm has M be a matrix of the form PHD, where D is a diagonal matrix that flips the signs of the vector randomly, H is a Fourier Transform (or Hadamard transform) and P a sparse matrix. This method was proposed and analysed in [AC09] and, roughly speaking, achieves a complexity of $\mathcal{O}(d \log d)$, instead of the classical $\mathcal{O}(\epsilon^{-2} \log(n) d)$.

There is a very interesting line of work proposing fast Johnson Lindenstrauss projections based on sparse matrices. In fact, this is, in some sense, the motivation for Open Problem 4.4. in [Ban15c]. We recommend these notes Jelani Nelson's notes for more on the topic [Nel].

5.2 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 $\Theta\left(\epsilon^{-2}\log n\right)$ 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 $\Theta\left(\epsilon^{-2}\log|S|\right)$ dimensions. The question this section is meant to answer is: can this 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 take on the projection to still approximately preserve the norms of points in S. Was we will see below, this will be captured, via Gordon's Theorem, by the so called Gaussian Width of S.

Definition 5.4 (Gaussian Width) *Given a closed set $S \subset \mathbb{R}^d$, its* gaussian width $\omega(S)$ *is define as:*

$$\omega(S) = \mathbb{E}\max_{x \in S}\left[g_d^T x\right],$$

where $g_d \sim \mathcal{N}(0, I_{d\times d})$.

Similarly to what we did in the proof of Theorem 5.1 we will restrict our attention to sets S of unit norm vectors, meaning that $S \subset \mathbb{S}^{d-1}$.

Also, we will focus our attention not in random projections but in the similar model of random linear maps $G : \mathbb{R}^d \to \mathbb{R}^k$ that are given by matrices with i.i.d. gaussian entries. For this reason the following Proposition will be useful:

Proposition 5.5 *Let $g_k \sim \mathcal{N}(0, I_{k\times k})$, and define*

$$a_k := \mathbb{E}\|g_k\|.$$

Then $\sqrt{\frac{k}{k+1}}\sqrt{k} \leq a_k \leq \sqrt{k}$.

We are now ready to present Gordon's Theorem.

Theorem 5.6 (Gordon's Theorem [Gor88]) *Let $G \in \mathbb{R}^{k\times d}$ a random matrix with independent $\mathcal{N}(0,1)$ entries and $S \subset \mathbb{S}^{d-1}$ be a closed subset of the unit sphere in d dimensions. Then*

$$\mathbb{E}\max_{x\in S}\left\|\frac{1}{a_k}Gx\right\| \leq 1 + \frac{\omega(S)}{a_k},$$

and

$$\mathbb{E}\min_{x\in S}\left\|\frac{1}{a_k}Gx\right\| \geq 1 - \frac{\omega(S)}{a_k},$$

where $a_k = \mathbb{E}\|g_k\|$ and $\omega(S)$ is the gaussian width of S. Recall that $\sqrt{\frac{k}{k+1}}\sqrt{k} \leq a_k \leq \sqrt{k}$.

Before proving Gordon's Theorem we'll note some of it's direct implications. It suggest that $\frac{1}{a_k}G$ preserves the norm of the points in S up to $1 \pm \frac{\omega(S)}{a_k}$, indeed we can make this precise with Gaussian Concentration.

Note that the function $F(G) = \max_{x\in S}\left\|\frac{1}{a_k}Gx\right\|$ is 1-Lipschitz. Indeed

$$\begin{aligned}\left|\max_{x_1\in S}\|G_1x_1\| - \max_{x_2\in S}\|G_2x_2\|\right| &\leq \max_{x\in S}\left|\|G_1x\| - \|G_2x\|\right| \leq \max_{x\in S}\|(G_1-G_2)\,x\| \\ &= \|G_1-G_2\| \leq \|G_1-G_2\|_F\,.\end{aligned}$$

Similarly, one can show that $F(G) = \min_{x\in S}\left\|\frac{1}{a_k}Gx\right\|$ is 1-Lipschitz. Thus, one can use Gaussian Concentration to get:

$$\operatorname{Prob}\left\{\max_{x\in S}\|Gx\| \geq a_k + \omega(S) + t\right\} \leq \exp\left(-\frac{t^2}{2}\right), \tag{51}$$

and

$$\operatorname{Prob}\left\{\min_{x\in S}\|Gx\| \leq a_k - \omega(S) - t\right\} \leq \exp\left(-\frac{t^2}{2}\right). \tag{52}$$

This gives us the following Theorem.

Theorem 5.7 *Let $G \in \mathbb{R}^{k\times d}$ a random matrix with independent $\mathcal{N}(0,1)$ entries and $S \subset \mathbb{S}^{d-1}$ be a closed subset of the unit sphere in d dimensions. Then, for $\varepsilon > \sqrt{\frac{\omega(S)^2}{a_k^2}}$, with probability $\geq 1 - 2\exp\left[-k\left(\varepsilon - \frac{\omega(S)}{a_k}\right)^2\right]$:*

$$(1-\varepsilon)\|x\| \leq \left\|\frac{1}{a_k}Gx\right\| \leq (1+\varepsilon)\|x\|,$$

for all $x \in S$.

Recall that $k - \frac{k}{k+1} \leq a_k^2 \leq k$.

Proof. This is readily obtained by taking $\varepsilon = \frac{\omega(S)+t}{a_k}$, using (51), (52), and recalling that $a_k^2 \leq k$. □

Remark 5.8 *Note that a simple use of a union bound*[23] *shows that $\omega(S) \lesssim \sqrt{2\log|S|}$, which means that taking k to be of the order of $\log|S|$ suffices to ensure that $\frac{1}{a_k}G$ to have the Johnson Lindenstrauss property. This observation shows that Theorem 5.7 essentially directly implies Theorem 5.1 (although not exacly, since $\frac{1}{a_k}G$ is not a projection).*

5.2.1 Gordon's Escape Through a Mesh Theorem

Theorem 5.7 suggests that, if $\omega(S) \le a_k$, a uniformly chosen random subspace of $\mathbb{R}^n$ of dimension $(n-k)$ (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, it's Corollary 3.4. in Gordon's original paper [Gor88]. See also [Mix14b] for a description of the proof. We include the Theorem below for the sake of completeness.

Theorem 5.9 (Corollary 3.4. in [Gor88]) *Let $S \subset \mathbb{S}^{d-1}$ be a closed subset of the unit sphere in d dimensions. If $\omega(S) < a_k$, then for a $(n-k)$-dimensional subspace Λ drawn uniformly from the Grassmanian manifold we have*

$$\operatorname{Prob}\left\{\Lambda \cap S \neq \emptyset\right\} \le \frac{7}{2}\exp\left(-\frac{1}{18}\left(a_k - \omega(S)\right)^2\right),$$

where $\omega(S)$ is the gaussian width of S and $a_k = \mathbb{E}\|g_k\|$ where $g_k \sim \mathcal{N}(0, I_{k\times k})$.

5.2.2 Proof of Gordon's Theorem

In order to prove this Theorem we will use extensions of the Slepian's Comparison Lemma.

Slepian's Comparison Lemma, 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. [24]

Theorem 5.10 (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 $u_1, u_2 \in U$:*

$$\mathbb{E}\left[X_{u_1} - X_{u_2}\right]^2 \le \mathbb{E}\left[Y_{u_1} - Y_{u_2}\right]^2, \tag{53}$$

then

$$\mathbb{E}\left[\max_{u\in U} X_u\right] \le \mathbb{E}\left[\max_{u\in U} Y_u\right].$$

The following extension is due to Gordon [Gor85, Gor88].

[23] This follows from the fact that the maximum of n standard gaussian random variables is $\lesssim \sqrt{2\log|S|}$.

[24] 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 5.11 *[Theorem A in [Gor88]] 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 $u_1, u_2 \in U$:*

$$\mathbb{E}\left[X_{t_1,u_1} - X_{t_1,u_2}\right]^2 \leq \mathbb{E}\left[Y_{t_1,u_1} - Y_{t_1,u_2}\right]^2, \tag{54}$$

and, for $t_1 \neq t_2$,

$$\mathbb{E}\left[X_{t_1,u_1} - X_{t_2,u_2}\right]^2 \geq \mathbb{E}\left[Y_{t_1,u_1} - Y_{t_2,u_2}\right]^2, \tag{55}$$

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 5.10 easily follows by setting $|T| = 1$.

We are now ready to prove Gordon's theorem.

Proof. [of Theorem 5.6]

Let $G \in \mathbb{R}^{k\times d}$ with i.i.d. $\mathcal{N}(0,1)$ entries. We define two gaussian processes: For $v \in S \subset \mathbb{S}^{d-1}$ and $u \in \mathbb{S}^{k-1}$ let $g \sim \mathcal{N}(0, I_{k\times k})$ and $h \sim \mathcal{N}(0, I_{d\times d})$ and define the following processes:

$$A_{u,v} = g^T u + h^T v,$$

and

$$B_{u,v} = u^T G v.$$

For all $v, v' \in S \subset \mathbb{S}^{d-1}$ and $u, u' \in \mathbb{S}^{k-1}$,

$$\begin{aligned}
\mathbb{E}\left|A_{v,u} - A_{v',u'}\right|^2 - \mathbb{E}\left|B_{v,u} - B_{v',u'}\right|^2 &= 4 - 2\left(u^T u' + v^T v'\right) - \sum_{ij}\left(v_i u_j - v_i' u_j'\right)^2 \\
&= 4 - 2\left(u^T u' + v^T v'\right) - \left[2 - 2\left(v^T v'\right)\left(u^T u'\right)\right] \\
&= 2 - 2\left(u^T u' + v^T v' - u^T u' v^T v'\right) \\
&= 2\left(1 - u^T u'\right)\left(1 - v^T v'\right).
\end{aligned}$$

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 means that we can use Theorem 5.11 with $X = A$ and $Y = B$, to get

$$\mathbb{E}\min_{v\in S}\max_{u\in\mathbb{S}^{k-1}} A_{v,u} \leq \mathbb{E}\min_{v\in S}\max_{u\in\mathbb{S}^{k-1}} B_{v,u}.$$

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

$$\mathbb{E}\left[\min_{v\in S}\max_{u\in\mathbb{S}^{k-1}} A_{v,u}\right] = \mathbb{E}\max_{u\in\mathbb{S}^{k-1}} g^T u + \mathbb{E}\min_{v\in S} h^T v = \mathbb{E}\max_{u\in\mathbb{S}^{k-1}} g^T u - \mathbb{E}\max_{v\in S}(-h^T v) = a_k - \omega(S),$$

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 5.10 with $X = B$ and $Y = A$, to get

$$\mathbb{E} \max_{v\in S} \max_{u\in\mathbb{S}^{k-1}} A_{v,u} \geq \mathbb{E} \max_{v\in S} \max_{u\in\mathbb{S}^{k-1}} B_{v,u}.$$

Noting that

$$\mathbb{E} \max_{v\in S} \max_{u\in\mathbb{S}^{k-1}} B_{v,u} = \mathbb{E} \max_{v\in S} \max_{u\in\mathbb{S}^{k-1}} u^T G v = \mathbb{E} \max_{v\in S} \|Gv\|,$$

and

$$\mathbb{E}\left[\max_{v\in S} \max_{u\in\mathbb{S}^{k-1}} A_{v,u}\right] = \mathbb{E} \max_{u\in\mathbb{S}^{k-1}} g^T u + \mathbb{E} \max_{v\in S} h^T v = a_k + \omega(S),$$

concludes the proof of the Theorem.

□

5.3 Sparse vectors and Low-rank matrices

In this Section we illustrate the utility of Gordon's theorem by undertanding which projections are expected to keep the norm of sparse vectors and low-rank matrices.

5.3.1 Gaussian width of k-sparse vectors

Say we have a signal (or image) $x \in \mathbb{R}^N$ that we are interested in measuring with linear measurements $y_i = a_i^T x$, for $a_i \in \mathbb{R}^N$. In general, it is clear that we would need N measurements to find x. The idea behind *Compressed Sensing* [CRT06a, Don06] 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 have few non-zero entries (being sparse). Sparse signals do arise in countless applications (for example, images are known to be sparse in the Wavelet basis; in fact this is the basis of the JPEG2000 compression method).

We'll revisit sparse recovery and Compressed Sensing next lecture but for now we'll see how Gordon's Theorem can suggest us 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 hope to be able to reconstruct x from y we need 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[25]).

[25] It is important to note that $\|\cdot\|_0$ is not actually a norm

It is also intuitive that, in order for reconstruction to be stable, one would like that not only A is injective in s-sparse vectors but actually 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 keep the norm of $2s$ sparse vectors. Gordon's Theorem above suggests that we can take $A \in \mathbb{R}^{M\times N}$ to have i.i.d. gaussian entries and to take $M \approx \omega\left(\mathcal{S}_{2s}\right)$, where $\mathcal{S}_k = \left\{x : \, x \in \mathbb{S}^{N-1}, \, \|x\|_0 \leq k\right\}$ is the set of $2s$ sparse vectors, and $\omega\left(\mathcal{S}_{2s}\right)$ the gaussian width of $\mathcal{S}_{2s}$.

Proposition 5.12 *If $s \leq N$, the Gaussian Width $\omega\left(\mathcal{S}_s\right)$ of $\mathcal{S}_s$, the set of unit-norm vectors that are at most s sparse, satisfies*

$$\omega\left(\mathcal{S}_s\right)^2 \lesssim s \log\left(\frac{N}{s}\right).$$

Proof.

$$\omega\left(\mathcal{S}_s\right) = \max_{v \in SS^{N-1}, \, \|v\|_0 \leq s} g^T v, \log\left(\frac{N}{s}\right),$$

where $g \sim \mathcal{N}(0, I_{N\times N})$. We have

$$\omega\left(\mathcal{S}_s\right) = \max_{\Gamma \subset [N], \, |\Gamma| = s} \|g_\Gamma\|,$$

where g_Γ is the restriction of g to the set of indices Γ.

Given a set Γ, Theorem 4.12 gives

$$\text{Prob}\left\{\|g_\Gamma\|^2 \geq s + 2\sqrt{s}\sqrt{t} + t\right\} \leq \exp(-t).$$

Union bounding over all $\Gamma \subset [N]$, $|\Gamma| = s$ gives

$$\text{Prob}\left\{\max_{\Gamma \subset [N], \, |\Gamma| = s} \|g_\Gamma\|^2 \geq s + 2\sqrt{s}\sqrt{t} + t\right\} \leq \binom{N}{s} \exp(-t)$$

Taking u such that $t = su$, gives

$$\text{Prob}\left\{\max_{\Gamma \subset [N], \, |\Gamma| = s} \|g_\Gamma\|^2 \geq s\left(1 + 2\sqrt{u} + u\right)\right\} \leq \exp\left[-su + s\log\left(\frac{N}{s}\right)\right]. \tag{56}$$

Taking $u > \log\left(\frac{N}{s}\right)$ it can be readily seen that the typical size of $\max_{\Gamma \subset [N], \, |\Gamma| = s} \|g_\Gamma\|^2$ is $\lesssim s\log\left(\frac{N}{s}\right)$. The proof can be finished by integrating (56) in order to get a bound of the expectation of $\sqrt{\max_{\Gamma \subset [N], \, |\Gamma| = s} \|g_\Gamma\|^2}$.

□

This suggests that $\approx 2s\log\left(\frac{N}{2s}\right)$ measurements suffice to identify a $2s$-sparse vector. As we'll see, not only such a number of measurements suffices to identify a sparse vector but also for certain efficient algorithms to do so.

5.3.2 The Restricted Isometry Property and a couple of open problems

Matrices perserving the norm of sparse vectors do play a central role in sparse recovery, they are said to satisfy the Restricted Isometry Property. More precisely:

Definition 5.13 (The Restricted Isometry Property) *An $M \times N$ 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.

Using Proposition 5.12 and Theorem 5.7 one can readily show that matrices with Gaussian entries satisfy the restricted isometry property with $M \approx s \log\left(\frac{N}{s}\right)$.

Theorem 5.14 *Let A be an $M \times N$ matrix with i.i.d. standard gaussian entries, there exists a constant C such that, if*

$$M \geq Cs \log\left(\frac{N}{s}\right),$$

then $\frac{1}{a_M}A$ satisfies the $\left(s, \frac{1}{3}\right)$-RIP, with high probability.

Theorem 5.14 suggests that RIP matrices are abundant for $s \approx \frac{M}{\log(N)}$, however it appears to be very difficult to deterministically construct matrices that are RIP for $s \gg \sqrt{M}$, known as the square bottleneck [Tao07, BFMW13, BFMM14, BMM14, B$^+$11, Mix14a]. The only known unconditional construction that is able to break this bottleneck is due to Bourgain et al. [B$^+$11] that achieves $s \approx M^{\frac{1}{2}+\varepsilon}$ for a small, but positive, ε. There is a conditional construction, based on the Paley Equiangular Tight Frame, that will be briefly described in the next Lecture [BFMW13, BMM14].

Open Problem 5.1 *Construct deterministic matrices $A \in \mathbb{C}^{M\times N}$ (or $A \in \mathbb{C}^{M\times N}$) satisfying $(s, \frac{1}{3})$-RIP for $s \gtrsim \frac{M^{0.6}}{\operatorname{polylog}(N}$.*

Open Problem 5.2 *Theorem 5.14 guarantees that if we take A to have i.i.d. Gaussian entries then it should be RIP for $s \approx \frac{M}{\log(N)}$. If we were able to, given A, certify that it indeed is RIP for some s then one could have a randomized algorithm to build RIP matrices (but that is guaranteed to eventually find one). This motives the following question*

1. *Let $N = 2M$, for which s is there a polynomial time algorithm that is guaranteed to, with high probability, certify that a gaussian matrix A is $\left(s, \frac{1}{3}\right)$-RIP?*
2. *In particular, a $\left(s, \frac{1}{3}\right)$-RIP matrix has to not have s sparse vectors in its nullspace. This motivates a second question: Let $N = 2M$, for which s is there a polynomial time algorithm that is guaranteed to, with high probability, certify that a gaussian matrix A does not have s-sparse vectors in its nullspace?*

The second question is tightly connected to the question of sparsest vector on a subspace (for which $s \approx \sqrt{M}$ is the best known answer), we refer the reader to [SWW12, QSW14, BKS13b] for more on this problem and recent advances. Note that checking whether a matrix has RIP or not is, in general, NP-hard [BDMS13, TP13].

5.3.3 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 the Netflix Prize. In that particular example the matrix in question is a matrix indexed by users of the Netflix service and movies. Given a user and a movie, the corresponding entry of the matrix should correspond to the score that user would attribute to that movie. This matrix is believed to be low-rank. The goal is then to estimate the score for user and movie 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 [CT10, CR09, Rec11].

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 measuremetns of it, of the form $y_i = \operatorname{Tr}(A_i^T X)$.

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 $X \in\in \mathbb{R}^{n_1 \times n_2}$ whose rank is smaller or equal to $2r$ (and use Gordon's Theorem).

Proposition 5.15

$$\omega\left(\left\{X \,:\, X \in \mathbb{R}^{n_1 \times n_2},\ \operatorname{rank}(X) \leq r\right\}\right) \lesssim \sqrt{r(d_1 + d_2)}.$$

Proof.

$$\omega\left(\left\{X \,:\, X \in \mathbb{R}^{n_1 \times n_2},\ \operatorname{rank}(X) \leq r\right\}\right) = \mathbb{E} \max_{\substack{\|X\|_F = 1 \\ \operatorname{rank}(X) \leq r}} \operatorname{Tr}(GX).$$

Let $X = U \Sigma V^T$ be the SVD decomposition of X, then

$$\omega\left(\left\{X \,:\, X \in \mathbb{R}^{n_1 \times n_2},\ \operatorname{rank}(X) \leq r\right\}\right) = \mathbb{E} \max_{\substack{U^T U = V^T V = I_{r \times r} \\ \Sigma \in \mathbb{R}^{r \times r} \text{ diagonal } \|\Sigma\|_F = 1}} \operatorname{Tr}(\Sigma\left(V^T G U\right)).$$

This implies that

$$\omega\left(\left\{X \,:\, X \in \mathbb{R}^{n_1 \times n_2},\ \operatorname{rank}(X) \leq r\right\}\right) \leq (\operatorname{Tr} \Sigma)\left(\mathbb{E}\|G\|\right) \lesssim \sqrt{r}\left(\sqrt{n_1} + \sqrt{n_1}\right),$$

where the last inequality follows from bounds on the largest eigenvalue of a Wishart matrix, such as the ones used on Lecture 1. □

6 Compressed Sensing and Sparse Recovery

Most of us have noticed how saving an image in JPEG dramatically reduces the space it occupies in our hard drives (as oppose to file types that save the pixel 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 pixel 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. This special structure tends to exploited via sparsity. Indeed, natural images are known to be 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 $x \in \mathbb{R}^N$ as the signal corresponding to the image already in the basis for which it is sparse. Let's say that x is s-sparse, or $\|x\|_0 \leq s$, meaning that x has, at most, s non-zero components and, usually, $s \ll N$. The ℓ_0 norm[26] $\|x\|_0$ of a vector x is the number of non-zero entries of x. This means that, when we take a picture, our camera makes N measurements (each corresponding to a pixel) but then, after an appropriate change of basis, it keeps only $s \ll N$ non-zero coefficients and drops the others. This motivates the question: "If only a few degrees of freedom are kept after compression, why not measure in a more efficient way and take considerably less than N measurements?". This question is in the heart of Compressed Sensing [CRT06a, CRT06b, CT05, CT06, Don06, FR13]. It is particularly important in MRI imaging [**?**] as less measurements potentially means less measurement time. The following book is a great reference on Compressed Sensing [FR13].

More precisely, given a s-sparse vector x, we take $s < M \ll N$ linear measurements $y_i = a_i^T x$ and the goal is to recover x from the underdetermined system:

$$\left[\begin{array}{c} \\ y \\ \\ \end{array}\right] = \left[\begin{array}{ccc} & & \\ & A & \\ & & \end{array}\right] \left[\begin{array}{c} \\ \\ \\ x \\ \\ \\ \\ \end{array}\right].$$

Last lecture we used Gordon's theorem to show that, if we took random measurements, on the order of $s\log\left(\frac{N}{s}\right)$ measurements suffice to have all considerably different s-sparse signals correspond to considerably different sets of measurements. This suggests that $\approx s\log\left(\frac{N}{s}\right)$ may be enough to recover x, we'll see (later) in this lecture that this intuition is indeed correct.

Since the system is underdetermined and we know x is sparse, the natural thing to try, in order to recover x, is to solve

$$\begin{array}{ll} \min & \|z\|_0 \\ s.t. & Az = y, \end{array} \tag{57}$$

and hope that the optimal solution z corresponds to the signal in question x. Unfortunately, (57) is known to be a computationally hard problem in general. Instead, the approach usually taken in sparse recovery is to consider a convex relaxation of the ℓ_0 norm, the ℓ_1 norm: $\|z\|_1 = \sum_{i=1}^N |z_i|$. Figure 19

[26] The ℓ_0 norm is not actually a norm though.

depicts how the ℓ_1 norm can be seen as a convex relaxation of the ℓ_0 norm and how it promotes sparsity.

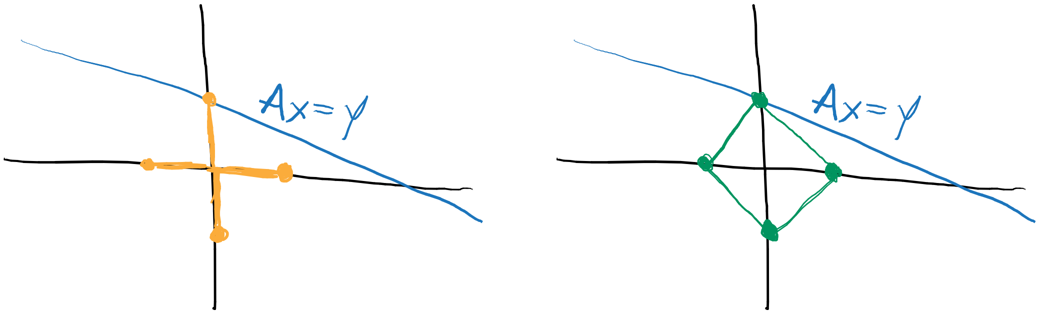


Figure 19: A two-dimensional depiction of ℓ_0 and ℓ_1 minimization. In ℓ_1 minimization (the picture of the right), one inflates the ℓ_1 ball (the diamond) until it hits the affine subspace of interest, this image conveys how this norm promotes sparsity, due to the pointy corners on sparse vectors.

This motivates one to consider the following optimization problem (surrogate to (57)):

$$\begin{array}{ll} \min & \|z\|_1 \\ s.t. & Az = y, \end{array} \tag{58}$$

In order for (58) we need two things, for the solution of it to be meaningful (hopefully to coincide with x) and for (58) to be efficiently solved.

We will formulate (58) as a Linear Program (and thus show that it is efficiently solvable). Let us think of ω^+ as the positive part of x and ω^- as the symmetric of the negative part of it, meaning that $x = \omega^+ - \omega^-$ and, for each i, either ω_i^- or ω_i^+ is zero. Note that, in that case,

$$\|x\|_1 = \sum_{i=1}^N \omega_i^+ + \omega_i^- = \mathbf{1}^T \left(\omega^+ + \omega^-\right).$$

Motivated by this we consider:

$$\begin{array}{ll} \min & \mathbf{1}^T \left(\omega^+ + \omega^-\right) \\ s.t. & A\left(\omega^+ - \omega^-\right) = y \\ & \omega^+ \geq 0 \\ & \omega^- \geq 0, \end{array} \tag{59}$$

which is a linear program. It is not difficult to see that the optimal solution of (59) will indeed satisfy that, for each i, either ω_i^- or ω_i^+ is zero and it is indeed equivalent to (58). Since linear programs are efficiently solvable [VB04], this means that (58) efficiently.

6.1 Duality and exact recovery

The goal now is to show that, under certain conditions, the solution of (58) indeed coincides with x. We will do this via duality (this approach is essentially the same as the one followed in [Fuc04] for the real case, and in [Tro05] for the complex case.)

Let us start by presenting duality in Linear Programming with a game theoretic view point. The idea will be to reformulate (59) 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 $u^T\left(y - A\left(\omega^+ - \omega^-\right)\right)$). More precisely consider the following

$$\min_{\substack{\omega^+\\ \omega^-}} \max_{\substack{u\\ v^+\geq 0\\ v^-\geq 0}} \mathbf{1}^T\left(\omega^+ + \omega^-\right) - \left(v^+\right)^T\omega^+ - \left(v^-\right)^T\omega^- + u^T\left(y - A\left(\omega^+ - \omega^-\right)\right). \tag{60}$$

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 (59) = (60).

Now image that we switch the order at 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 (60) $\geq$ (61), where (61) is given by:

$$\max_{\substack{u\\ v^+\geq 0\\ v^-\geq 0}} \min_{\substack{\omega^+\\ \omega^-}} \mathbf{1}^T\left(\omega^+ + \omega^-\right) - \left(v^+\right)^T\omega^+ - \left(v^-\right)^T\omega^- + u^T\left(y - A\left(\omega^+ - \omega^-\right)\right). \tag{61}$$

Rewriting

$$\max_{\substack{u\\ v^+\geq 0\\ v^-\geq 0}} \min_{\substack{\omega^+\\ \omega^-}} \left(\mathbf{1} - v^+ - A^Tu\right)^T\omega^+ + \left(\mathbf{1} - v^- + A^Tu\right)^T\omega^- + u^Ty \tag{62}$$

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 (62) is equivalent to

$$\max_{\substack{u\\ v^+\geq 0\\ v^-\geq 0\\ \mathbf{1}-v^+-A^Tu=0\\ \mathbf{1}-v^-+A^Tu=0}} \quad u^Ty$$

or equivalently,

$$\begin{array}{ll} \max_u & u^Ty \\ s.t. & -\mathbf{1} \leq A^Tu \leq \mathbf{1}. \end{array} \tag{63}$$

The linear program (63) is known as the dual program to (59). The discussion above shows that (63) $\leq$ (59) 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

$$0 \leq \left(\mathbf{1}^T - u^T A\right)\omega^+ + \left(\mathbf{1}^T + u^T A\right)\omega^- = 1^T\left(\omega^+ + \omega^-\right) - u^T\left[A\left(\omega^+ - \omega^-\right)\right] = 1^T\left(\omega^+ + \omega^-\right) - u^T y, \tag{64}$$

showing that (63) $\leq$ (59).

6.2 Finding a dual certificate

In order to show that $\omega^+ - \omega^- = x$ is an optimal solution[27] to (59), 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 (64) 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 have 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 6.1 *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 6.2 *Consider the problem of sparse recovery discussed this lecture. 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 program 58.

Since we know that $\left(A^T u\right)_S = \operatorname{sign}\left(x_S\right)$ (and that A_S is injective), we'll try to construct 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 6.3 *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 program 58.

[27] For now we will focus on showing that it is an optimal solution, see Remark 6.1 for a brief discussion of how to strengthen the argument to show uniqueness

Recall the definition of $A \in \mathbb{R}^{M \times N}$ satisfying the restricted isometry property from last Lecture.

Definition 6.4 (Restricted Isometry Property) *A matrix $A \in \mathbb{R}^{M \times N}$ is said to satisfy the (s, δ) restricted isometry property if*

$$(1-\delta)\|x\|^2 \leq \|Ax\|^2 \leq (1+\delta)\|x\|^2,$$

for all two s-sparse vectors x.

Last lecture (Lecture 5) we showed that if $M \gg s \log\left(\frac{N}{s}\right)$ and $A \in \mathbb{R}^{M \times N}$ is drawn with i.i.d. gaussian entries $\mathcal{N}\left(0, \frac{1}{M}\right)$ then it will, with high probability, satisfy the $(s, 1/3)$-RIP. Note that, if A satisfies the (s, δ)-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

$$\|A_S\left(A_S^T A_S\right)^{-1} \operatorname{sign}(x_S)\| \leq \sqrt{1+\frac{1}{3}}\frac{3}{2} = \sqrt{3}\sqrt{s},$$

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

$$\operatorname{Prob}\left(\left|A_j^T A_S\left(A_S^T A_S\right)^{-1} \operatorname{sign}(x_S)\right| \geq \frac{1}{\sqrt{M}}\sqrt{3}\sqrt{s}t\right) \leq 2\exp\left(-\frac{t^2}{2}\right),$$

where A_j is the j-th column of A.

Union bound gives

$$\operatorname{Prob}\left(\left\|A_S^T A_S\left(A_S^T A_S\right)^{-1} \operatorname{sign}(x_S)\right\|_\infty \geq \frac{1}{\sqrt{M}}\sqrt{3}\sqrt{s}t\right) \leq 2N\exp\left(-\frac{t^2}{2}\right),$$

which implies

$$\operatorname{Prob}\left(\left\|A_S^T A_S\left(A_S^T A_S\right)^{-1} \operatorname{sign}(x_S)\right\|_\infty \geq 1\right) \leq 2N\exp\left(-\frac{\left(\frac{\sqrt{M}}{\sqrt{3s}}\right)^2}{2}\right) = \exp\left(-\frac{1}{2}\left[\frac{M}{3s} - 2\log(2N)\right]\right),$$

which means that we expect to exactly recover x via ℓ_1 minimization when $M \gg s\log(N)$, similarly to what was predicted by Gordon's Theorem last Lecture.

6.3 A different approach

Given x a sparse vector, we want to show that x is the unique optimal solution to

$$\begin{array}{ll} \min & \|z\|_1 \\ s.t. & Az = y, \end{array} \tag{65}$$

Let $S = \operatorname{supp}(x)$ and suppose that $z \neq x$ is an optimal solution of the ℓ_1 minimization problem. Let $v = z - x$, it satisfies

$$\|v + x\|_1 \leq \|x\|_1 \quad \text{and} \quad A(v+x) = Ax,$$

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 triangular 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 this concentrated on such few entries. This motivates the following definition.

Definition 6.5 (Null Space Property) *A is said to satisfy the s-Null Space Property 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 (58). Also, now that recovery is formulated in terms of certain vectors not belonging to the nullspace of A, one could again resort to Gordon's theorem. And indeed, Gordon's Theorem can be used to understand the number of necessary measurements to do sparse recovery[28] [CRPW12]. There is also an interesting approach based on Integral Geometry [ALMT14].

As it turns out one can show that the $\left(2s, \frac{1}{3}\right)$-RIP implies s-NSP [FR13]. We omit that proof as it does not appear to be as enlightening (or adaptable) as the approach that was shown here.

6.4 Partial Fourier matrices satisfying the Restricted Isometry Property

While the results above are encouraging, rarely one has the capability of designing random gaussian measurements. A more realistic measurement design is to use rows of the Discrete Fourier Transform: Consider the random $M \times N$ matrix obtained by drawing rows uniformly with replacement from the $N \times N$ discrete Fourier transform matrix. It is known [CT06] that if $M = \Omega_\delta(K \operatorname{polylog} N)$, then the resulting partial Fourier matrix satisfies the restricted isometry property with high probability.

A fundamental problem in compressed sensing is determining the order of the smallest number M of random rows necessary. To summarize the progress to date, Candès and Tao [CT06] first found that $M = \Omega_\delta(K \log^6 N)$ rows suffice, then Rudelson and Vershynin [RV08] proved $M = \Omega_\delta(K \log^4 N)$, and more recently, Bourgain [Bou14] achieved $M = \Omega_\delta(K \log^3 N)$; Nelson, Price and Wootters [NPW14] also achieved $M = \Omega_\delta(K \log^3 N)$, but using a slightly different measurement matrix. The latest result is due to Haviv and Regev [HR] giving an upper bound of $M = \Omega_\delta(K \log^2 k \log N)$. As far as lower bounds, in [BLM15] it was shown that $M = \Omega_\delta(K \log N)$ is necessary. This draws a contrast with random Gaussian matrices, where $M = \Omega_\delta(K \log(N/K))$ is known to suffice.

Open Problem 6.1 *Consider the random $M \times N$ matrix obtained by drawing rows uniformly with replacement from the $N \times N$ discrete Fourier transform matrix. How large does M need to be so that, with high probability, the result matrix satisfies the Restricted Isometry Property (for constant δ)?*

6.5 Coherence and Gershgorin Circle Theorem

Last lectures we discussed the problem of building deterministic RIP matrices (building deterministic RIP matrices is particularly important because checking whether a matrix is RIP is computationally

[28] In these references the sets considered are slightly different than the one described here, as the goal is to ensure recovery of just one sparse vector, and not all of them simultaneously.

hard [BDMS13, TP13]). Despite suboptimal, coherence based methods are still among the most popular ways of building RIP matrices, we'll briefly describe some of the ideas involved here.

Recall the definition of the Restricted Isometry Property (Definition 6.4). Essentially, it asks that any $S \subset [N]$, $|S| \leq s$ 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 $\left\|A_S^T A_S - I\right\| \leq \delta(s-1)$.

More precisely, given a symmetric matrix B, the Gershgorin's circle theorem [HJ85] tells 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: $\|B\| \leq \max_i |B_{ij}|$.

Given a set of N vectors $a_1, \ldots, a_N \in \mathbb{R}^M$ we define its worst-case coherence μ as

$$\mu = \max_{i \neq j} \left|a_i^T a_j\right|$$

Given a set of unit-norm vectors $a_1, \ldots, a_N \mathbb{R}^M$ with worst-case coherence μ, if we form a matrix with these vectors as columns, then it will be $(s, \mu(s-1)\mu)$-RIP, meaning that it will be $\left(s, \frac{1}{3}\right)$- RIP for $s \leq \frac{1}{3}\frac{1}{\mu}$.

6.5.1 Mutually Unbiased Bases

We note that now we will consider our vectors to be complex valued, rather than real valued, but all of the results above hold for either case.

Consider the following $2d$ vectors: the d vectors from the identity basis and the d orthonormal vectors corresponding to columns of the Discrete Fourier Transform F. Since these basis are both orthonormal the vectors in question are unit-norm and within the basis are orthogonal, it is also easy to see that the inner product between any two vectors, one from each basis, has absolute value $\frac{1}{\sqrt{d}}$, meaning that the worst case coherence of this set of vectors is $\mu = \frac{1}{\sqrt{d}}$ this corresponding matrix $[I\ F]$ is RIP for $s \approx \sqrt{d}$.

It is easy to see that $\frac{1}{\sqrt{d}}$ coherence is the minimum possible between two orthonormal bases in $\mathbb{C}^d$, such bases are called unbiased (and are important in Quantum Mechanics, see for example [BBRV01]) This motivates the question of how many orthonormal basis can be made simultaneously (or mutually) unbiased in $\mathbb{C}^d$, such sets of bases are called mutually unbiased bases. Let $\mathcal{M}(d)$ denote the maximum number of such bases. It is known that $\mathcal{M}(d) \leq d+1$ and that this upper bound is achievable when d is a prime power, however even determining the value of $\mathcal{M}(6)$ is open [BBRV01].

Open Problem 6.2 *How many mutually unbiased bases are there in 6 dimensions? Is it true that* $\mathcal{M}(6) < 7$*?*[29]

6.5.2 Equiangular Tight Frames

Another natural question is whether one can get better coherence (or more vectors) by relaxing the condition that the set of vectors have to be formed by taking orthonormal basis. A tight frame (see, for example, [CK12] for more on Frame Theory) is a set of N vectors in $\mathbb{C}^M$ (with $N \geq M$) that spans $\mathbb{C}^M$ "equally". More precisely:

Definition 6.6 (Tight Frame) $v_1, \dots, v_N \in \mathbb{C}^M$ *is a tight frame if there exists a constant* α *such that*

$$\sum_{k=1}^{N} |\langle v_k, x\rangle|^2 = \alpha \|x\|^2, \quad \forall_{x\in\mathbb{C}^M},$$

or equivalently

$$\sum_{k=1}^{N} v_k v_k^T = \alpha I.$$

The smallest coherence of a set of N unit-norm vectors in M dimensions is bounded below by the Welch bound (see, for example, [BFMW13]) which reads:

$$\mu \geq \sqrt{\frac{N-M}{M(N-1)}}.$$

One can check that, due to this limitation, deterministic constructions based on coherence cannot yield matrices that are RIP for $s \gg \sqrt{M}$, known as the square-root bottleneck [BFMW13, Tao07].

There are constructions that achieve the Welch bound, known as Equiangular Tight Frames (ETFs), these are tight frames for which all inner products between pairs of vectors have the same modulus $\mu = \sqrt{\frac{N-M}{M(N-1)}}$, meaning that they are "equiangular". It is known that for there to exist an ETF in $\mathbb{C}^M$ one needs $N \leq M^2$ and ETF's for which $N = M^2$ are important in Quantum Mechanics, and known as SIC-POVM's. However, they are not known to exist in every dimension (see here for some recent computer experiments [SG10]). This is known as Zauner's conjecture.

Open Problem 6.3 *Prove or disprove the SIC-POVM / Zauner's conjecture: For any* d*, there exists an Equiangular tight frame with* d^2 *vectors in* $\mathbb{C}^d$ *dimensions. (or, there exist* d^2 *equiangular lines in* $\mathbb{C}^d$*).*

We note that this conjecture was recently shown [Chi15] for $d = 17$ and refer the reader to this interesting remark [Mix14c] on the description length of the constructions known for different dimensions.

[29] The author thanks Bernat Guillen Pegueroles for suggesting this problem.

6.5.3 The Paley ETF

There is a simple construction of an ETF made of $2M$ vectors in M dimensions (corresponding to a $M \times 2M$ matrix) known as the Paley ETF that is essentially a partial Discrete Fourier Transform matrix. While we refer the reader to [BFMW13] for the details the construction consists of picking rows of the $p \times p$ Discrete Fourier Transform (with $p \cong 1 \mod 4$ a prime) with indices corresponding to quadratic residues modulo p. Just by coherence considerations this construction is known to be RIP for $s \approx \sqrt{p}$ but conjectured [BFMW13] to be RIP for $s \approx \frac{p}{\text{polylog} p}$, which would be predicted if the choice os rows was random (as discussed above)[30]. Although partial conditional (conditioned on a number theory conjecture) progress on this conjecture has been made [BMM14] no unconditional result is known for $s \ll \sqrt{p}$. This motivates the following Open Problem.

Open Problem 6.4 *Does the Paley Equiangular tight frame satisfy the Restricted Isometry Property pass the square root bottleneck? (even by logarithmic factors?).*

We note that [BMM14] shows that improving polynomially on this conjecture implies an improvement over the Paley clique number conjecture (Open Problem 8.4.)

6.6 The Kadison-Singer problem

The Kadison-Singer problem (or the related Weaver's conjecture) was one of the main questions in frame theory, it was solved (with a non-constructive proof) in the recent breakthrough of Marcus, Spielman, and Srivastava [MSS15b], using similar techniques to their earlier work [MSS15a]. Their theorem guarantees the existence of universal constants $\eta \geq 2$ and $\theta > 0$ s.t. for any tight frame $\omega_1, \dots, \omega_N \in \mathbb{C}^M$ satisfying $\|\omega_k\| \leq 1$ and

$$\sum_{k=1}^{N} \omega_k \omega_k^T = \eta I,$$

there exists a partition of this tight frame $S_1, S_2 \subset [N]$ such that each is "almost a tight frame" in the sense that,

$$\sum_{k \in S_j} \omega_k \omega_k^T \preceq (\eta - \theta) I.$$

However, a constructive prove is still not known and there is no known (polynomial time) method that is known to construct such partitions.

Open Problem 6.5 *Give a (polynomial time) construction of the tight frame partition satisfying the properties required in the Kadison-Singer problem (or the related Weaver's conjecture).*

[30] We note that the quadratic residues are known to have pseudorandom properties, and indeed have been leveraged to reduce the randomness needed in certain RIP constructions [BFMM14]

7 Group Testing and Error-Correcting Codes

7.1 Group Testing

During the Second World War the United States was interested in weeding out all syphilitic soldiers called up for the army. However, syphilis testing back then was expensive and testing every soldier individually would have been very costly and inefficient. A basic breakdown of a test is: 1) Draw sample from a given individual, 2) Perform required tests, and 3) Determine presence or absence of syphilis.

If there are n soldiers, this method of testing leads to n tests. If a significant portion of the soldiers were infected then the method of individual testing would be reasonable. The goal however, is to achieve effective testing in the more likely scenario where it does not make sense to test n (say $n = 100,000$) people to get k (say $k = 10$) positives.

Let's say that it was believed that there is only one soldier infected, then one could mix the samples of half of the soldiers and with a single test determined in which half the infected soldier is, proceeding with a binary search we could pinpoint the infected individual in $\log n$ tests. If instead of one, one believes that there are at most k infected people, then one could simply run k consecutive binary searches and detect all of the infected individuals in $k \log n$ tests. Which would still be potentially much less than n.

For this method to work one would need to observe the outcome of the previous tests before designing the next test, meaning that the samples have to be prepared adaptively. This is often not practical, if each test takes time to run, then it is much more efficient to run them in parallel (at the same time). This means that one has to non-adaptively design T tests (meaning subsets of the n individuals) from which it is possible to detect the infected individuals, provided there are at most k of them. Constructing these sets is the main problem in (Combinatorial) Group testing, introduced by Robert Dorfman [Dor43] with essentially the motivation described above.[31]

Let A_i be a subset of $[T] = \{1, \dots, T\}$ that indicates the tests for which soldier i participates. Consider $\mathbb{A}$ the family of n such sets $\mathbb{A} = \{A_1, \dots, A_n\}$. We say that $\mathbb{A}$ satisfies the k-disjunct property if no set in $\mathbb{A}$ is contained in the union of k other sets in $\mathbb{A}$. A test set designed in such a way will succeed at identifying the (at most k) infected individuals – the set of infected tests is also a subset of $[T]$ and it will be the union of the A_i's that correspond to the infected soldiers. If the set of infected tests contains a certain A_i then this can only be explained by the soldier i being infected (provided that there are at most k infected people).

Theorem 7.1 *Given n and k, there exists a family $\mathbb{A}$ satisfying the k-disjunct property for a number of tests*

$$T = \mathcal{O}\left(k^2 \log n\right).$$

Proof. We will use the probabilistic method. We will show that, for $T = Ck^2 \log n$ (where C is a universal constant), by drawing the family $\mathbb{A}$ from a (well-chosen) distribution gives a $k-$disjunct family with positive probability, meaning that such a family must exist (otherwise the probability would be zero).

[31] in fact, our description for the motivation of Group Testing very much follows the description in [Dor43].

Let $0 \leq p \leq 1$ and let $\mathbb{A}$ be a collection of n random (independently drawn) subsets of $[T]$. The distribution for a random set A is such that each $t \in [T]$ belongs to A with probability p (and independently of the other elements).

Consider $k+1$ independent draws of this random variable, $A_0, \dots, A_k$. The probability that A_0 is contained in the union of A_1 through A_k is given by

$$\Pr\left[A_0 \subseteq (A_1 \cup \cdots \cup A_k)\right] = \left(1 - p(1-p)^k\right)^T .$$

This is minimized for $p = \frac{1}{k+1}$. For this choice of p, we have

$$1 - p(1-p)^k = 1 - \frac{1}{k+1}\left(1 - \frac{1}{k+1}\right)^k$$

Given that there are n such sets, there are $(k+1)\binom{n}{k+1}$ different ways of picking a set and k others to test whether the first is contained in the union of the other k. Hence, using a union bound argument, the probability that $\mathbb{A}$ is k-disjunct can be bounded as

$$\Pr[k - \text{disjunct}] \geq 1 - (k+1)\binom{n}{k+1}\left(1 - \frac{1}{k+1}\left(1 - \frac{1}{k+1}\right)^k\right)^T .$$

In order to show that one of the elements in $\mathbb{A}$ is k-disjunct we show that this probability is strictly positive. That is equivalent to

$$\left(1 - \frac{1}{k+1}\left(1 - \frac{1}{k+1}\right)^k\right)^T \leq \frac{1}{(k+1)\binom{n}{k+1}}.$$

Note that $\left(1 - \frac{1}{k+1}\right)^k \to e^{-1}\frac{1}{1-\frac{1}{k+1}} = e^{-1}\frac{k+1}{k}$, as $k \to \infty$. Thus, we only need

$$T \geq \frac{\log\left((k+1)\binom{n}{k+1}\right)}{-\log\left(1 - \frac{1}{k+1}e^{-1}\frac{k+1}{k}\right)} = \frac{\log\left(k\binom{n}{k+1}\right)}{-\log\left(1-(ek)^{-1}\right)} = \mathcal{O}(k^2\log(n/k)),$$

where the last inequality uses the fact that $\log\left(\binom{n}{k+1}\right) = \mathcal{O}\left(k\log\left(\frac{n}{k}\right)\right)$ due to Stirling's formula and the Taylor expansion $-\log(1-x^{-1})^{-1} = \mathcal{O}(x)$ □

This argument simply shows the existence of a family satisfying the k-disjunt property. However, it is easy to see that by having T slightly larger one can ensure that the probability that the random family satisfies the desired property can be made very close to 1.

Remarkably, the existence proof presented here is actually very close to the best known lower bound.

Theorem 7.2 *Given n and k, if there exists a family $\mathbb{A}$ of subsets of $[T]$ satisfying the k-disjunct property, then*

$$T = \Omega\left(\frac{k^2\log n}{\log k}\right).$$

Proof.

Fix a u such that $0 < u < \frac{T}{2}$; later it will be fixed to $u := \left\lceil (T-k)/\binom{k-1}{2} \right\rceil$. We start by constructing a few auxiliary family of sets. Let

$$\mathbb{A}_0 = \{A \in \mathbb{A} : |A| < u\},$$

and let $\mathbb{A}_1 \subseteq \mathbb{A}$ denote the family of sets in $\mathbb{A}$ that contain their own unique u-subset,

$$\mathbb{A}_1 := \left\{A \in \mathbb{A} : \exists F \subseteq A : |F| = u \text{ and, for all other } A' \in \mathbb{A},\ F \not\subseteq A'\right\}.$$

We will procede by giving an upper bound to $\mathbb{A}_0 \cup \mathbb{A}_1$. For that, we will need a couple of auxiliary family of sets. Let $\mathbb{F}$ denote the family of sets F in the definition of $\mathbb{A}_1$. More precisely,

$$\mathbb{F} := \{F \in [T] : |F| = u \text{ and } \exists! A \in \mathbb{A} : F \subseteq A\}.$$

By construction $|\mathbb{A}_1| \leq |\mathbb{F}|$

Also, let $\mathbb{B}$ be the family of subsets of $[T]$ of size u that contain an element of $\mathbb{A}_0$,

$$\mathbb{B} = \{B \subseteq [T] : |B| = u \text{ and } \exists A \in \mathbb{A}_0 \text{ such that } A \subseteq B\}.$$

We now prove that $|\mathbb{A}_0| \leq |\mathbb{B}|$. Let $\mathbb{B}'$ denote the family of subsets of $[T]$ of size u that are not in $\mathbb{B}$,

$$\mathbb{B}' = \left\{B' \subseteq [T] : |B'| = u \text{ and } B' \notin \mathbb{B}\right\}.$$

By construction of $\mathbb{A}_0$ and $\mathbb{B}$, no set in $\mathbb{B}'$ contains a set in $\mathbb{A}_0$ nor does a set in $\mathbb{A}_0$ contain a set in $\mathbb{B}'$. Also, both $\mathbb{A}_0$ and $\mathbb{B}'$ are antichains (or Sperner family), meaning that no pair of sets in each family contains each other. This implies that $\mathbb{A}_0 \cup \mathbb{B}'$ is an antichain containing only sets with u or less elements. The Lubell-Yamamoto-Meshalkin inequality [Yam54] directly implies that (as long as $u < \frac{T}{2}$) the largest antichain whose sets contain at most u elements is the family of subsets of $[T]$ of size u. This means that

$$|\mathbb{A}_0| + \left|\mathbb{B}'\right| = \left|\mathbb{A}_0 \cup \mathbb{B}'\right| \leq \binom{T}{u} = \left|\mathbb{B} \cup \mathbb{B}'\right| = |\mathbb{B}| + \left|\mathbb{B}'\right|.$$

This implies that $|\mathbb{A}_0| \leq |\mathbb{B}|$.

Because $\mathbb{A}$ satisfies the k-disjunct property, no two sets in $\mathbb{A}$ can contain eachother. This implies that the families $\mathbb{B}$ and $\mathbb{F}$ of sets of size u are disjoint which implies that

$$|\mathbb{A}_0 \cup \mathbb{A}_1| = |\mathbb{A}_0| + |\mathbb{A}_1| \leq |\mathbb{B}| + |\mathbb{F}| \leq \binom{T}{u}.$$

Let $\mathbb{A}_2 := \mathbb{A} \setminus (\mathbb{A}_0 \cup \mathbb{A}_1)$. We want to show that if $A \in \mathbb{A}_2$ and $A_1, \ldots, A_j \in \mathbb{A}$ we have

$$\left| A \setminus \bigcup_{i=1}^{j} A_i \right| > u(k-j). \tag{66}$$

This is readily shown by noting that if (66) did not hold then one could find $B_{j+1}, \ldots, B_k$ subsets of A of size t such that $A \setminus \bigcup_{i=1}^{j} A_i \subseteq \bigcup_{i=j+1}^{k} B_i$. Since A as no unique subsets of size t there must exist

$A_{j+1}, \dots, A_k \in \mathbb{A}$ such that $B_i \subseteq A_i$ for $i = j+1, \dots, k$. This would imply that $A \subseteq \bigcup_{i=1}^k A_i$ which would contradict the k-disjunct property.

If $|\mathbb{A}_2| > k$ then we can take $A_0, A_1, \dots, A_k$ distinct elements of $\mathbb{A}_2$. For this choice and any $j = 0, \dots, k$

$$\left| A_j \setminus \bigcup_{0 \le i < j} A_i \right| \ge 1 + u(k-j).$$

This means that

$$\left| \bigcup_{j=0}^{k} A_j \right| = \sum_{j=0,\dots,k} \left| A_j \setminus \bigcup_{0 \le i < j} A_i \right| \ge \sum_{j=0,\dots,k} [1 + u(k-j)] = 1 + k + u\binom{k+1}{2}.$$

Since all sets in $\mathbb{A}$ are subsets of $[T]$ we must have $1 + k + u\binom{k+1}{2} \le \left| \bigcup_{j=0}^k A_j \right| \le T$. On the other hand, taking

$$u := \left\lceil (T-k) / \binom{k+1}{2} \right\rceil$$

gives a contradition (note that this choice of u is smaller than $\frac{T}{2}$ as long as $k > 2$). This implies that $|\mathbb{A}_2| \le k$ which means that

$$n = |\mathbb{A}| = |\mathbb{A}_0| + |\mathbb{A}_1| + |\mathbb{A}_2| \le k + \binom{T}{u} = k + \binom{T}{\left\lceil (T-k)/\binom{k+1}{2} \right\rceil}.$$

This means that

$$\log n \le \log \left(k + \binom{T}{\left\lceil (T-k)/\binom{k+1}{2} \right\rceil} \right) = O\left(\frac{T}{k^2} \log k \right),$$

which concludes the proof of the theorem.

□

We essentially borrowed the proof of Theorem 7.2 from [Fur96]. We warn the reader however that the notation in [Fur96] is drastically different than ours, T corresponds to the number of people and n to the number of tests.

There is another upper bound, incomparable to the one in Theorem 7.1 that is known.

Theorem 7.3 *Given n and k, there exists a family $\mathbb{A}$ satisfying the k-disjunct property for a number of tests*

$$T = \mathcal{O}\left(k^2 \left(\frac{\log n}{\log k} \right)^2 \right).$$

The proof of this Theorem uses ideas of Coding Theory (in particular Reed-Solomon codes) so we will defer it for next section, after a crash course on coding theory.

The following Corollary follows immediately.

Corollary 7.4 *Given n and k, there exists a family $\mathbb{A}$ satisfying the k-disjunct property for a number of tests*

$$T = \mathcal{O}\left(\frac{k^2 \log n}{\log k} \min\left\{\log k, \frac{\log n}{\log k}\right\}\right).$$

While the upper bound in Corollary 7.4 and the lower bound in Theorem 7.2 are quite close, there is still a gap. This gap was recently closed and Theorem 7.2 was shown to be optimal [DVPS14] (original I was not aware of this reference and closing this gap was posed as an open problem).

Remark 7.5 *We note that the lower bounds established in Theorem 7.2 are not an artifact of the requirement of the sets being k-disjunct. For the measurements taken in Group Testing to uniquely determine a group of k infected individuals it must be that the there are no two subfamilies of at most k sets in $\mathbb{A}$ that have the same union. If $\mathbb{A}$ is not $k-1$-disjunct then there exists a subfamily of $k-1$ sets that contains another set A, which implies that the union of that subfamily is the same as the union of the same subfamily together with A. This means that a measurement system that is able to uniquely determine a group of k infected individuals must be $k-1$-disjunct.*

7.2 Some Coding Theory and the proof of Theorem 7.3

In this section we (very) briefly introduce error-correcting codes and use Reed-Solomon codes to prove Theorem 7.3. We direct the reader to [GRS15] for more on the subject.

Lets say Alice wants to send a message to Bob but they can only communicate through a channel that erases or replaces some of the letters in Alice's message. If Alice and Bob are communicating with an alphabet Σ and can send messages with lenght N they can pre-decide a set of allowed messages (or codewords) such that even if a certain number of elements of the codeword gets erased or replaced there is no risk for the codeword sent to be confused with another codeword. The set C of codewords (which is a subset of Σ^N) is called the codebook and N is the blocklenght.

If every two codewords in the codebook differs in at least d coordinates, then there is no risk of confusion with either up to $d-1$ erasures or up to $\lfloor \frac{d-1}{2} \rfloor$ replacements. We will be interested in codebooks that are a subset of a finite field, meanign that we will take Σ to be $\mathbb{F}_q$ for q a prime power and C to be a linear subspace of $\mathbb{F}_q$.

The dimension of the code is given by

$$m = \log_q |C|,$$

and the rate of the code by

$$R = \frac{m}{N}.$$

Given two code words c_1, c_2 the Hamming distance $\Delta(c_1, c_2)$ is the number of entries where they differ. The distance of a code is defined as

$$d = \min_{c_1 \neq c_2 \in C} \Delta(c_1, c_2).$$

For linear codes, it is the same as the minimum weight

$$\omega(C) = \min_{c \in C \setminus \{0\}} \Delta(c).$$

We say that a linear code C is a $[N,m,d]_q$ code (where N is the blocklenght, m the dimension, d the distance, and $\mathbb{F}^q$ the alphabet.

One of the main goals of the theory of error-correcting codes is to understand the possible values of rates, distance, and q for which codes exist. We simply briefly mention a few of the bounds and refer the reader to [GRS15]. An important parameter is given by the entropy function:

$$H_q(x) = x\frac{\log(q-1)}{\log q} - x\frac{\log x}{\log q} - (1-x)\frac{\log(1-x)}{\log q}.$$

- Hamming bound follows essentially by noting that if a code has distance d then balls of radius $\lfloor \frac{d-1}{2} \rfloor$ centered at codewords cannot intersect. It says that

$$R \leq 1 - H_q\left(\frac{1}{2}\frac{d}{N}\right) + o(1)$$

- Another particularly simple bound is Singleton bound (it can be easily proven by noting that the first $n+d+2$ of two codewords need to differ in at least 2 coordinates)

$$R \leq 1 - \frac{d}{N} + o(1).$$

There are probabilistic constructions of codes that, for any $\epsilon > 0$, satisfy

$$R \geq 1 - H_q\left(\frac{d}{N}\right) - \epsilon.$$

This means that R^* the best rate achievable satisties

$$R^* \geq 1 - H_q\left(\frac{d}{N}\right), \tag{67}$$

known as the GilbertVarshamov (GV) bound [Gil52, Var57]. Even for $q=2$ (corresponding to binary codes) it is not known whether this bound is tight or not, nor are there deterministic constructions achieving this Rate. This motivates the following problem.

Open Problem 7.1 *1. Construct an explicit (deterministic) binary code ($q=2$) satisfying the GV bound* (67).

2. Is the GV bound tight for binary codes ($q=2$)?

7.2.1 Boolean Classification

A related problem is that of Boolean Classification [AABS15]. Let us restrict our attention to In error-correcting codes one wants to build a linear codebook that does not contain a codeword with weight $\leq d-1$. In other words, one wants a linear codebook C that does intersect $B(d-1) = \{x \in \{0,1\}^n : 0 < \Delta(x) \leq d-1\}$ the pinched Hamming ball of radius d (recall that $\Delta(d)$ is the Hamming weight of x, meaning the number of non-zero entries). In the Boolean Classification problem one is willing to confuse two codewords as long as they are sufficiently close (as this is likely to mean they are in the same group, and so they are the same from the point of view of classification). The objective then becomes understanding what is the largest possible rate of a codebook that avoids an Annulus $A(a,b) = \{x \in \{0,1\}^n : a \leq \Delta(x) \leq b\}$. We refer the reader to [AABS15] for more details. Let us call that rate

$$R_A^*(a,b,n).$$

Note that $R_A^*(1,d-1,n)$ corresponds to the optimal rate for a binary error-correcting code, conjectured to be $1 - H_q\left(\frac{d}{N}\right)$ (The GV bound).

Open Problem 7.2 *It is conjectured in [AABS15] (Conjecture 3 in [AABS15]) that the optimal rate in this case is given by*

$$R_A^*(\alpha n, \beta n, n) = \alpha + (1-\alpha)\, R_A^*\left(1, \beta n, (1-\alpha)\right) + o(1),$$

where $o(1)$ goes to zero as n goes to infinity.

This is established in [AABS15] for $\beta \geq 2\alpha$ but open in general.

7.2.2 The proof of Theorem 7.3

Reed-Solomon codes[RS60] are $[n, m, n-m+1]_q$ codes, for $m \leq n \leq q$. They meet the Singleton bound, the drawback is that they have very large q ($q > n$). We'll use their existence to prove Theorem 7.3

Proof. [of Theorem 7.3]

We will construct a family $\mathbb{A}$ of sets achieving the upper bound in Theorem 7.3. We will do this by using a Reed-Solomon code $[q, m, q-m+1]_q$. This code has q^m codewords. To each codework c we will correspond a binary vector a of length q^2 where the i-th q-block of a is the indicator of the value of $c(i)$. This means that a is a vector with exactly q ones (and a total of q^2 entries)[32]. We construct the family $\mathbb{A}$ for $T = q^2$ and $n = q^m$ (meaning q^m subsets of $\left[q^2\right]$) by constructing, for each codeword c, the set of non-zero entries of the corresponding binary vector a.

These sets have the following properties,

$$\min_{j\in[n]} |A_j| = q,$$

and

$$\max_{j_1 \neq j_2 \in [n]} |A_{j_1} \cap A_{j_2}| = q - \min_{c_1 \neq c_2 \in C} \Delta(c_1, c_2) \leq q - (q-m+1) = m-1.$$

This readily implies that $\mathbb{A}$ is k-disjunct for

$$k = \left\lfloor \frac{q-1}{m-1} \right\rfloor,$$

because the union of $\left\lfloor \frac{q-1}{m-1} \right\rfloor$ sets can only contain $(m-1)\left\lfloor \frac{q-1}{m-1} \right\rfloor < q$ elements of another set.

Now we pick $q \approx 2k\frac{\log n}{\log k}$ (q has to be a prime but there is always a prime between this number and its double by Bertrand's postulate (see [?] for a particularly nice proof)). Then $m = \frac{\log n}{\log q}$ (it can be taken to be the ceiling of this quantity and then n gets updated accordingly by adding dummy sets).

[32] This is precisely the idea of code concatenation [GRS15]

This would give us a family (for large enough parameters) that is k-disjunct for

$$\begin{aligned}\left\lfloor \frac{q-1}{m-1} \right\rfloor &\geq \left\lfloor \frac{2k\frac{\log n}{\log k}-1}{\frac{\log n}{\log q}+1-1} \right\rfloor \\ &= \left\lfloor 2k\frac{\log q}{\log k} - \frac{\log q}{\log n} \right\rfloor \\ &\geq k.\end{aligned}$$

Noting that

$$T \approx \left(2k\frac{\log n}{\log k}\right)^2.$$

concludes the proof. □

7.3 In terms of ~~linear~~ Bernoulli algebra

We can describe the process above in terms of something similar to a sparse linear system. let 1_{A_i} be the $t-\textit{dimensional}$ indicator vector of A_i, $1_{i:n}$ be the (unknown) $n-$dimensional vector of infected soldiers and $1_{t:T}$ the $T-$dimensional vector of infected (positive) tests. Then

$$\begin{bmatrix} | & & | \\ 1_{A_1} & \cdots & 1_{A_n} \\ | & & | \end{bmatrix} \otimes \begin{bmatrix} | \\ | \\ 1_{i:n} \\ | \\ | \end{bmatrix} = \begin{bmatrix} | \\ 1_{t:T} \\ | \end{bmatrix},$$

where $\otimes$ is matrix-vector multiplication in the Bernoulli algebra, basically the only thing that is different from the standard matrix-vector multiplications is that the addition operation is replaced by binary "or", meaning $1 \oplus 1 = 1$.

This means that we are essentially solving a linear system (with this non-standard multiplication). Since the number of rows is $T = \mathcal{O}(k^2 \log(n/k))$ and the number or columns $n \gg T$ the system is underdetermined. Note that the unknown vector, $1_{i:n}$ has only k non-zero components, meaning it is $k-$sparse. Interestingly, despite the similarities with the setting of sparse recovery discussed in a previous lecture, in this case, $\tilde{\mathcal{O}}(k^2)$ measurements are needed, instead of $\tilde{\mathcal{O}}(k)$ as in the setting of Compressed Sensing.

7.3.1 Shannon Capacity

The goal Shannon Capacity is to measure the amount of information that can be sent through a noisy channel where some pairs of messages may be confused with eachother. Given a graph G (called the confusion graph) whose vertices correspond to messages and edges correspond to messages that may be confused with each other. A good example is the following: say one has a alphabet of five symbols $1,2,3,4,5$ and that each digit can be confused with the immediately before and after (and 1 and 5 can be confused with eachother). The confusion graph in this case is C_5, the cyclic graph on 5 nodes. It is easy to see that one can at most send two messages of one digit each without confusion, this corresponds to the independence number of C_5, $\alpha(C_5) = 2$. The interesting question arises when asking how many different words of two digits can be sent, it is clear that one can send at least $\alpha(C_5)^2 = 4$ but the remarkable fact is that one can send 5 (for example: "11","23","35","54", or "42"). The confusion graph for the set of two digit words $C_5^{\oplus 2}$ can be described by a product of the original graph C_5 where for a graph G on n nodes $G^{\oplus 2}$ is a graph on n nodes where the vertices are indexed by pairs ij of vertices of G and

$$(ij, kl) \in E\left(G^{\oplus 2}\right)$$

if both a) $i = k$ or $i, k \in E$ and b) $j = l$ or $j, l \in E$ hold.

The above observation can then be written as $\alpha\left(C_5^{\oplus 2}\right) = 5$. This motivates the definition of Shannon Capacity [Sha56]

$$\theta_S(G) \sup_k \left(G^{\oplus k}\right)^{\frac{1}{k}}.$$

Lovasz, in a remarkable paper [Lov79], showed that $\theta_S(C_5) = \sqrt{5}$, but determining this quantity is an open problem for many graphs of interested [AL06], including C_7.

Open Problem 7.3 *What is the Shannon Capacity of the 7 cycle?*

7.3.2 The deletion channel

In many applications the erasures or errors suffered by the messages when sent through a channel are random, and not adversarial. There is a beautiful theory understanding the amount of information that can be sent by different types of noisy channels, we refer the reader to [CT] and references therein for more information.

A particularly challenging channel to understand is the deletion channel. The following open problem will envolve a particular version of it. Say we have to send a binary string "10010" through a deletion channel and the first and second bits get deleted, then the message receive would be "010" and the receiver would not know which bits were deleted. This is in contrast with the erasure channel where bits are erased but the receiver knows which bits are missing (in the case above the message received would be "??010"). We refer the reader to this survey on many of the interesting questions (and results) regarding the Deletion channel [Mit09].

A particularly interesting instance of the problem is the Trace Reconstruction problem, where the same message is sent multiple times and the goal of the receiver is to find exactly the original message sent from the many observed corrupted version of it. We will be interested in the following quantity: Draw a random binary string with n bits, suppose the channel has a deletion probability of $\frac{1}{2}$ for each bit (independently), define $\mathcal{D}\left(n; \frac{1}{2}\right)$ has the number of times the receiver needs to receive the message (with independent corruptions) so that she can decode the message exactly, with high probability. It is easy to see that $\mathcal{D}\left(n; \frac{1}{2}\right) \leq 2^n$, since roughly once in every 2^n times the whole message will go through the channel unharmed. It is possible to show (see [HMPW]) that $\mathcal{D}\left(n; \frac{1}{2}\right) \leq 2^{\sqrt{n}}$ but it is not known whether this bound is tight.

Open Problem 7.4 *1. What are the asymptotics of $\mathcal{D}\left(n; \frac{1}{2}\right)$?*

2. *An interesting aspect of the Deletion Channel is that different messages may have different difficulties of decoding. This motivates the following question: What are the two (distinct) binary sequences* $x^{(2)}$ *and* $x^{(2)}$ *that are more difficult to distinguish (let's say that the receiver knows that either* $x^{(1)}$ *or* $x^{(2)}$ *was sent but not which)?*

8 Approximation Algorithms and Max-Cut

8.1 The Max-Cut problem

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 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 the typical ones!), or to design polynomial algorithms that, in any instance, produce guaranteed approximate solutions. This section is about the third option. The second is discussed in later in the course, in the context of community detection.

The `Max-Cut` problem is the following: Given a graph $G = (V, E)$ with non-negative weights w_{ij} on the edges, find a set $S \subset V$ for which $\mathrm{cut}(S)$ is maximal. Goemans and Williamson [GW95] introduced an approximation algorithm that runs in polynomial time and has a randomized component to 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

$$\begin{array}{ll} \max & \frac{1}{2}\sum_{i<j} w_{ij}(1 - y_i y_j) \\ s.t. & \quad |y_i| = 1 \end{array} \tag{68}$$

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:

$$\begin{array}{ll} \max & \frac{1}{2}\sum_{i<j} w_{ij}(1 - u_i^T u_j) \\ s.t. & \quad u_i \in \mathbb{R}^n, \|u_i\| = 1. \end{array} \tag{69}$$

This is in fact a relaxation because if we restrict u_i to be a multiple of e_1, the first element of the canonical basis, then (79) is equivalent to (68). For this to be a useful approach, the following two properties should hold:

(a) Problem (79) needs to be easy to solve.

(b) The solution of (79) needs to be, in some way, related to the solution of (68).

Definition 8.1 *Given a graph* G*, we define* $\mathrm{MaxCut}(G)$ *as the optimal value of* (68) *and* $\mathcal{R}\mathrm{MaxCut}(G)$ *as the optimal value of* (79).

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):

$$\begin{array}{ll} \max & \frac{1}{2}\sum_{i<j} w_{ij}(1 - X_{ij}) \\ s.t. & X \succeq 0 \text{ and } X_{ii} = 1,\ i = 1, \ldots, n. \end{array} \tag{70}$$

This SDP can be solved (up to ϵ accuracy) in time polynomial on the input size and $\log(\epsilon^{-1})$[VB96].

There is an alternative way of viewing (70) as a relaxation of (68). By taking $X = yy^T$ one can formulate a problem equivalent to (68)

$$\begin{array}{ll} \max & \frac{1}{2}\sum_{i<j} w_{ij}(1 - X_{ij}) \\ s.t. & X \succeq 0\ ,\ X_{ii} = 1,\ i = 1, \ldots, n, \text{ and } \operatorname{rank}(X) = 1. \end{array} \tag{71}$$

The SDP (70) can be regarded as a relaxation of (71) 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.

We now turn to property (b), and consider the problem of forming a solution to (68) from a solution to (70). From the solution $\{u_i\}_{i=1,\ldots,n}$ of the relaxed problem (70), 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

$$S' = \{i | r^T u_i \geq 0\}$$

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.

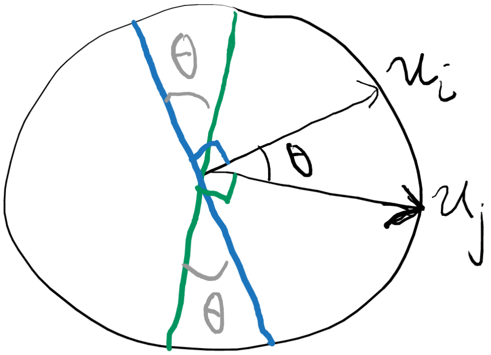


Figure 20: $\theta = \arccos(u_i^T u_j)$

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 20) to be given by

$$\begin{aligned} \mathbb{E}[W] &= \sum_{i<j} w_{ij} \Pr\left\{\operatorname{sign}(r^T u_i) \neq \operatorname{sign}(r^T u_j)\right\} \\ &= \sum_{i<j} w_{ij} \frac{1}{\pi} \arccos(u_i^T u_j). \end{aligned}$$

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 $\alpha_{GW} > 0.87$.

It is then clear that

$$\mathbb{E}[W] = \sum_{i<j} w_{ij}\frac{1}{\pi}\arccos(u_i^T u_j) \geq \alpha_{GW}\frac{1}{2}\sum_{i<j} w_{ij}(1-u_i^T u_j). \tag{72}$$

Let MaxCut(G) be the maximum cut of G, meaning the maximum of the original problem (68). Naturally, the optimal value of (79) is larger or equal than MaxCut(G). Hence, an algorithm that solves (79) and uses the random rounding procedure described above produces a cut W satisfying

$$\text{MaxCut}(G) \geq \mathbb{E}[W] \geq \alpha_{GW}\frac{1}{2}\sum_{i<j} w_{ij}(1-u_i^T u_j) \geq \alpha_{GW}\text{MaxCut}(G), \tag{73}$$

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.

Note that the above gives

$$\text{MaxCut}(G) \geq \mathbb{E}[W] \geq \alpha_{GW}\mathcal{R}\text{MaxCut}(G) \geq \alpha_{GW}\text{MaxCut}(G)$$

8.2 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}.

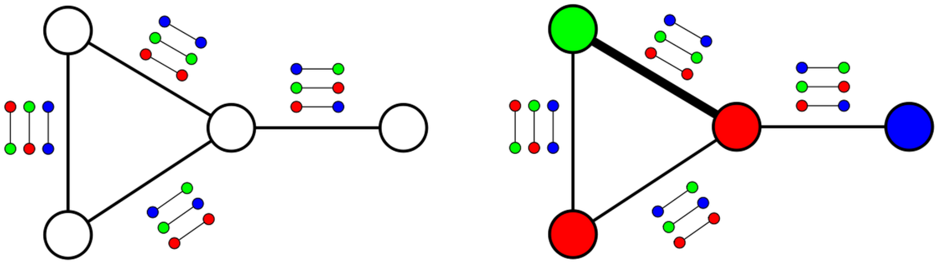


Figure 21: The Unique Games Problem

The unique games problem (as depicted in Figure 21) 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 21 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 [Kho02], has played a major role in hardness of approximation results.

Conjecture 8.2 *For any $\epsilon > 0$, the problem of distinguishing whether an instance of the Unique Games Problem is such that it is possible to agree with a $\geq 1 - \epsilon$ 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 [ABS10], 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.

Open Problem 8.1 *Is the Unique Games conjecture true? In particular, can it be refuted by a constant degree Sum-of-squares relaxation?*

Remarkably, approximating `Max-Cut` with an approximation ratio better than α_{GW} is has hard as refuting the Unique Games Conjecture (UG-hard) [KKMO05]. 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 [Rag08].

Not depending on the Unique Games Conjecture, there is a NP-hardness of approximation of $\frac{16}{17}$ for `Max-Cut` [Has02].

Remark 8.3 *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).*

8.3 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 $w_{ii} = 0$).

$$\begin{aligned}
\max_{y_i=\pm 1} \frac{1}{2}\sum_{i<j} w_{ij}(1 - y_iy_j) &= \max_{y_i=\pm 1} \frac{1}{4}\sum_{i,j} w_{ij}(1 - y_iy_j) \\
&= \max_{y_i=\pm 1} \frac{1}{4}\sum_{i,j} w_{ij}\left(\frac{y_i^2 + y_j^2}{2} - y_iy_j\right) \\
&= \max_{y_i=\pm 1} \frac{1}{4}\left(-\sum_{i,j} w_{ij}y_iy_j + \frac{1}{2}\sum_i\left[\sum_j w_{ij}\right]y_i^2 + \frac{1}{2}\sum_j\left[\sum_i w_{ij}\right]y_j^2\right) \\
&= \max_{y_i=\pm 1} \frac{1}{4}\left(-\sum_{i,j} w_{ij}y_iy_j + \frac{1}{2}\sum_i \deg(i)y_i^2 + \frac{1}{2}\sum_j \deg(j)y_j^2\right) \\
&= \max_{y_i=\pm 1} \frac{1}{4}\left(-\sum_{i,j} w_{ij}y_iy_j + \sum_i \deg(i)y_i^2\right) \\
&= \max_{y_i=\pm 1} \frac{1}{4}y^T L_G y,
\end{aligned}$$

where $L_G = D_G - W$ is the Laplacian matrix, D_G is a diagonal matrix with $(D_G)_{ii} = \deg(i) = \sum_j w_{ij}$ and $W_{ij} = w_{ij}$.

This means that we rewrite (68) as

$$\begin{array}{ll} \max & \frac{1}{4} y^T L_G y \\ & y_i = \pm 1, \; i = 1, \ldots, n. \end{array} \tag{74}$$

Similarly, (70) can be written (by taking $X = yy^T$) as

$$\begin{array}{ll} \max & \frac{1}{4} \operatorname{Tr}(L_G X) \\ s.t. & X \succeq 0 \\ & X_{ii} = 1, \; i = 1, \ldots, n. \end{array} \tag{75}$$

Indeed, given

Next lecture we derive the formulation of the dual program to (75) in the context of recovery in the Stochastic Block Model. Here we will simply show weak duality. The dual is given by

$$\begin{array}{ll} \min & \operatorname{Tr}(D) \\ s.t. & D \text{ is a diagonal matrix} \\ & D - \frac{1}{4} L_G \succeq 0. \end{array} \tag{76}$$

Indeed, if X is a feasible solution to (75) and D a feasible solution to (76) 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 (76) is larger than the one of (75).

If certain conditions, the so called Slater conditions [VB04, VB96], 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 (75) and the identity is such a matrix. This means that there exists $D^\natural$ feasible for (76) such that

$$\operatorname{Tr}(D^\natural) = \mathcal{R}\text{MaxCut}.$$

Hence, for any $y \in \mathbb{R}^n$ we have

$$\frac{1}{4} y^T L_G y = \mathcal{R}\text{MaxCut} - y^T \left(D^\natural - \frac{1}{4}L_G\right)^T + \sum_{i=1}^n D_{ii}\left(y_i^2 - 1\right). \tag{77}$$

Note that (77) certifies that no cut of G is larger than $\mathcal{R}$MaxCut. Indeed, if $y \in \{\pm 1\}^2$ then $y_i^2 = 1$ and so

$$\mathcal{R}\text{MaxCut} - \frac{1}{4} y^T L_G y = y^T \left(D^\natural - \frac{1}{4}L_G\right)^T.$$

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 $v_1, \dots, v_n$. This means that meaning that $y^T \left(D^\natural - \frac{1}{4}L_G\right)^T = \left\|V^T y\right\|^2 = \sum_{k=1}^n (v_k^T y)^2$. This means that, for $y \in \{\pm 1\}^2$,

$$\mathcal{R}\text{MaxCut} - \frac{1}{4}y^T L_G y = \sum_{k=1}^{n} (v_k^T y)^2.$$

In other words, $\mathcal{R}\text{MaxCut} - \frac{1}{4}y^T L_G y$ is, in the hypercube ($y \in \{\pm 1\}^2$) a sum-of-squares of degree 2. This is known as a sum-of-squares certificate [BS14, Bar14, Par00, Las01, Sho87, Nes00]; indeed, if a polynomial is a sum-of-squares naturally it is non-negative.

Note that, by definition, $\text{MaxCut} - \frac{1}{4}y^T L_G y$ is always non-negative on the hypercube. This does not mean, however, that it needs to be a sum-of-squares[33] of degree 2.

(A Disclaimer: the next couple of paragraphs are a bit hand-wavy, they contain some of intuition for the Sum-of-squares hierarchy but for details and actual formulations, please see the references.)

The remarkable fact is that, if one bounds the degree of the sum-of-squares certificate, it can be found using Semidefinite programming [Par00, Las01]. In fact, SDPs (76) and (76) 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 (77) and the primal is constraining the moments of degree 2 of y of the form $X_{ij} = y_i y_j$ (see [Bar14] for some nice lecture notes on Sum-of-Squares, see also Remark 8.4). This raises a natural question of whether, by allowing a sum-of-squares certificate of degree 4 (which corresponds to another, larger, SDP that involves all monomials of degree ≤ 4 [Bar14]) one can improve the approximation of α_{GW} to Max-Cut. Remarkably this is open.

Open Problem 8.2

1. *What is the approximation ratio achieved by (or the integrality gap of) the Sum-of-squares degree* 4 *relaxation of the Max-Cut problem?*
2. *The relaxation described above (of degree* 2*)* (76) *is also known to produce a cut of* $1 - \mathcal{O}\left(\sqrt{\epsilon}\right)$ *when a cut of* $1 - \epsilon$ *exists. Can the degree* 4 *relaxation improve over this?*
3. *What about other (constant) degree relaxations?*

Remark 8.4 (triangular inequalities and Sum of squares level 4) *A (simpler) natural question is wether the relaxation of degree* 4 *is actually strictly 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.*

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 (75) *of degree 2 is only able to constraint*

$$X_{ij} + X_{jk} + X_{ik} \geq -\frac{3}{2},$$

[33] This is related with Hilbert's 17th problem [Sch12] and Stengle's Positivstellensatz [Ste74]

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 $y_iy_j + y_jy_k + y_ky_i \geq -\frac{3}{2}$ can be proven using sum-of-squares proof with degree 2:

$$(y_i + y_j + y_k)^2 \geq 0 \quad \Rightarrow \quad y_iy_j + y_jy_k + y_ky_i \geq -\frac{3}{2}$$

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 $y_iy_j + y_jy_k + y_ky_i \geq -1$ can indeed be proven using sum-of-squares proof with degree 4 (recall that $y_i^2 = 1$):

$$(1 + y_iy_j + y_jy_k + y_ky_i)^2 \geq 0 \quad \Rightarrow \quad y_iy_j + y_jy_k + y_ky_i \geq -1.$$

Interestingly, it is known [KV13] that these extra inequalities alone will not increase the approximation power (in the worst case) of (70).

8.4 The Grothendieck Constant

There is a somewhat similar remarkable problem, known as the Grothendieck problem [AN04, AMMN05]. Given a matrix $A \in \mathbb{R}^{n\times m}$ the goal is to maximize

$$\begin{array}{ll} \max & x^TAy \\ s.t. & x_i = \pm, \forall_i \\ s.t. & y_j = \pm, \forall_j \end{array} \tag{78}$$

Note that this is similar to problem (68). In fact, if $A \succeq 0$ it is not difficult to see that the optimal solution of (78) satisfies $y = x$ and so if $A = L_G$, since $L_G \succeq 0$, (78) reduces to (68). In fact, when $A \succeq 0$ this problem is known as the little Grothendieck problem [AN04, CW04, BKS13a].

Even when A is not positive semidefinite, one can take $z^T = [x^T\, y^T]$ and the objective can be written as

$$z^T \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} z.$$

Similarly to the approximation ratio in Max-Cut, the Grothendieck constant [Pis11] K_G is the maximum ratio (over all matrices A) between the SDP relaxation

$$\begin{array}{ll} \max & \sum_{ij} A_{ij} u_i^T v_j \\ s.t. & u_i \in \mathbb{R}^{n+m}, \|u_i\| = 1, \\ & v_j \in \mathbb{R}^{n+m}, \|v_j\| = 1 \end{array} \tag{79}$$

and 78, and its exact value is still unknown, the best known bounds are available here [] and are $1.676 < K_G < \frac{\pi}{2\log(1+\sqrt{2})}$. See also page 21 here [F$^+$14]. There is also a complex valued analogue [Haa87].

Open Problem 8.3 *What is the real Grothendieck constant K_G?*

8.5 The Paley Graph

Let p be a prime such that $p \cong 1 \mod 4$. The Paley graph of order p is a graph on p nodes (each node associated with an element of $\mathbb{Z}_p$) where (i,j) is an edge if $i-j$ is a quadratic residue modulo p. In other words, (i,j) is an edge is there exists a such that $a^2 \cong i-j \mod p$. Let $\omega(p)$ denote the clique number of the Paley graph of order p, meaning the size of its largest clique. It is conjectured that $\omega(p) \lesssim \text{pollywog}(n)$ but the best known bound is $\omega(p) \leq \sqrt{p}$ (which can be easily obtained). The only improvement to date is that, infinitely often, $\omega(p) \leq \sqrt{p} - 1$, see [BRM13].

The theta function of a graph is a Semidefinite programming based relaxation of the independence number [Lov79] (which is the clique number of the complement graph). As such, it provides an upper bound on the clique number. In fact, this upper bound for Paley graph matches $\omega(p) \leq \sqrt{p}$.

Similarly to the situation above, one can define a degree 4 sum-of-squares analogue to $\theta(G)$ that, in principle, has the potential to giving better upper bounds. Indeed, numerical experiments in [GLV07] seem to suggest that this approach has the potential to improve on the upper bound $\omega(p) \leq \sqrt{p}$

Open Problem 8.4 *What are the asymptotics of the Paley Graph clique number $\omega(p)$? Can the the SOS degree 4 analogue of the theta number help upper bound it?* [34]

Interestingly, a polynomial improvement on Open Problem 6.4. is known to imply an improvement on this problem [BMM14].

8.6 An interesting conjecture regarding cuts and bisections

Given d and n let $G^{reg}(n,d)$ be a random d-regular graph on n nodes, drawn from the uniform distribution on all such graphs. An interesting question is to understand the typical value of the Max-Cut such a graph. The next open problem is going to involve a similar quantity, the Maximum Bisection. Let n be even, the Maximum Bisection of a graph G on n nodes is

$$\text{MaxBis}(G) = \max_{S\colon |S| = \frac{n}{2}} \text{cut}(S),$$

and the related Minimum Bisection (which will play an important role in next lectures), is given by

$$\text{MinBis}(G) = \min_{S\colon |S| = \frac{n}{2}} \text{cut}(S),$$

[34] The author thanks Dustin G. Mixon for suggesting this problem.

A typical bisection will cut half the edges, meaning $\frac{d}{4}n$. It is not surprising that, for large n, $\mathrm{MaxBis}(G)$ and $\mathrm{MinBis}(G)$ will both fluctuate around this value, the amazing conjecture [ZB09] is that their fluctuations are the same.

Conjecture 8.5 ([ZB09]) *Let $G \sim G^{reg}(n,d)$, then for all d, as n grows*

$$\frac{1}{n}\left(\mathrm{MaxBis}(G) + \mathrm{MinBis}(G)\right) = \frac{d}{2} + o(1),$$

where $o(1)$ is a term that goes to zero with n.

Open Problem 8.5 *Prove or disprove Conjecture 8.5.*

Recently, it was shown that the conjecture holds up to $o(\sqrt{d})$ terms [DMS15]. We also point the reader to this paper [Lyo14], that contains bounds that are meaningful already for $d = 3$.

9 Community detection and the Stochastic Block Model

9.1 Community Detection

Community detection in a network is a central problem in data science. A few lectures ago we discussed clustering and gave a performance guarantee for spectral clustering (based on Cheeger's Inequality) that was guaranteed to hold for any graph. While these guarantees are remarkable, they are worst-case guarantees and hence pessimistic in nature. In what follows we analyze the performance of a convex relaxation based algorithm on typical instances of the community detection problem (where typical is defined through some natural distribution of the input).

We focus on the problem of minimum graph bisection. The objective is to partition a graph in two equal-sized disjoint sets (S, S^c) while minimizing $\mathrm{cut}(S)$ (note that in the previous lecture, for the Max-Cut problem, we were maximizing it instead!).

9.2 Stochastic Block Model

We consider a random graph model that produces graphs that have a clustering structure. Let n be an even positive integer. Given two sets of $m = \frac{n}{2}$ nodes consider the following random graph G: For each pair (i, j) of nodes, (i, j) is an edge of G with probability p if i and j are in the same set, and with probability q if they are in different sets. Each edge is drawn independently and $p > q$. This is known as the Stochastic Block Model on two communities.

(Think of nodes as habitants of two different towns and edges representing friendships, in this model, people leaving in the same town are more likely to be friends)

The goal will be to recover the original partition. This problem is clearly easy if $p = 1$ and $q = 0$ and hopeless if $p = q$. The question we will try to answer is for which values of p and q is it possible to recover the partition (perhaps with high probability). As $p > q$, we will try to recover the original partition by attempting to find the minimum bisection of the graph.

9.3 What does the spike model suggest?

Motivated by what we saw in previous lectures, one approach could be to use a form of spectral clustering to attempt to partition the graph.

Let A be the adjacency matrix of G, meaning that

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise.} \end{cases} \tag{80}$$

Note that in our model, A is a random matrix. We would like to solve

$$\begin{aligned} \max \ & \sum_{i,j} A_{ij} x_i x_j \\ \text{s.t. } & x_i = \pm 1, \forall_i \\ & \sum_j x_j = 0, \end{aligned} \tag{81}$$

The intended solution x takes the value $+1$ in one cluster and -1 in the other.

Relaxing the condition $x_i = \pm 1,\ \forall_i$ to $\|x\|_2^2 = n$ would yield a spectral method

$$\begin{aligned} \max & \sum_{i,j} A_{ij} x_i x_j \\ \text{s.t. } & \|x\|_2 = \sqrt{n} \\ & \mathbf{1}^T x = 0 \end{aligned} \tag{82}$$

The solution consists of taking the top eigenvector of the projection of A on the orthogonal of the all-ones vector $\mathbf{1}$.

The matrix A is a random matrix whose expectation is given by

$$\mathbb{E}[A] = \begin{cases} p & \text{if } (i,j) \in E(G) \\ q & \text{otherwise.} \end{cases}$$

Let g denote a vector that is $+1$ in one of the clusters and -1 in the other (note that this is the vector we are trying to find!). Then we can write

$$\mathbb{E}[A] = \frac{p+q}{2}\mathbf{1}\mathbf{1}^T + \frac{p-q}{2} g g^T,$$

and

$$A = \big(A - \mathbb{E}[A]\big) + \frac{p+q}{2}\mathbf{1}\mathbf{1}^T + \frac{p-q}{2} g g^T.$$

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

$$\mathcal{A} = \big(\mathcal{A} - \mathbb{E}[\mathcal{A}]\big) + \frac{p-q}{2} g g^T.$$

This means that $\mathcal{A}$ is a superposition of a random matrix whose expected value is zero and a rank-1 matrix, i.e.

$$\mathcal{A} = W + \lambda v v^T$$

where $W = \big(\mathcal{A} - \mathbb{E}[\mathcal{A}]\big)$ and $\lambda v v^T = \frac{p-q}{2} n \left(\frac{g}{\sqrt{n}}\right)\left(\frac{g}{\sqrt{n}}\right)^T$. In previous lectures we saw that for large enough λ, the eigenvalue associated with λ pops outside the distribution of eigenvalues of W and whenever this happens, the leading eigenvector has a non-trivial correlation with g (the eigenvector associated with λ).

Note that since to obtain $\mathcal{A}$ we simply subtracted a multiple of $\mathbf{1}\mathbf{1}^T$ from A, problem (82) is equivalent to

$$\begin{aligned} \max & \sum_{i,j} \mathcal{A}_{ij} x_i x_j \\ \text{s.t. } & \|x\|_2 = \sqrt{n} \\ & \mathbf{1}^T x = 0 \end{aligned} \tag{83}$$

Now that we removed a suitable multiple of $\mathbf{1}\mathbf{1}^T$ we will even drop the constraint $\mathbf{1}^T x = 0$, yielding

$$\begin{aligned} \max \ & \sum_{i,j} \mathcal{A}_{ij} x_i x_j \\ \text{s.t. } & \|x\|_2 = \sqrt{n}, \end{aligned} \tag{84}$$

whose solution is the top eigenvector of $\mathcal{A}$.

Recall that if $\mathcal{A} - \mathbb{E}[\mathcal{A}]$ was a Wigner matrix with i.i.d entries with zero mean and variance σ^2 then its empirical spectral density would follow the semicircle law and it will essentially be 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

$$\frac{p-q}{2} n > \frac{2\sigma\sqrt{n}}{2}. \tag{85}$$

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 $q(1-q)$.

If we were to take $\sigma^2 = \frac{p(1-p)+q(1-q)}{2}$ and use (85) it would suggest that the leading eigenvector of $\mathcal{A}$ correlates with the true partition vector g as long as

$$\frac{p-q}{2} > \frac{1}{\sqrt{n}} \sqrt{\frac{p(1-p)+q(1-q)}{2}}, \tag{86}$$

However, this argument is not necessarily valid because the matrix is not a Wigner matrix. For the special case $q = 1-p$, 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 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 (86) 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 is

$$(a-b)^2 > 2(a+b). \tag{87}$$

Remarkably this was posed as conjecture by Decelle et al. [DKMZ11] and proved in a series of works by Mossel et al. [MNS14b, MNS14a] and Massoulie [Mas14].

9.3.1 Three of more communities

The stochastic block model can be similarly defined for any $k \geq 2$ communities: G is a graph on $n = km$ nodes divided on k groups of m nodes each. Similarly to the $k = 2$ case, for each pair (i,j) of nodes, (i,j) is an edge of G with probability p if i and j are in the same set, and with probability q if they are in different sets. Each edge is drawn independently and $p > q$. In the sparse regime, $p = \frac{a}{n}$ and $q = \frac{b}{n}$, the threshold at which it is possible to make an estimate that correlates with the original partition is open.

Open Problem 9.1 *Consider the balanced Stochastic Block Model for $k > 3$ (constant) communities with inner probability $p = \frac{a}{n}$ and outer probability $q = \frac{b}{n}$, what is the threshold at which it becomes possible to make an estimate that correlates with the original partition is open (known as the partial recovery or detection threshold). We refer the reader to [DKMZ11, ZMZ14, GZC+15] for more information on this and many other interesting conjectures often motivated from statistical physics.*

9.4 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'll restrict to two communities for now. If the probability of intra-cluster edges 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 $p \ll \frac{2\log n}{n}$. For that reason we focus on the regime

$$p = \frac{\alpha \log(n)}{n} \text{ and } q = \frac{\beta \log(n)}{n}, \tag{88}$$

for some constants $\alpha > \beta$.

Let $x \in \mathbb{R}^n$ with $x_i = \pm 1$ representing the partition (note there is an ambiguity in the sense that x and $-x$ represent the same partition). Then, if we did not worry about efficiency then our guess (which corresponds to the Maximum Likelihood Estimator) would be the solution of the minimum bissection problem (81).

In fact, one can show (but this will not be the main focus of this lecture, see [ABH14] for a proof) that if

$$\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}, \tag{89}$$

then, with high probability, (81) recovers the true partition. Moreover, if

$$\sqrt{\alpha} - \sqrt{\beta} < \sqrt{2},$$

no algorithm (efficient or not) can, with high probability, recover the true partition.

We'll consider a semidefinite programming relaxation algorithm for SBM and derive conditions for exact recovery. The main ingredient for the proof will be duality theory.

9.5 The algorithm

Note that if we remove the constraint that $\sum_j x_j = 0$ in (81) then the optimal solution becomes $x = \mathbf{1}$. Let us define $B = 2A - (\mathbf{1}\mathbf{1}^T - I)$, meaning that

$$B_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } (i,j) \in E(G) \\ -1 & \text{otherwise} \end{cases} \tag{90}$$

It is clear that the problem

$$
\begin{aligned}
\max \ & \sum_{i,j} B_{ij} x_i x_j \\
\text{s.t. } & x_i = \pm 1, \forall_i \\
& \sum_j x_j = 0
\end{aligned} \tag{91}
$$

has the same solution as (81). However, when the constraint is dropped,

$$
\begin{aligned}
\max \ & \sum_{i,j} B_{ij} x_i x_j \\
\text{s.t. } & x_i = \pm 1, \forall_i,
\end{aligned} \tag{92}
$$

$x = \mathbf{1}$ is no longer an optimal solution. Intuitively, there is enough "-1" contribution to discourage unbalanced partitions. In fact, (92) is the problem we'll set ourselves to solve.

Unfortunately (92) 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 (92) 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 (92) is equivalent to

$$
\begin{aligned}
\max \quad & \mathrm{Tr}(BX) \\
\text{s.t.} \quad & X_{ii} = 1, \forall_i \\
& X = xx^T \text{ for some } x \in \mathbb{R}^n.
\end{aligned} \tag{93}
$$

The fact that $X = xx^T$ for some $x \in \mathbb{R}^n$ is equivalent to $\mathrm{rank}(X) = 1$ and $X \succeq 0$.This means that (92) is equivalent to

$$
\begin{aligned}
\max \quad & \mathrm{Tr}(BX) \\
\text{s.t.} \quad & X_{ii} = 1, \forall_i \\
& X \succeq 0 \\
& \mathrm{rank}(X) = 1.
\end{aligned} \tag{94}
$$

We now relax the problem by removing the non-convex rank constraint

$$
\begin{aligned}
\max \quad & \mathrm{Tr}(BX) \\
\text{s.t.} \quad & X_{ii} = 1, \forall_i \\
& X \succeq 0.
\end{aligned} \tag{95}
$$

This is an SDP that can be solved (up to arbitrary precision) in polynomial time [VB96].

Since we removed the rank constraint, the solution to (95) is no longer guaranteed to be rank-1. We will take a different approach from the one we used before 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 (95) not only satisfies the rank constraint but it coincides with $X = gg^T$ where g corresponds to the true partition. After X is computed, g is simply obtained as its leading eigenvector.

9.6 The analysis

Without loss of generality, we can assume that $g = (1, \ldots, 1, -1, \ldots, -1)^T$, meaning that the true partition corresponds to the first $\frac{n}{2}$ nodes on one side and the other $\frac{n}{2}$ on the other.

9.6.1 Some preliminary definitions

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 neighbours of vertex i and that $\lambda_2(M)$ is the second smallest eigenvalue of a symmetric matrix M.

Definition 9.1 *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*

$$L_{SBM} = D_{\mathcal{G}}^+ - D_{\mathcal{G}}^- - A$$

9.7 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 (95) without imposing constraints on X but rather have the constraints be implicitly enforced. Consider the following optimization problem.

$$\max_{X} \min_{\substack{Z,\, Q \\ Z \text{ is diagonal} \\ Q \succeq 0}} \operatorname{Tr}(BX) + \operatorname{Tr}(QX) + \operatorname{Tr}\left(Z\left(I_{n\times n} - X\right)\right) \tag{96}$$

Let us give it a game theoretical interpretation. Suppose that 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 satistying the constraints of (95) 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 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 satisfy the constraints of (95) then

$$\mathrm{Tr}(BX) \leq \min_{\substack{Z,\ Q \\ Z \text{ is diagonal} \\ Q \succeq 0}} \mathrm{Tr}(BX) + \mathrm{Tr}(QX) + \mathrm{Tr}\left(Z\left(I_{n\times n} - X\right)\right),$$

since equality can be achieve if, for example, the dual player picks $Q = 0_{n\times n}$, then it is clear that the values of (95) and (96) are the same:

$$\max_{\substack{X, \\ X_{ii}\ \forall_i \\ X\succeq 0}} \mathrm{Tr}(BX) = \max_{X} \min_{\substack{Z,\ Q \\ Z \text{ is diagonal} \\ Q \succeq 0}} \mathrm{Tr}(BX) + \mathrm{Tr}(QX) + \mathrm{Tr}\left(Z\left(I_{n\times n} - X\right)\right)$$

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:

$$\max_{\substack{X, \\ X_{ii}\ \forall_i \\ X\succeq 0}} \mathrm{Tr}(BX) \leq \min_{\substack{Z,\ Q \\ Z \text{ is diagonal} \\ Q \succeq 0}} \max_{X} \mathrm{Tr}(BX) + \mathrm{Tr}(QX) + \mathrm{Tr}\left(Z\left(I_{n\times n} - X\right)\right).$$

Note that we can rewrite

$$\mathrm{Tr}(BX) + \mathrm{Tr}(QX) + \mathrm{Tr}\left(Z\left(I_{n\times n} - X\right)\right) = \mathrm{Tr}\left(\left(B + Q - Z\right)X\right) + \mathrm{Tr}(Z).$$

When playing:

$$\min_{\substack{Z,\ Q \\ Z \text{ is diagonal} \\ Q \succeq 0}} \max_{X} \mathrm{Tr}\left(\left(B + Q - Z\right)X\right) + \mathrm{Tr}(Z),$$

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

$$\min_{\substack{Z,\ Q \\ Z \text{ is diagonal} \\ Q \succeq 0}} \max_{X} \mathrm{Tr}\left(\left(B + Q - Z\right)X\right) + \mathrm{Tr}(Z) = \min_{\substack{Z, \\ Z \text{ is diagonal} \\ Z - B \succeq 0}} \max_{X} \mathrm{Tr}(Z),$$

which clearly does not depend on the choices of the primal player. This means that

$$\max_{\substack{X, \\ X_{ii}\ \forall_i \\ X\succeq 0}} \mathrm{Tr}(BX) \leq \min_{\substack{Z, \\ Z \text{ is diagonal} \\ Z - B \succeq 0}} \mathrm{Tr}(Z).$$

This is known as weak duality (strong duality says that, under some conditionsm the two optimal values actually match, see, for example, [VB96], 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 [Ban16]).

Also, the problem

$$\begin{array}{rc} \min & \operatorname{Tr}(Z) \\ \text{s.t.} & Z \text{ is diagonal} \\ & Z - B \succeq 0 \end{array} \tag{97}$$

is called the dual problem of (95).

The derivation above explains why the objective value of the dual is always larger or equal to the primal. Nevertheless, there is a much simpler proof (although not as enlightening): let X, Z be respectively a feasible point of (95) and (97). Since Z is diagonal and $X_{ii} = 1$ then $\operatorname{Tr}(ZX) = \operatorname{Tr}(Z)$. Also, $Z - B \succeq 0$ and $X \succeq 0$, therefore $\operatorname{Tr}[(Z - B)X] \geq 0$. Altogether,

$$\operatorname{Tr}(Z) - \operatorname{Tr}(BX) = \operatorname{Tr}[(Z - B)X] \geq 0,$$

as stated.

Recall that we want to show that gg^T is the optimal solution of (95). Then, if we find Z diagonal, such that $Z - B \succeq 0$ and

$$\operatorname{Tr}[(Z - B)gg^T] = 0, \quad \text{(this condition is known as complementary slackness)}$$

then $X = gg^T$ must be an optimal solution of (95). 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. $\operatorname{Tr}[(Z - B)gg^T] = 0$
3. $Z - B \succeq 0$
4. $\lambda_2(Z - B) > 0$,

then gg^T is the unique optima of (95) and so recovery of the true partition is possible (with an efficient algorithm).

Z is known as the dual certificate, or dual witness.

9.8 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.

If indeed $Z - B \succeq 0$ then (2) becomes equivalent to $(Z - B)g = 0$. This means that we need to construct Z such that $Z_{ii} = \frac{1}{g_i} B[i, :]g$. 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

$$Z = 2(D_{\mathcal{G}}^{+} - D_{\mathcal{G}}^{-}) + I$$

is our guess for the 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

$$(Z - B)g = 0.$$

Therefore

Lemma 9.2 *If*

$$\lambda_2(2L_{SBM} + 11^T) > 0, \tag{98}$$

then the relaxation recovers the true partition.

Note that $2L_{SBM} + 11^T$ is a random matrix and so this boils down to "an exercise" in random matrix theory.

9.9 Matrix Concentration

Clearly,

$$\mathbb{E}\left[2L_{SBM} + 11^T\right] = 2\mathbb{E}L_{SBM} + 11^T = 2\mathbb{E}D_{\mathcal{G}}^{+} - 2\mathbb{E}D_{\mathcal{G}}^{-} - 2\mathbb{E}A + 11^T,$$

and $\mathbb{E}D_{\mathcal{G}}^{+} = \frac{n}{2}\frac{\alpha\log(n)}{n}I$, $\mathbb{E}D_{\mathcal{G}}^{-} = \frac{n}{2}\frac{\beta\log(n)}{n}I$, and $\mathbb{E}A$ is a matrix such with 4 $\frac{n}{2}\times\frac{n}{2}$ blocks where the diagonal blocks have $\frac{\alpha\log(n)}{n}$ and the off-diagonal blocks have $\frac{\beta\log(n)}{n}$. We can write this as $\mathbb{E}A = \frac{1}{2}\left(\frac{\alpha\log(n)}{n} + \frac{\beta\log(n)}{n}\right)11^T + \frac{1}{2}\left(\frac{\alpha\log(n)}{n} - \frac{\beta\log(n)}{n}\right)gg^T$

This means that

$$\mathbb{E}\left[2L_{SBM} + 11^T\right] = \left((\alpha-\beta)\log n\right)I + \left(1 - (\alpha+\beta)\frac{\log n}{n}\right)11^T - (\alpha-\beta)\frac{\log n}{n}gg^T.$$

Since $2L_{SBM}g = 0$ we can ignore what happens in the span of g and it is not hard to see that

$$\lambda_2\left[\left((\alpha-\beta)\log n\right)I + \left(1 - (\alpha+\beta)\frac{\log n}{n}\right)11^T - (\alpha-\beta)\frac{\log n}{n}gg^T\right] = (\alpha-\beta)\log n.$$

This means that it is enough to show that

$$\|L_{SBM} - \mathbb{E}\left[L_{SBM}\right]\| < \frac{\alpha-\beta}{2}\log n, \tag{99}$$

which is a large deviations inequality. ($\|\cdot\|$ denotes operator norm)

We will skip the details here (and refer the reader to [Ban15b] for the details), but the main idea is to use an inequality similar to the ones presented in the lecture about concentration of measure (and, in particular, matrix concentration). The main idea is to separate the diagonal from the non-diagonal part of $L_{SBM} - \mathbb{E}\left[L_{SBM}\right]$. The diagonal part depends on in and out-degrees of each node and can be handled with scalar concentration inequalities for trinomial distributions (as it was in [ABH14] to obtain the information theoretical bounds). The non-diagonal part has independent entries and so its spectral norm can be controlled by the following inequality:

Lemma 9.3 (Remark 3.13 in [BvH15]) *Let X be the $n \times n$ symmetric matrix with independent centered entries. Then there exists a universal constant c', such that for every $t \geq 0$*

$$\text{Prob}[\|X\| > 3\sigma + t] \leq n e^{-t^2/c'\sigma_\infty^2}, \tag{100}$$

where we have defined

$$\sigma := \max_i \sqrt{\sum_j \mathbb{E}[X_{ij}^2]}, \qquad \sigma_\infty := \max_{ij} \|X_{ij}\|_\infty .$$

Using these techniques one can show (this result was independently shown in [Ban15b] and [HWX14], with a slightly different approach)

Theorem 9.4 *Let G be a random graph with n nodes drawn accordingly 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*

$$\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}, \tag{101}$$

the semidefinite program considered above coincides with the true partition with high probability.

Note that, if

$$\sqrt{\alpha} - \sqrt{\beta} < \sqrt{2},$$

then exact recovery of the communities is impossible, meaning that the SDP algorithm is optimal. Furthermore, in this regime one can show that there will be a node on each community that is more connected to the other community that to its own, meaning that a partition that swaps them would have more likelihood. In fact, the fact that the SDP will start working essentially when this starts happening appears naturally in the analysis; the diagonal part corresponds exactly to differences between in and out-degrees and Lemma 9.3 allows one to show that the contributions of the off-diagonal part are of lower order.

Remark 9.5 *A simpler analysis (and seemingly more adaptable to other problems) can be carried out by using by Matrix Bernstein's inequality [Tro12] (described in the lecture about Matrix Concentration). The idea is simply to write $L_{SBM} - \mathbb{E}\left[L_{SBM}\right]$ as a sum of independent matrices (where each matrix corresponds to a pair of nodes) and to apply Matrix Bernstein (see [ABH14]). Unfortunately, this only shows exact recovery of a suboptimal threshold (suboptimal essentially by a factor of 2).*

9.10 More communities

A natural question is to understand what is the exact recovery threshold for the Stochastic Block Model on $k \geq 2$ communities. Recall the definition: The stochastic block model can be similarly defined for any $k \geq 2$ communities: G is a graph on $n = km$ nodes divided on k groups of m nodes each. Similarly to the $k = 2$ case, for each pair (i, j) of nodes, (i, j) is an edge of G with probability p if i and j are in the same set, and with probability q if they are in different sets. Each edge is drawn independently and $p > q$. In the logarithmic degree regime, we'll define the parameters in a slightly different way: $p = \frac{\alpha' \log m}{m}$ and $q = \frac{\beta' \log m}{m}$. Note that, for $k = 2$, we roughly have $\alpha = 2\alpha'$ and $\beta = 2\beta'$, which means that the exact recovery threshold, for $k = 2$, reads as: for

$$\sqrt{\alpha'} - \sqrt{\beta'} > 1$$

recovery is possible (and with the SDP algorithm), and for $\sqrt{\alpha'} - \sqrt{\beta'} < 1$ exact recovery is impossible.

Clearly, for any $k > 2$, if $\sqrt{\alpha'} - \sqrt{\beta'} < 1$ then exact recovery will also be impossible (simply imagine that n oracle tells us all of the community memberships except for those of two of the clusters, then the problem reduces to the $k = 2$ case). The remarkable fact is that, for $k = o(\log m)$ this is enough, not only for exact recovery to be possible, but also for an SDP based algorithm (very similar to the one above) to achieve exact recovery (see [AS15, ABKK15, HWX15, PW15]). However, for $k \approx \log n$, the situation is not understood.

Open Problem 9.2 *What is the threshold for exact recovery on the balanced symmetric Stochastic Block Model in $k \approx \log n$ communities and at what threshold does the SDP succeed at exactly determining the communities? (see [ABKK15]).*

9.11 Euclidean Clustering

The stochastic block model, although having fascinating phenomena, is not always an accurate model for clustering. The independence assumption assumed on the connections between pairs of vertices may sometimes be too unrealistic. Also, the minimum bisection of multisection objective may not be the most relevant in some applications.

One particularly popular form of clustering is k-means clustering. Given n points $x_1, \dots, x_n$ and pairwise distances $d(x_i, x_j)$, the k-means objective attempts to partition the points in k clusters $A_1, \dots, A_k$ (not necessarily of the same size) as to minimize the following objective[35]

$$\min \sum_{t=1}^{k} \frac{1}{|A_t|} \sum_{x_i, x_j \in A_t} d^2(x_i, x_j).$$

A similar objective is the one in k-medians clustering, where for each cluster a center is picked (the center has to be a point in the cluster) and the sum of the distances from all points in the cluster to the center point are to be minimized, in other words, the objective to be minimized is:

$$\min \sum_{t=1}^{k} \min_{c_t \in A_t} \sum_{x_i \in A_t} d(x_i, c_t).$$

In [ABC$^+$15] both an Linear Programming (LP) relaxation for k-medians and a Semidefinite Programming (SDP) relaxation for k-means are analyzed for a points in a generative model on which there are k disjoint balls in $\mathbb{R}^d$ and, for every ball, points are drawn according to a isotropic distribution on each of the balls. The goal is to establish exact recovery of these convex relaxations requiring the least distance between the balls. This model (in this context) was first proposed and analyzed for k-medians in [NW13], the conditions for k-medians were made optimal in [ABC$^+$15] and conditions for k-means were also given. More recently, the conditions on k-means were improved (made optimal for large dimensions) in [IMPV15a, IMPV15b] which also coined the term “Stochastic Ball Model”.

For P the set of points, in order to formulate the k-medians LP we use variables y_p indicating whether p is a center of its cluster or not and z_{pq} indicating whether q is assigned to p or not (see [ABC$^+$15] for details), the LP then reads:

[35] When the points are in Euclidean space there is an equivalent more common formulation in which each cluster is assign a mean and the objective function is the sum of the distances squared to the center.

$$
\begin{array}{ll}
\min & \sum_{p,q} d(p,q) z_{pq}, \\
s.t. & \sum_{p\in P} z_{pq} = 1, \quad \forall q \in P \\
& z_{pq} \leq y_p \\
& \sum_{p\in P} y_p = k \\
& z_{pq}, y_p \in [0,1], \quad \forall p, q \in P.
\end{array}
$$

the solution corresponds to an actual k-means solution if it is integral.

The semidefinite program for k-means is written in terms of a PSD matrix $X \in \mathbb{R}^{n\times n}$ (where n is the total number of points), see [ABC+15] for details. The intended solution is

$$
X = \frac{1}{n} \sum_{t=1}^{k} \mathbf{1}_{A_t} \mathbf{1}_{A_t}^T,
$$

where $\mathbf{1}_{A_t}$ is the indicator vector of the cluster A_t. The SDP reads as follows:

$$
\begin{array}{ll}
\min_X & \sum_{i,j} d(i,j) X_{ij}, \\
s.t. & \mathrm{Tr}(X) = k, \\
& X\mathbf{1} = \mathbf{1} \\
& X \geq 0 \\
& X \succeq 0.
\end{array}
$$

Inspired by simulations in the context of [NW13] and [ABC+15], Rachel Ward observed that the k-medians LP tends to be integral even for point configurations where no planted partition existed, and proposed the conjecture that k-medians is tight for typical point configurations. This was recorded as Problem 6 in [Mix15]. We formulate it as an open problem here:

Open Problem 9.3 *Is the LP relaxation for k-medians tight for a natural (random) generative model of points even without a clustering planted structure (such as, say, gaussian independent points)?*

Ideally, one would like to show that these relaxations (both the k-means SDP and the k-medians LP) are integral in instances that have clustering structure and not necessarily arising from generative random models. It is unclear however how to define what is meant by "clustering structure". A particularly interesting approach is through stability conditions (see, for example [AJP13]), the idea is that if a certain set of data points has a much larger $k-1$-means (or medians) objective than a k-means (or medians) one, and there is not much difference between the k and the $k+1$ objectives, then this is a good suggestion that the data is well explained by k clusters.

Open Problem 9.4 *Give integrality conditions to either the k-medians LP or the k-means SDP based on stability like conditions, as described above.*

9.12 Probably Certifiably Correct algorithms

While the SDP described in this lecture for recovery in the Stochastic Block Model achieves exact recovery in the optimal regime, SDPs (while polynomial time) tend to be slow in practice. There are faster (quasi-linear) methods that are also able to achieve exact recovery at the same threshold.

However, the SDP has an added benefit of producing a posteriori certificates. Indeed, if the solution from the SDP is integral (rank 1) then one is (a posteriori) sure to have found the minimum bisection. This means that the SDP (above the threshold) will, with high probability, not only find the minimum bisection but will also produce a posteriori certificate of such,. Such an algorithms are referred to as Probably Certifiably Correct (PCC) [Ban16]. Fortunately, one can get (in this case) get the best of both worlds and get a fast PCC method for recovery in the Stochastic Block Model essentially by using a fas method to find the solution and then using the SDP to only certify, which can be done considerably faster (see [Ban16]). More recently, a PCC algorithm was also analyzed for k-means clustering (based on the SDP described above) [IMPV15b].

9.13 Another conjectured instance of tightness

The following problem is posed, by Andrea Montanari, in [Mon14], a description also appears in [Ban15a]. We briefly describe it here as well:

Given a symmetric matrix $W \in \mathbb{R}^{n\times n}$ the positive principal component analysis problem can be written as

$$\begin{array}{ll} \max & x^T W x \\ \text{s. t.} & \|x\| = 1 \\ & x \geq 0 \\ & x \in \mathbb{R}^n. \end{array} \tag{102}$$

In the flavor of the semidefinite relaxations considered in this section, (102) can be rewritten (for $X \in \mathbb{R}^{n\times n}$) as

$$\begin{array}{ll} \max & \operatorname{Tr}(WX) \\ \text{s. t.} & \operatorname{Tr}(X) = 1 \\ & X \geq 0 \\ & X \succeq 0 \\ & \operatorname{rank}(X) = 1, \end{array}$$

and further relaxed to the semidefinite program

$$\begin{array}{ll} \max & \operatorname{Tr}(WX) \\ \text{s. t.} & \operatorname{Tr}(X) = 1 \\ & X \geq 0 \\ & X \succeq 0. \end{array} \tag{103}$$

This relaxation appears to have a remarkable tendency to be tight. In fact, numerical simulations suggest that if W is taken to be a Wigner matrix (symmetric with i.i.d. standard Gaussian entries), then the solution to (103) is rank 1 with high probability, but there is no explanation of this phenomenon. If the Wigner matrix is normalized to have entries $\mathcal{N}(0, 1/n)$, it is known that the typical value of the rank constraint problem is $\sqrt{2}$ (see [MR14]).

This motivates the last open problem of this section.

Open Problem 9.5 *Let W be a gaussian Wigner matrix with entries $\mathcal{N}(0, 1/n)$. Consider the fol-*

lowing Semidefinite Program:

$$\begin{array}{ll} \max & \mathrm{Tr}(WX) \\ s.\ t. & \mathrm{Tr}(X) = 1 \\ & X \geq 0 \\ & X \succeq 0. \end{array} \tag{104}$$

Prove or disprove the following conjectures.

1. *The expected value of this program is* $\sqrt{2} + o(1)$.

2. *With high probability, the solution of this SDP is rank 1.*

Remark 9.6 *The dual of this SDP motivates a particularly interesting statement which is implied by the conjecture. By duality, the value of the SDP is the same as the value of*

$$\min_{\Lambda \geq 0} \lambda_{\max} \left(W + \Lambda \right),$$

which is thus conjectured to be $\sqrt{2} + o(1)$*, although no bound better than* 2 *(obtained by simply taking* $\Lambda = 0$*) is known.*

10 Synchronization Problems and Alignment

10.1 Synchronization-type problems

This section will focuses on synchronization-type problems.[36] These are problems where the goal is to estimate a set of parameters from data concerning relations or interactions between pairs of them. A good example to have in mind is an important problem in computer vision, known as structure from motion: the goal is to build a three-dimensional model of an object from several two-dimensional photos of it taken from unknown positions. Although one cannot directly estimate the positions, one can compare pairs of pictures and gauge information on their relative positioning. The task of estimating the camera locations from this pairwise information is a synchronization-type problem. Another example, from signal processing, is multireference alignment, which is the problem of estimating a signal from measuring multiple arbitrarily shifted copies of it that are corrupted with noise.

We will formulate each of these problems as an estimation problem on a graph $G = (V, E)$. More precisely, we will associate each data unit (say, a photo, or a shifted signal) to a graph node $i \in V$. The problem can then be formulated as estimating, for each node $i \in V$, a group element $g_i \in \mathcal{G}$, where the group $\mathcal{G}$ is a group of transformations, such as translations, rotations, or permutations. The pairwise data, which we identify with edges of the graph $(i, j) \in E$, reveals information about the ratios $g_i(g_j)^{-1}$. In its simplest form, for each edge $(i, j) \in E$ of the graph, we have a noisy estimate of $g_i(g_j)^{-1}$ and the synchronization problem consists of estimating the individual group elements $g : V \to \mathcal{G}$ that are the most consistent with the edge estimates, often corresponding to the Maximum Likelihood (ML) estimator. Naturally, the measure of "consistency" is application specific. While there is a general way of describing these problems and algorithmic approaches to them [BCS15, Ban15a], for the sake of simplicity we will illustrate the ideas through some important examples.

10.2 Angular Synchronization

The angular synchronization problem [Sin11, BSS13] consist in estimating n unknown angles $\theta_1, \dots, \theta_n$ from m noisy measurements of their offsets $\theta_i - \theta_j \mod 2\pi$. This problem easily falls under the scope of synchronization-type problem by taking a graph with a node for each θ_i, an edge associated with each measurement, and taking the group to be $\mathcal{G} \cong SO(2)$, the group of in-plane rotations. Some of its applications include time-synchronization of distributed networks [GK06], signal reconstruction from phaseless measurements [ABFM12], surface reconstruction problems in computer vision [ARC06] and optics [RW01].

Let us consider a particular instance of this problem (with a particular noise model).

Let $z_1, \dots, z_n \in \mathbb{C}$ satisfying $|z_a| = 1$ be the signal (angles) we want to estimate ($z_a = \exp(i\theta_a)$). Suppose for every pair (i, j) we make a noisy measurement of the angle offset

$$Y_{ij} = z_i \overline{z_j} + \sigma W_{ij},$$

where $W_{ij} \sim \mathcal{N}(0, 1)$. The maximum likelihood estimator for z is given by solving (see [Sin11, BBS14])

$$\max_{|x_i|^2=1} x^* Y x. \tag{105}$$

[36] And it will follow somewhat the structure in Chapter 1 of [Ban15a]

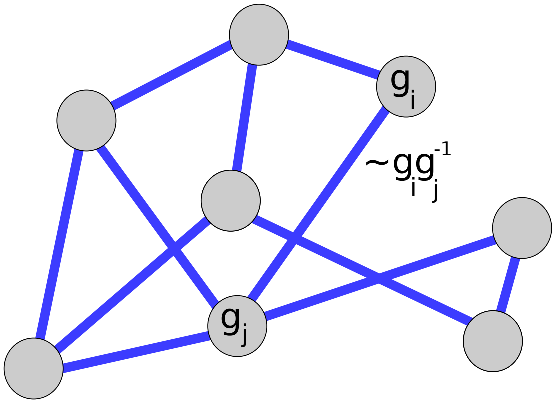


Figure 22: Given a graph $G = (V, E)$ and a group $\mathcal{G}$, the goal in synchronization-type problems is to estimate node labels $g : V \to \mathcal{G}$ from noisy edge measurements of offsets $g_i g_j^{-1}$.

There are several approaches to try to solve (105). Using techniques very similar to the study of the spike model in PCA on the first lecture one can (see [Sin11]), for example, understand the performance of the spectral relaxation of (105) into

$$\max_{\|x\|^2=n} x^* Y x. \tag{106}$$

Notice that, since the solution to (106) will not necessarily be a vector with unit-modulus entries, a rounding step will, in general, be needed. Also, to compute the leading eigenvector of A one would likely use the power method. An interesting adaptation to this approach is to round after each iteration of the power method, rather than waiting for the end of the process, more precisely:

Algorithm 10.1 *Given Y. Take a original (maybe random) vector $x^{(0)}$. For each iteration k (until convergence or a certain number of iterations) take $x^{(k+1)}$ to be the vector with entries:*

$$\left(x^{(k+1)}\right)_i = \frac{\left(Yx^{(k)}\right)_i}{\left|\left(Yx^{(k)}\right)_i\right|}.$$

Although this method appears to perform very well in numeric experiments, its analysis is still an open problem.

Open Problem 10.1 *In the model where $Y = zz^* + \sigma W$ as described above, for which values of σ will the Projected Power Method (Algorithm 10.1) converge to the optimal solution of* (105) *(or at least to a solution that correlates well with z), with high probability?*[37]

[37] We thank Nicolas Boumal for suggesting this problem.

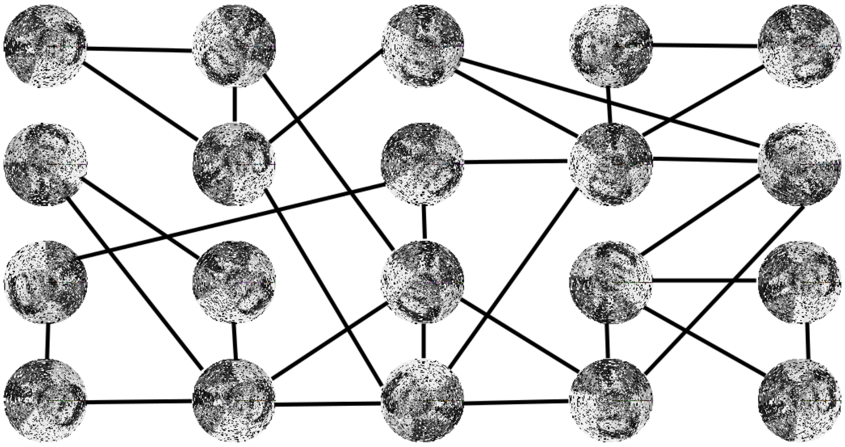


Figure 23: An example of an instance of a synchronization-type problem. Given noisy rotated copies of an image (corresponding to vertices of a graph), the goal is to recover the rotations. By comparing pairs of images (corresponding to edges of the graph), it is possible to estimate the relative rotations between them. The problem of recovering the rotation of each image from these relative rotation estimates is an instance of Angular synchronization.

We note that Algorithm 10.1 is very similar to the Approximate Message Passing method presented, and analyzed, in [MR14] for the positive eigenvector problem.

Another approach is to consider an SDP relaxation similar to the one for Max-Cut and minimum bisection.

$$\begin{aligned} \max \quad & \operatorname{Tr}(YX) \\ \text{s.t.} \quad & X_{ii} = 1, \forall_i \\ & X \succeq 0. \end{aligned} \tag{107}$$

In [BBS14] it is shown that, in the model of $Y = zz^* + \sigma W$, as long as $\sigma = \tilde{\mathcal{O}}(n^{1/4})$ then (107) is tight, meaning that the optimal solution is rank 1 and thus it corresponds to the optimal solution of (105).[38]. It is conjecture [BBS14] however that $\sigma = \tilde{\mathcal{O}}(n^{1/2})$ should suffice. It is known (see [BBS14]) that this is implied by the following conjecture:

If $x^\natural$ is the optimal solution to (105), then with high probability $\|Wx^\natural\|_\infty = \tilde{\mathcal{O}}(n^{1/2})$. This is the content of the next open problem.

Open Problem 10.2 *Prove or disprove: With high probability the SDP relaxation* (107) *is tight as long as* $\sigma = \tilde{\mathcal{O}}(n^{1/2})$. *This would follow from showing that, with high probability* $\|Wx^\natural\|_\infty = \tilde{\mathcal{O}}(n^{1/2})$, *where* $x^\natural$ *is the optimal solution to* (105).

[38] Note that this makes (in this regime) the SDP relaxation a Probably Certifiably Correct algorithm [Ban16]

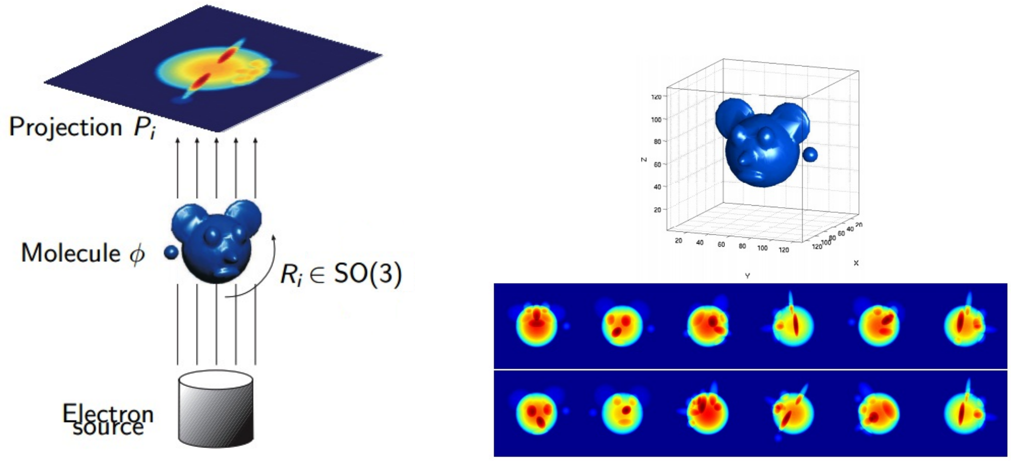


Figure 24: Illustration of the Cryo-EM imaging process: A molecule is imaged after being frozen at a random (unknown) rotation and a tomographic 2-dimensional projection is captured. Given a number of tomographic projections taken at unknown rotations, we are interested in determining such rotations with the objective of reconstructing the molecule density. Images courtesy of Amit Singer and Yoel Shkolnisky [SS11].

We note that the main difficulty seems to come from the fact that W and $x^\natural$ are not independent random variables.

10.2.1 Orientation estimation in Cryo-EM

A particularly challenging application of this framework is the orientation estimation problem in Cryo-Electron Microscopy [SS11].

Cryo-EM is a technique used to determine the three-dimensional structure of biological macromolecules. The molecules are rapidly frozen in a thin layer of ice and imaged with an electron microscope, which gives 2-dimensional projections. One of the main difficulties with this imaging process is that these molecules are imaged at different unknown orientations in the sheet of ice and each molecule can only be imaged once (due to the destructive nature of the imaging process). More precisely, each measurement consists of a tomographic projection of a rotated (by an unknown rotation) copy of the molecule. The task is then to reconstruct the molecule density from many such measurements. As the problem of recovering the molecule density knowing the rotations fits in the framework of classical tomography—for which effective methods exist— the problem of determining the unknown rotations, the orientation estimation problem, is of paramount importance. While we will not go into details here, there is a mechanism that, from two such projections, obtains information between their orientation. The problem of finding the orientation of each projection from such pairwise information naturally fits in the framework of synchronization and some of the techniques described here can be adapted to this setting [BCS15].

10.2.2 Synchronization over $\mathbb{Z}_2$

This particularly simple version already includes many applications of interest. Similarly to before, given a graph $G = (V, E)$, the goal is recover unknown node labels $g : V \to \mathbb{Z}_2$ (corresponding to memberships to two clusters) from pairwise information. Each pairwise measurement either suggests the two involved nodes are in the same cluster or in different ones (recall the problem of recovery in the stochastic block model). The task of clustering the graph in order to agree, as much as possible, with these measurements is tightly connected to *correlation clustering* [BBC04] and has applications to determining the orientation of a manifold [SW11].

In the case where all the measurements suggest that the involved nodes belong in different communities, then this problem essentially reduces to the `Max-Cut` problem.

10.3 Signal Alignment

In signal processing, the multireference alignment problem [BCSZ14] consists of recovering an unknown signal $u \in \mathbb{R}^L$ from n observations of the form

$$y_i = R_{l_i} u + \sigma \xi_i, \tag{108}$$

where R_{l_i} is a circulant permutation matrix that shifts u by $l_i \in \mathbb{Z}_L$ coordinates, ξ_i is a noise vector (which we will assume standard gaussian i.i.d. entries) and l_i are unknown shifts.

If the shifts were known, the estimation of the signal u would reduce to a simple denoising problem. For that reason, we will focus on estimating the shifts $\{l_i\}_{i=1}^n$. By comparing two observations y_i and y_j we can obtain information about the relative shift $l_i - l_j \mod L$ and write this problem as a Synchronization problem

10.3.1 The model bias pitfall

In some of the problems described above, such as the multireference alignment of signals (or the orientation estimation problem in Cryo-EM), the alignment step is only a subprocedure of the estimation of the underlying signal (or the 3d density of the molecule). In fact, if the underlying signal was known, finding the shifts would be nearly trivial: for the case of the signals, one could simply use match-filtering to find the most likely shift l_i for measurement y_i (by comparing all possible shifts of it to the known underlying signal).

When the true signal is not known, a common approach is to choose a reference signal z that is not the true template but believed to share some properties with it. Unfortunately, this creates a high risk of model bias: the reconstructed signal $\hat{u}$ tends to capture characteristics of the reference z that are not present on the actual original signal u (see Figure 10.3.1 for an illustration of this phenomenon). This issue is well known among the biological imaging community [SHBG09, Hen13] (see, for example, [Coh13] for a particularly recent discussion of it). As the experiment shown on Figure 10.3.1 suggests, the methods treated in this paper, based solely on pairwise information between observations, do not suffer from model bias as they do not use any information besides the data itself.

In order to recover the shifts l_i from the shifted noisy signals (108) we will consider the following estimator

$$\operatorname{argmin}_{l_1,\dots,l_n \in \mathbb{Z}_L} \sum_{i,j \in [n]} \left\| R_{-l_i} y_i - R_{-l_j} y_j \right\|^2, \tag{109}$$

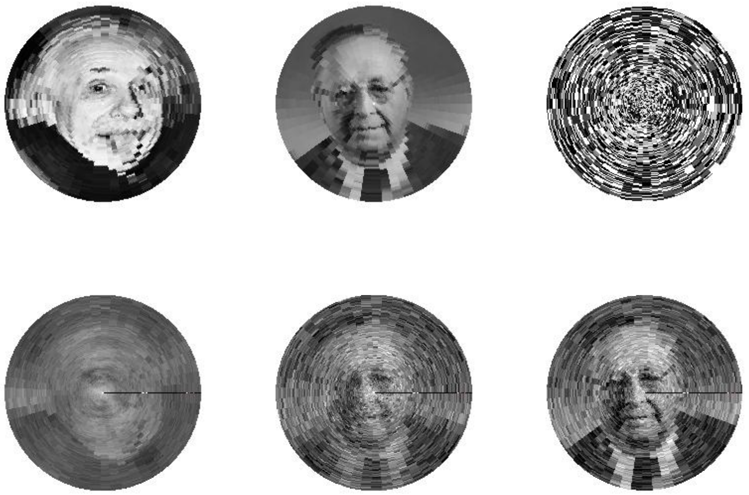


Figure 25: A simple experiment to illustrate the model bias phenomenon: Given a picture of the mathematician Hermann Weyl (second picture of the top row) we generate many images consisting of random rotations (we considered a discretization of the rotations of the plane) of the image with added gaussian noise. An example of one such measurements is the third image in the first row. We then proceeded to align these images to a reference consisting of a famous image of Albert Einstein (often used in the model bias discussions). After alignment, an estimator of the original image was constructed by averaging the aligned measurements. The result, first image on second row, clearly has more resemblance to the image of Einstein than to that of Weyl, illustration the model bias issue. One the other hand, the method based on the synchronization approach produces the second image of the second row, which shows no signs of suffering from model bias. As a benchmark, we also include the reconstruction obtained by an oracle that is given the true rotations (third image in the second row).

which is related to the maximum likelihood estimator of the shifts. While we refer to [Ban15a] for a derivation we note that it is intuitive that if l_i is the right shift for y_i and l_j for y_j then $R_{-l_i}y_i - R_{-l_j}y_j$ should be random gaussian noise, which motivates the estimator considered.

Since a shift does not change the norm of a vector, (109) is equivalent to

$$\operatorname*{argmax}_{l_1,\dots,l_n \in \mathbb{Z}_L} \sum_{i,j\in[n]} \langle R_{-l_i}y_i, R_{-l_j}y_j \rangle, \tag{110}$$

we will refer to this estimator as the quasi-MLE.

It is not surprising that solving this problem is NP-hard in general (the search space for this optimization problem has exponential size and is nonconvex). In fact, one can show [BCSZ14] that, conditioned on the Unique Games Conjecture, it is hard to approximate up to any constant.

10.3.2 The semidefinite relaxation

We will now present a semidefinite relaxation for (110) (see [BCSZ14]).

Let us identify R_l with the $L \times L$ permutation matrix that cyclicly permutes the entries fo a vector by l_i coordinates:

$$R_l \left[\begin{array}{c} u_1 \\ \vdots \\ u_L \end{array} \right] = \left[\begin{array}{c} u_{1-l} \\ \vdots \\ u_{L-l} \end{array} \right].$$

This corresponds to an L-dimensional representation of the cyclic group. Then, (110) can be rewritten:

$$\begin{aligned} \sum_{i,j\in[n]} \langle R_{-l_i} y_i, R_{-l_j} y_j \rangle &= \sum_{i,j\in[n]} \left(R_{-l_i} y_i\right)^T R_{-l_j} y_j \\ &= \sum_{i,j\in[n]} \operatorname{Tr}\left[\left(R_{-l_i} y_i\right)^T R_{-l_j} y_j\right] \\ &= \sum_{i,j\in[n]} \operatorname{Tr}\left[y_i^T R_{-l_i}^T R_{-l_j} y_j\right] \\ &= \sum_{i,j\in[n]} \operatorname{Tr}\left[\left(y_i y_j^T\right)^T R_{l_i} R_{l_j}^T\right]. \end{aligned}$$

We take

$$X = \left[\begin{array}{c} R_{l_1} \\ R_{l_2} \\ \vdots \\ R_{l_n} \end{array} \right] \left[\begin{array}{cccc} R_{l_1}^T & R_{l_2}^T & \cdots & R_{l_n}^T \end{array} \right] \in \mathbb{R}^{nL \times nL}, \tag{111}$$

and can rewrite (110) as

$$\begin{array}{ll} \max & \operatorname{Tr}(CX) \\ \text{s. t.} & X_{ii} = I_{L\times L} \\ & X_{ij} \text{ is a circulant permutation matrix} \\ & X \succeq 0 \\ & \operatorname{rank}(X) \le L, \end{array} \tag{112}$$

where C is the rank 1 matrix given by

$$C = \left[\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array} \right] \left[\begin{array}{cccc} y_1^T & y_2^T & \cdots & y_n^T \end{array} \right] \in \mathbb{R}^{nL \times nL}, \tag{113}$$

with blocks $C_{ij} = y_i y_j^T$.

The constraints $X_{ii} = I_{L\times L}$ and $\mathrm{rank}(X) \leq L$ imply that $\mathrm{rank}(X) = L$ and $X_{ij} \in O(L)$. Since the only doubly stochastic matrices in $O(L)$ are permutations, (112) can be rewritten as

$$\begin{array}{ll} \max & \mathrm{Tr}(CX) \\ \text{s. t.} & X_{ii} = I_{L\times L} \\ & X_{ij}\mathbf{1} = \mathbf{1} \\ & X_{ij} \text{ is circulant} \\ & X \geq 0 \\ & X \succeq 0 \\ & \mathrm{rank}(X) \leq L. \end{array} \tag{114}$$

Removing the nonconvex rank constraint yields a semidefinite program, corresponding to (**??**),

$$\begin{array}{ll} \max & \mathrm{Tr}(CX) \\ \text{s. t.} & X_{ii} = I_{L\times L} \\ & X_{ij}\mathbf{1} = \mathbf{1} \\ & X_{ij} \text{ is circulant} \\ & X \geq 0 \\ & X \succeq 0. \end{array} \tag{115}$$

Numerical simulations (see [BCSZ14, BKS14]) suggest that, below a certain noise level, the semidefinite program (115) is tight with high probability. However, an explanation of this phenomenon remains an open problem [BKS14].

Open Problem 10.3 *For which values of noise do we expect that, with high probability, the semidefinite program* (115) *is tight? In particular, is it true that for any σ by taking arbitrarily large n the SDP is tight with high probability?*

10.3.3 Sample complexity for multireference alignment

Another important question related to this problem is to understand its sample complexity. Since the objective is to recover the underlying signal u, a larger number of observations n should yield a better recovery (considering the model in (**??**)). Another open question is the consistency of the quasi-MLE estimator, it is known that there is some bias on the power spectrum of the recovered signal (that can be easily fixed) but the estimates for phases of the Fourier transform are conjecture to be consistent [BCSZ14].

Open Problem 10.4

1. *Is the quasi-MLE (or the MLE) consistent for the Multireference alignment problem? (after fixing the power spectrum appropriately).*
2. *For a given value of L and σ, how large does n need to be in order to allow for a reasonably accurate recovery in the multireference alignment problem?*

Remark 10.2 *One could design a simpler method based on angular synchronization: for each pair of signals take the best pairwise shift and then use angular synchronization to find the signal shifts from these pairwise measurements. While this would yield a smaller SDP, the fact that it is not* *using all of the information renders it less effective [BCS15]. This illustrates an interesting trade-off between size of the SDP and its effectiveness. There is an interpretation of this through dimensions of representations of the group in question (essentially each of these approaches corresponds to a different representation), we refer the interested reader to [BCS15] for more one that.*

References

[AABS15] E. Abbe, N. Alon, A. S. Bandeira, and C. Sandon. Linear boolean classification, coding and "the critical problem". *Available online at arXiv:1401.6528v3 [cs.IT]*, 2015.

[ABC$^+$15] P. Awasthi, A. S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, and R. Ward. Relax, no need to round: integrality of clustering formulations. *6th Innovations in Theoretical Computer Science (ITCS 2015)*, 2015.

[ABFM12] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon. Phase retrieval with polarization. *available online*, 2012.

[ABG12] L. Addario-Berry and S. Griffiths. The spectrum of random lifts. *available at arXiv:1012.4097 [math.CO]*, 2012.

[ABH14] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *Available online at arXiv:1405.3267 [cs.SI]*, 2014.

[ABKK15] N. Agarwal, A. S. Bandeira, K. Koiliaris, and A. Kolla. Multisection in the stochastic block model using semidefinite programming. *Available online at arXiv:1507.02323 [cs.DS]*, 2015.

[Abr16] L. D. Abreu. The average singular value of a complex random matrix decreases with dimension. *Available online at arXiv:1606.00494 [math.PR]*, 2016.

[ABS10] S. Arora, B. Barak, and D. Steurer. Subexponential algorithms for unique games related problems. 2010.

[AC09] Nir Ailon and Bernard Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput*, pages 302–322, 2009.

[AGZ10] 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.

[AJP13] M. Agarwal, R. Jaiswal, and A. Pal. k-means++ under approximation stability. *The 10th annual conference on Theory and Applications of Models of Computation*, 2013.

[AL06] N. Alon and E. Lubetzky. The shannon capacity of a graph and the independence numbers of its powers. *IEEE Transactions on Information Theory*, 52:21722176, 2006.

[ALMT14] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. 2014.

[Alo86] N. Alon. Eigenvalues and expanders. *Combinatorica*, 6:83–96, 1986.

[Alo03] N. Alon. Problems and results in extremal combinatorics i. *Discrete Mathematics*, 273(1–3):31–53, 2003.

[AM85] N. Alon and V. Milman. Isoperimetric inequalities for graphs, and superconcentrators. *Journal of Combinatorial Theory*, 38:73–88, 1985.

[AMMN05] N. Alon, K. Makarychev, Y. Makarychev, and A. Naor. Quadratic forms on graphs. *Invent. Math*, 163:486–493, 2005.

[AN04] N. Alon and A. Naor. Approximating the cut-norm via Grothendieck's inequality. In *Proc. of the 36 th ACM STOC*, pages 72–80. ACM Press, 2004.

[ARC06] A. Agrawal, R. Raskar, and R. Chellappa. What is the range of surface reconstructions from a gradient field? In A. Leonardis, H. Bischof, and A. Pinz, editors, *Computer Vision – ECCV 2006*, volume 3951 of *Lecture Notes in Computer Science*, pages 578–591. Springer Berlin Heidelberg, 2006.

[AS15] E. Abbe and C. Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *to appear in FOCS 2015, also available online at arXiv:1503.00609 [math.PR]*, 2015.

[B^+11] J. Bourgain et al. Explicit constructions of RIP matrices and related problems. *Duke Mathematical Journal*, 159(1), 2011.

[Bai99] Z. D. Bai. Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistics Sinica*, 9:611–677, 1999.

[Ban15a] A. S. Bandeira. *Convex relaxations for certain inverse problems on graphs.* PhD thesis, Program in Applied and Computational Mathematics, Princeton University, 2015.

[Ban15b] A. S. Bandeira. Random Laplacian matrices and convex relaxations. *Available online at arXiv:1504.03987 [math.PR]*, 2015.

[Ban15c] A. S. Bandeira. Relax and Conquer BLOG: Ten Lectures and Forty-two Open Problems in Mathematics of Data Science. 2015.

[Ban16] A. S. Bandeira. A note on probably certifiably correct algorithms. *Comptes Rendus Mathematique, to appear*, 2016.

[Bar14] 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.

[BBAP05] 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.

[BBC04] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning*, 56(1-3):89–113, 2004.

[BBRV01] S. Bandyopadhyay, P. O. Boykin, V. Roychowdhury, and F. Vatan. A new proof for the existence of mutually unbiased bases. *Available online at arXiv:quant-ph/0103162*, 2001.

[BBS14] A. S. Bandeira, N. Boumal, and A. Singer. Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Available online at arXiv:1411.3272 [math.OC]*, 2014.

[BCS15] A. S. Bandeira, Y. Chen, and A. Singer. Non-unique games over compact groups and orientation estimation in cryo-em. *Available online at arXiv:1505.03840 [cs.CV]*, 2015.

[BCSZ14] A. S. Bandeira, M. Charikar, A. Singer, and A. Zhu. Multireference alignment using semidefinite programming. *5th Innovations in Theoretical Computer Science (ITCS 2014)*, 2014.

[BDMS13] 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.

[BFMM14] 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.

[BFMW13] 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.

[BGN11] F. Benaych-Georges and R. R. Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 2011.

[BGN12] 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.

[BKS13a] A. S. Bandeira, C. Kennedy, and A. Singer. Approximating the little grothendieck problem over the orthogonal group. *Available online at arXiv:1308.5207 [cs.DS]*, 2013.

[BKS13b] B. Barak, J. Kelner, and D. Steurer. Rounding sum-of-squares relaxations. *Available online at arXiv:1312.6652 [cs.DS]*, 2013.

[BKS14] A. S. Bandeira, Y. Khoo, and A. Singer. Open problem: Tightness of maximum likelihood semidefinite relaxations. In *Proceedings of the 27th Conference on Learning Theory*, volume 35 of *JMLR W&CP*, pages 1265–1267, 2014.

[BLM15] 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.

[BMM14] A. S. Bandeira, D. G. Mixon, and J. Moreira. A conditional construction of restricted isometries. *Available online at arXiv:1410.6457 [math.FA]*, 2014.

[Bou14] J. Bourgain. An improved estimate in the restricted isometry problem. *Lect. Notes Math.*, 2116:65–70, 2014.

[BR13] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. *Conference on Learning Theory (COLT)*, 2013.

[BRM13] C. Bachoc, I. Z. Ruzsa, and M. Matolcsi. Squares and difference sets in finite fields. *Available online at arXiv:1305.0577 [math.CO]*, 2013.

[BS05] J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. 2005.

[BS14] B. Barak and D. Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. *Survey, ICM 2014*, 2014.

[BSS13] A. S. Bandeira, A. Singer, and D. A. Spielman. A Cheeger inequality for the graph connection Laplacian. *SIAM J. Matrix Anal. Appl.*, 34(4):1611–1630, 2013.

[BvH15] 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.

[Che70] 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.

[Chi15] T.-Y. Chien. *Equiangular lines, projective symmetries and nice error frames.* PhD thesis, 2015.

[Chu97] F. R. K. Chung. *Spectral Graph Theory.* AMS, 1997.

[Chu10] F. Chung. Four proofs for the cheeger inequality and graph partition algorithms. *Fourth International Congress of Chinese Mathematicians, pp. 331–349*, 2010.

[Chu13] M. Chudnovsky. The erdos-hajnal conjecture – a survey. 2013.

[CK12] P. G. Casazza and G. Kutyniok. *Finite Frames: Theory and Applications.* 2012.

[Coh13] J. Cohen. Is high-tech view of HIV too good to be true? *Science*, 341(6145):443–444, 2013.

[Coh15] G. Cohen. Two-source dispersers for polylogarithmic entropy and improved ramsey graphs. *Electronic Colloquium on Computational Complexity*, 2015.

[Con09] David Conlon. A new upper bound for diagonal ramsey numbers. *Annals of Mathematics*, 2009.

[CR09] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[CRPW12] 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.

[CRT06a] 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.

[CRT06b] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59:1207–1223, 2006.

[CT] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience.

[CT05] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51:4203–4215, 2005.

[CT06] E. J. Candès and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52:5406–5425, 2006.

[CT10] 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.

[CW04] M. Charikar and A. Wirth. Maximizing quadratic programs: Extending grothendieck's inequality. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '04, pages 54–60, Washington, DC, USA, 2004. IEEE Computer Society.

[CZ15] E. Chattopadhyay and D. Zuckerman. Explicit two-source extractors and resilient functions. *Electronic Colloquium on Computational Complexity*, 2015.

[DG02] S. Dasgupta and A. Gupta. An elementary proof of the johnson-lindenstrauss lemma. Technical report, 2002.

[DKMZ11] 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.

[DM13] Y. Deshpande and A. Montanari. Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *Available online at arXiv:1304.7047 [math.PR]*, 2013.

[DMS15] A. Dembo, A. Montanari, and S. Sen. Extremal cuts of sparse random graphs. *Available online at arXiv:1503.03923 [math.PR]*, 2015.

[Don06] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52:1289–1306, 2006.

[Dor43] R. Dorfman. The detection of defective members of large populations. 1943.

[Duc12] J. C. Duchi. Commentary on "towards a noncommutative arithmetic-geometric mean inequality" by b. recht and c. re. 2012.

[Dur06] R. Durrett. *Random Graph Dynamics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, New York, NY, USA, 2006.

[DVPS14] A. G. D'yachkov, I. V. Vorob'ev, N. A. Polyansky, and V. Y. Shchukin. Bounds on the rate of disjunctive codes. *Problems of Information Transmission*, 2014.

[EH89] P. Erdos and A. Hajnal. Ramsey-type theorems. *Discrete Applied Mathematics*, 25, 1989.

[F$^+$14] Y. Filmus et al. Real analysis in computer science: A collection of open problems. *Available online at http://simons.berkeley.edu/sites/default/files/openprobsmerged.pdf*, 2014.

[Fei05] U. Feige. On sums of independent random variables with unbounded variance, and estimating the average degree in a graph. 2005.

[FP06] 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.

[FR13] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing.* Birkhauser, 2013.

[Fuc04] J. J. Fuchs. On sparse representations in arbitrary redundant bases. *Information Theory, IEEE Transactions on*, 50(6):1341–1344, 2004.

[Fur96] Z. Furedia. On r-cover-free families. *Journal of Combinatorial Theory, Series A*, 1996.

[Gil52] E. N. Gilbert. A comparison of signalling alphabets. *Bell System Technical Journal*, 31:504–522, 1952.

[GK06] A. Giridhar and P.R. Kumar. Distributed clock synchronization over wireless networks: Algorithms and analysis. In *Decision and Control, 2006 45th IEEE Conference on*, pages 4915–4920. IEEE, 2006.

[GLV07] N. Gvozdenovic, M. Laurent, and F. Vallentin. Block-diagonal semidefinite programming hierarchies for 0/1 programming. *Available online at arXiv:0712.3079 [math.OC]*, 2007.

[Gol96] G. H. Golub. *Matrix Computations.* Johns Hopkins University Press, third edition, 1996.

[Gor85] Y. Gordon. Some inequalities for gaussian processes and applications. *Israel J. Math*, 50:109–110, 1985.

[Gor88] Y. Gordon. On milnan's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. 1988.

[GRS15] V. Guruswami, A. Rudra, and M. Sudan. *Essential Coding Theory.* Available at: http://www.cse.buffalo.edu/faculty/atri/courses/coding-theory/book/, 2015.

[GW95] 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.

[GZC$^+$15] 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.

[Haa87] U. Haagerup. A new upper bound for the complex Grothendieck constant. *Israel Journal of Mathematics*, 60(2):199–224, 1987.

[Has02] J. Hastad. Some optimal inapproximability results. 2002.

[Hen13] R. Henderson. Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proceedings of the National Academy of Sciences*, 110(45):18037–18041, 2013.

[HJ85] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[HMPW] T. Holenstein, T. Mitzenmacher, R. Panigrahy, and U. Wieder. Trace reconstruction with constant deletion probability and related results. *In Proceedings of the Nineteenth Annual ACM-SIAM*.

[HMT09] 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.

[HR] I. Haviv and O. Regev. The restricted isometry property of subsampled fourier matrices. *SODA 2016*.

[HWX14] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *Available online at arXiv:1412.6156*, 2014.

[HWX15] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *Available online at arXiv:1502.07738*, 2015.

[IKW14] A. Israel, F. Krahmer, and R. Ward. An arithmetic-geometric mean inequality for products of three matrices. *Available online at arXiv:1411.0333 [math.SP]*, 2014.

[IMPV15a] T. Iguchi, D. G. Mixon, J. Peterson, and S. Villar. On the tightness of an sdp relaxation of k-means. *Available online at arXiv:1505.04778 [cs.IT]*, 2015.

[IMPV15b] T. Iguchi, D. G. Mixon, J. Peterson, and S. Villar. Probably certifiably correct k-means clustering. *Available at arXiv*, 2015.

[JL84] 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.

[JMRT15] A. Javanmard, A. Montanari, and F. Ricci-Tersenghi. Phase transitions in semidefinite relaxations. *arXiv preprint arXiv:1511.08769*, 2015.

[Joh01] I. M. Johnston. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[Kar05] N. E. Karoui. Recent results about the largest eigenvalue of random covariance matrices and statistical application. *Acta Physica Polonica B*, 36(9), 2005.

[Kho02] S. Khot. On the power of unique 2-prover 1-round games. *Thiry-fourth annual ACM symposium on Theory of computing*, 2002.

[Kho10] S. Khot. On the unique games conjecture (invited survey). In *Proceedings of the 2010 IEEE 25th Annual Conference on Computational Complexity*, CCC '10, pages 99–121, Washington, DC, USA, 2010. IEEE Computer Society.

[KKMO05] S. Khot, G. Kindler, E. Mossel, and R. O'Donnell. Optimal inapproximability results for max-cut and other 2-variable csps? 2005.

[KV13] 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.

[KW92] J. Kuczynski and H. Wozniakowski. Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. *SIAM Journal on Matrix Analysis and Applications*, 13(4):1094–1122, 1992.

[Las01] J. B. Lassere. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[Lat05] R. Latała. Some estimates of norms of random matrices. *Proc. Amer. Math. Soc.*, 133(5):1273–1282 (electronic), 2005.

[LGT12] 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.

[Llo82] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, 1982.

[LM00] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 2000.

[Lov79] L. Lovasz. On the shannon capacity of a graph. *IEEE Trans. Inf. Theor.*, 25(1):1–7, 1979.

[LRTV12] A. Louis, P. Raghavendra, P. Tetali, and S. Vempala. Many sparse cuts via higher eigenvalues. *STOC*, 2012.

[LT16] A. E. Litvak and K. Tikhomirov. Order statistics of vectors with dependent coordinates, and the karhunen-loeve basis. *Available online at arXiv:1609.02126 [math.PR]*, 2016.

[Lyo14] R. Lyons. Factors of IID on trees. *Combin. Probab. Comput.*, 2014.

[Mas00] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2), 2000.

[Mas14] 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.

[Mek14] R. Meka. Windows on Theory BLOG: Discrepancy and Beating the Union Bound. *http://windowsontheory.org/2014/02/07/discrepancy-and-beating-the-union-bound/*, 2014.

[Mit09] M. Mitzenmacher. A survey of results for deletion channels and related synchronization channels. *Probability Surveys*, 2009.

[Mix14a] D. G. Mixon. Explicit matrices with the restricted isometry property: Breaking the square-root bottleneck. *available online at arXiv:1403.3427 [math.FA]*, 2014.

[Mix14b] D. G. Mixon. Short, Fat matrices BLOG: Gordon's escape through a mesh theorem. 2014.

[Mix14c] D. G. Mixon. Short, Fat matrices BLOG: Gordon's escape through a mesh theorem. 2014.

[Mix15] D. G. Mixon. Applied harmonic analysis and sparse approximation. *Short, Fat Matrices Web blog*, 2015.

[MM15] 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.

[MNS14a] 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.

[MNS14b] E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. *Probability Theory and Related Fields (to appear)*, 2014.

[Mon14] A. Montanari. Principal component analysis with nonnegativity constraints. *http://sublinear.info/index.php?title=Open_Problems:62*, 2014.

[Mos11] M. S. Moslehian. Ky Fan inequalities. *Available online at arXiv:1108.1467 [math.FA]*, 2011.

[MP67] 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.

[MR14] A. Montanari and E. Richard. Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *Available online at arXiv:1406.4775v1 [cs.IT]*, 2014.

[MS15] A. Montanari and S. Sen. Semidefinite programs on sparse random graphs. *Available online at arXiv:1504.05910 [cs.DM]*, 2015.

[MSS15a] A. Marcus, D. A. Spielman, and N. Srivastava. Interlacing families i: Bipartite ramanujan graphs of all degrees. *Annals of Mathematics*, 2015.

[MSS15b] A. Marcus, D. A. Spielman, and N. Srivastava. Interlacing families ii: Mixed characteristic polynomials and the kadison-singer problem. *Annals of Mathematics*, 2015.

[MZ11] S. Mallat and O. Zeitouni. A conjecture concerning optimality of the karhunen-loeve basis in nonlinear reconstruction. *Available online at arXiv:1109.0489 [math.PR]*, 2011.

[Nel] J. Nelson. Johnson-lindenstrauss notes. `http://web.mit.edu/minilek/www/jl_notes.pdf`.

[Nes00] Y. Nesterov. Squared functional systems and optimization problems. *High performance optimization*, 13(405-440), 2000.

[Nik13] A. Nikolov. The komlos conjecture holds for vector colorings. *Available online at arXiv:1301.4039 [math.CO]*, 2013.

[NN] J. Nelson and L. Nguyen. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. *Available at arXiv:1211.1002 [cs.DS].*

[NPW14] J. Nelson, E. Price, and M. Wootters. New constructions of RIP matrices with fast multiplication and fewer rows. *SODA*, pages 1515–1528, 2014.

[NSZ09] B. Nadler, N. Srebro, and X. Zhou. Semi-supervised learning with the graph laplacian: The limit of infinite unlabelled data. 2009.

[NW13] A. Nellore and R. Ward. Recovery guarantees for exemplar-based clustering. *Available online at arXiv:1309.3256v2 [stat.ML]*, 2013.

[Oli10] R. I. Oliveira. The spectrum of random k-lifts of large graphs (with possibly large k). *Journal of Combinatorics*, 2010.

[Pan13] D. Panchenko. *The Sherrington-Kirkpatrick Model*. Springer Monographs in Mathematics, 2013.

[Par00] P. A. Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, 2000.

[Pau] D. Paul. Asymptotics of the leading sample eigenvalues for a spiked covariance model. *Available online at* `http://anson.ucdavis.edu/~debashis/techrep/eigenlimit.pdf`.

[Pau07] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistics Sinica*, 17:1617–1642, 2007.

[Pea01] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6*, 2(11):559–572, 1901.

[Pis03] G. Pisier. *Introduction to operator space theory*, volume 294 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 2003.

[Pis11] G. Pisier. Grothendieck's theorem, past and present. *Bull. Amer. Math. Soc.*, 49:237–323, 2011.

[PW15] W. Perry and A. S. Wein. A semidefinite program for unbalanced multisection in the stochastic block model. *Available online at arXiv:1507.05605 [cs.DS]*, 2015.

[PWBM16] 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.

[QSW14] Q. Qu, J. Sun, and J. Wright. Finding a sparse vector in a subspace: Linear sparsity using alternating directions. *Available online at arXiv:1412.4659v1 [cs.IT]*, 2014.

[Rag08] 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.

[Ram28] F. P. Ramsey. On a problem of formal logic. 1928.

[Rec11] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[RR12] B. Recht and C. Re. Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. *Conference on Learning Theory (COLT)*, 2012.

[RS60] I. S. Reed and G. Solomon. Polynomial codes over certain finite fields. *Journal of the Society for Industrial and Applied Mathematics (SIAM)*, 8(2):300–304, 1960.

[RS10] P. Raghavendra and D. Steurer. Graph expansion and the unique games conjecture. *STOC*, 2010.

[RS13] S. Riemer and C. Schütt. On the expectation of the norm of random matrices with non-identically distributed entries. *Electron. J. Probab.*, 18, 2013.

[RST09] V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *Available at arXiv:0809.2274 [stat.CO]*, 2009.

[RST12] P. Raghavendra, D. Steurer, and M. Tulsiani. Reductions between expansion problems. *IEEE CCC*, 2012.

[RV08] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61:1025–1045, 2008.

[RW01] J. Rubinstein and G. Wolansky. Reconstruction of optical surfaces from ray data. *Optical Review*, 8(4):281–283, 2001.

[Sam66] S. M. Samuels. On a chebyshev-type inequality for sums of independent random variables. *Ann. Math. Statist.*, 1966.

[Sam68] S. M. Samuels. More on a chebyshev-type inequality. 1968.

[Sam69] S. M. Samuels. The markov inequality for sums of independent random variables. *Ann. Math. Statist.*, 1969.

[Sch12] K. Schmudgen. Around hilbert's 17th problem. *Documenta Mathematica - Extra Volume ISMP*, pages 433–438, 2012.

[Seg00] Y. Seginer. The expected norm of random matrices. *Combin. Probab. Comput.*, 9(2):149–166, 2000.

[SG10] A. J. Scott and M. Grassl. Sic-povms: A new computer study. *J. Math. Phys.*, 2010.

[Sha56] C. E. Shannon. The zero-error capacity of a noisy channel. *IRE Transactions on Information Theory*, 2, 1956.

[SHBG09] M. Shatsky, R. J. Hall, S. E. Brenner, and R. M. Glaeser. A method for the alignment of heterogeneous macromolecules from electron microscopy. *Journal of Structural Biology*, 166(1), 2009.

[Sho87] N. Shor. An approach to obtaining global extremums in polynomial mathematical programming problems. *Cybernetics and Systems Analysis*, 23(5):695–700, 1987.

[Sin11] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.*, 30(1):20 – 36, 2011.

[Spe75] J. Spencer. Ramsey's theorem – a new lower bound. *J. Combin. Theory Ser. A*, 1975.

[Spe85] J. Spencer. Six standard deviations suffice. *Trans. Amer. Math. Soc.*, (289), 1985.

[Spe94] J. Spencer. *Ten Lectures on the Probabilistic Method: Second Edition.* SIAM, 1994.

[SS11] A. Singer and Y. Shkolnisky. Three-dimensional structure determination from common lines in Cryo-EM by eigenvectors and semidefinite programming. *SIAM J. Imaging Sciences*, 4(2):543–572, 2011.

[Ste74] G. Stengle. A nullstellensatz and a positivstellensatz in semialgebraic geometry. *Math. Ann. 207*, 207:87–97, 1974.

[SW11] A. Singer and H.-T. Wu. Orientability and diffusion maps. *Appl. Comput. Harmon. Anal.*, 31(1):44–58, 2011.

[SWW12] D. A Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries. *COLT*, 2012.

[Tal95] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Etudes Sci. Publ. Math.*, (81):73–205, 1995.

[Tao07] T. Tao. What's new blog: Open question: deterministic UUP matrices. 2007.

[Tao12] T. Tao. *Topics in Random Matrix Theory.* Graduate studies in mathematics. American Mathematical Soc., 2012.

[TdSL00] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[TP13] 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.

[Tre11] L. Trevisan. in theory BLOG: CS369G Llecture 4: Spectral Partitionaing. 2011.

[Tro05] J. A. Tropp. Recovery of short, complex linear combinations via ℓ_1 minimization. *IEEE Transactions on Information Theory*, 4:1568–1570, 2005.

[Tro12] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[Tro15a] J. A. Tropp. The expected norm of a sum of independent random matrices: An elementary approach. *Available at arXiv:1506.04711 [math.PR]*, 2015.

[Tro15b] J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 2015.

[Tro15c] J. A. Tropp. Second-order matrix concentration inequalities. *In preparation*, 2015.

[Var57] R. R. Varshamov. Estimate of the number of signals in error correcting codes. *Dokl. Acad. Nauk SSSR*, 117:739–741, 1957.

[VB96] L. Vanderberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38:49–95, 1996.

[VB04] L. Vanderberghe and S. Boyd. *Convex Optimization.* Cambridge University Press, 2004.

[vH14] R. van Handel. Probability in high dimensions. *ORF 570 Lecture Notes, Princeton University*, 2014.

[vH15] R. van Handel. On the spectral norm of inhomogeneous random matrices. *Available online at arXiv:1502.05003 [math.PR]*, 2015.

[Yam54] K. Yamamoto. Logarithmic order of free distributive lattice. *Journal of the Mathematical Society of Japan*, 6:343–353, 1954.

[ZB09] L. Zdeborova and S. Boettcher. Conjecture on the maximum cut and bisection width in random regular graphs. *Available online at arXiv:0912.4861 [cond-mat.dis-nn]*, 2009.

[Zha14] T. Zhang. A note on the non-commutative arithmetic-geometric mean inequality. *Available online at arXiv:1411.5058 [math.SP]*, 2014.

[ZMZ14] Pan Zhang, Cristopher Moore, and Lenka Zdeborova. Phase transitions in semisupervised clustering of sparse networks. *Phys. Rev. E*, 90, 2014.