Lecture 4: Random Graphs and Random Matrices

In which we introduce the Stochastic Block Model and the planted clique problem, we build some intuition on linear algebraic approaches to these problem, and we introduce some random matrix theory.

1 The Stochastic Block Model and Planted Clique

In the first week, we studied connections between combinatorial properties of graphs and the spectrum of the Laplacian or normalized Laplacian matrix associated to the graph. Such connections hold for *every graph*, and the proofs are *constructive*, leading to spectral *algorithms* and to a *worst case analysis* of such algorithms.

This week we will study the spectrum of matrices (mostly, we will look at the adjacency matrix) associated to *random graphs*. Such results are proved via *random matrix theory* and lead to the *average-case analysis* of spectral algorithms in certain interesting generative models of graph instances.

We will focus on recovering the largest clique in graphs coming from the *planted clique* distribution and on community detection in the *stochastic block model*.

We will use $G_{n,p}$ to denote the distribution of Erdös-Renyi undirected random graphs on n vertices in which each pair of vertices $\{u, v\}$ is an edge with probability p, and the choices for different pairs are mutually independent.

The planted clique model with parameters n and k is a distribution of undirected graphs with n vertices sampled in the following way: we first pick a random graph from $G_{n,\frac{1}{2}}$, then we select a random subset k of k vertices and we add edges, if necessary, among pairs of elements of K in order to make K a clique. (In an undirected graph, a clique is a subset of vertices such that there is an edge between any pair of elements from the set.) Given a graph sampled from this distribution, the algorithmic problem of interest is to retrieve the set of K.

The *Stochastic Block Model* is a generic model for graphs having a cluster structure. The simplest model and one we will consider today is the $G_{n,p,q}$ problem. The $G_{n,p,q}$

distribution is a distribution on graphs of n vertices which is sampled in the following way: first V is partitioned into two 2 subsets of equal size: $V = V_1 \sqcup V_2$. Then for $\{u, v\}$ pair of vertices in the same subset, $\Pr((u, v) \in E) = p$ and otherwise $\Pr((u, jv \in E) = q$.

We will only consider the regime under which $p > q$. If we want to find the partition $V = V_1 \sqcup V_2$, it is intuitive to look at the problem of finding the minimum balanced cut. The cut (V_1, V_2) has expected size $qn^2/4$ and any other cut will have greater expected size. A justification of looking at the sparsest balanced cut can also be obtained in a Bayesian way. If we start from a prior that (V_1, V_2) is a random partition, then we see the graph and we update our probabilities, we see that, in the posterior for a given graph, each partition has a probability that depends on the number of edges crossing the cut (V_1, V_2) and that the most likely partition is the one crossed by the smallest number of edges.

We have already seen a worst-case analysis of how to find sparse cuts using spectral techniques, and we can see the results that we will present as an improved average-case analysis of that worst-case analysis. It will be more natural, however, to follow a different intuition, in which we think about how we would approach the problem spectrally if had access to the *expected* adjacency matrix, and then we reason about how close an empirical adjacency matrix is to the expected adjacency matrix, and how robust the spectral techniques are.

2 The Average Matrix for a Fixed Hidden Choice

For a fixed choice of partition (V_1, V_2) in the stochastic block model, the expectation of the adjacency matrix that we sample is

$$R := \left(\begin{array}{c|c} \mathbf{p} & \mathbf{q} \\ \hline \mathbf{q} & \mathbf{p} \end{array}\right) \tag{1}$$

Where $\mathbf{p}$ denotes an $\frac{n}{2} \times \frac{n}{2}$ sub-matrix all whose entries are p, and similarly $\mathbf{q}$, and we re-arranged the entries so that V_1 indexes the first $n/2$ rows and columns and V_2 indexes the remaining ones. (Actually, the diagonal in the expected adjacency matrix is zero, while the diagonal of R is p; we will gloss over this detail in this motivating discussion)

The matrix R is a rank-2 matrix with a very clean decomposition:

$$R = \left(\frac{p+q}{2}\right) J + \frac{p-q}{2} \left(\begin{array}{c|c} \mathbf{1} & -\mathbf{1} \\ \hline -\mathbf{1} & \mathbf{1} \end{array}\right) \tag{2}$$

The above decomposition shows that the largest eigenvalue of R is $\frac{p+q}{2} \cdot n$ with eigen-

vector $\mathbf{1}$, the second largest eigenvalue is $\frac{p-q}{2} \cdot n$ with eigenvector $\mathbf{1}_{V_1} - \mathbf{1}_{V_2}$, and all other eigenvalues are zero. We used the notation $\mathbf{1}$ for the vector that is 1 in all coordinate and, for a subset $S \subseteq V$ of vertices, we used the notation $\mathbf{1}_S$ for the vector that is 1 in the coordinates indexed by S and 0 in all other coordinates.

If we had direct access to the expected matrix, we could compute the eigenvector of the second eigenvalue, and then the partition defined by the vertices that are positive versus negative in the eigenvector would be identical to the hidden partition (V_1, V_2). Instead of the expected matrix, we have instead access to an empirical adjacency matrix of an empirical graph sampled from the distribution. Our analysis will show that the empirical adjacency matrix is close with high probability to the expected adjacency matrix, where "close" is defined in such a way as to approximately preserve the eigenvector of the second eigenvalue, so that the eigenvector of the second eigenvalue of the empirical adjacency matrix can be used to recover the hidden partition.

Regarding the planted clique problem, we can analogously consider, for a fixed choice of the set K, what is the expectation of the adjacency matrix of the graph. Ignoring, as before, some small error in the diagonal, the expected adjacency matrix will look like

$$Q := \left(\begin{array}{c|c} \mathbf{1} & \frac{\mathbf{1}}{\mathbf{2}} \\ \hline \frac{\mathbf{1}}{\mathbf{2}} & \frac{\mathbf{1}}{\mathbf{2}} \end{array}\right) \tag{3}$$

where we arranged rows and columns so that the first k rows and columns of the matrix are indexed by the elements of K. If we subtract $1/2$ from all entries, we have the rank-1 decomposition

$$Q - \frac{1}{2}J = \left(\begin{array}{c|c} \frac{\mathbf{1}}{\mathbf{2}} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right) = \frac{1}{2}\mathbf{1}_K \mathbf{1}_K^T$$

and $\mathbf{1}_K$ is the only non-trivial eigenvector of $Q - \frac{1}{2}J$, with eigenvalue $\frac{k}{2}$. The intuition for the spectral approach that we will study is as follows: we take the adjacency matrix of the graph that we are given, we subtract $1/2$ from each coordinate, and we find the eigenvector of the largest eigenvalue. If the empirical adjacency matrix is close enough, in the appropriate way, from the expected adjacency matrix, then the vertices corresponding to entries of large magnitude in the eigenvector will approximately be the hidden clique.

In both examples, the notion of closeness that we will be interested in is closenes in operator norm. The operator of a matrix M, denoted $||M||$ is defined as

$$||M|| := \sup_{x:||x||=1} ||Mx||$$

in the case of Hermitian matrices and, in particular, of symmetric real matrices which are the only matrices we study in this course, the operator norm of the matrix is also

equal to the largest magnitude of the eigenvalues of the matrix:

$$||M|| = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } M\}$$

Next time we will see the Davis-Kahan theorem which gives us that, under certain additional conditions, when two real symmetric matrices are close in operator norm then their largest eigenvalues and corresponding eigenvectors are approximately the same.

To apply such a result, we need to be able to prove that certain random matrices (specifically, the adjacency matrices of the graphs from our stochastic block model and planted clique distributions) are, with high probability, close, in operator norm, to their expectations.

3 Matrix Chernoff Bounds

When we deal with random matrices whose entries are chosen mutually independently, the *matrix Chernoff bounds* are useful and versatile tools to argue that the random matrices are, with high probability, close in operator norm to their expectation. We will state matrix Chernoff bounds without proofs.

In our statement we use the notation $A \preceq B$, that denotes that A comes before B in the positive semidefinite partial order defined on real symmetric matrices; this is equivalent as saying that all the eigenvalues of $B - A$ are non-negative.

Lemma 1 (Matrix Chernoff Bounds) *Let X be a random real $n \times n$ symmetric matrix that can be written as a sum*

$$X = X_1 + X_2 + \cdots + X_m$$

of mutually independent random real symmetric matrices X_i. Assume that $\mathbb{E}\, X_i = 0$ for all i. Then we have the following bounds.

- *(Matrix Hoeffding Bound) Suppose that we have matrices M_i that satisfy, with probability 1, $X_i^2 \preceq M_i^2$. Define $\sigma^2 := ||\sum_i M_i^2||$. Then for every $t > 0$ we have*

$$\mathbb{P}[||X|| > t] \leq 2n \cdot e^{-t^2/8\sigma^2}$$

- *(Matrix Bernstein Bound) Suppose that there is a positive real R such that, with probability 1, $||X_i|| \leq R$. Define $\sigma^2 := ||\sum_i \mathbb{E}\, X_i^2||$. Then for every $t > 0$ we have*

$$\mathbb{P}[||X|| > t] \leq 2n \cdot e^{-\frac{t^2/2}{\sigma^2 + Rt/3}}$$

To start getting a bit of practice with the use of these tools, consider the question of how close the adjacency matrix of a random graph from the Erdös-Renyi distribution $G_{n,\frac{1}{2}}$ is to its expectation. We can answer this question the Matrix Hoeffding Bound, although the answer will be slightly suboptimal.

If A is the adjacency matrix of a graph sampled from $G_{n,\frac{1}{2}}$, then the matrix $W := 2 \cdot (A - \mathbb{E}\, A)$ has a very clean description: the diagonal elements zero, and the off-diagonal elements are equally likely to be $+1$ or -1, and they are mutually independent, subject to the constraint that the matrix is symmetric. Thus W is an example of a *Wigner* matrix, a very well studied generative model of random symmetric matrices. We can write

$$W = \sum_{\{u,v\}} X^{u,v}$$

where the sum is over all unordered pairs u, v, and $X^{u,v}$ is either the matrix $M^{u,v}$ that is zero everywhere except $(M^{u,v})_{u,v} = (M^{u,v})_{v,u} = 1$, or $X^{u,v} = -M^{u,v}$, and either possibility has probability $1/2$. We see that, with probability 1, $(X^{u,v})^2 = (M^{u,v})^2$, so in particular $(X^{u,v})^2 \preceq (M^{u,v})^2$. Furthermore, we have

$$\sum_{\{u,v\}} (M^{u,v})^2 = (n-1) \cdot I$$

so

$$\sigma^2 = || \sum_{\{u,v\}} (M^{u,v})^2 || = n - 1$$

and, applying the Matrix Hoeffding Bound,

$$\mathbb{P}[||X|| > t] \leq 2n \cdot e^{-t^2/8(n-1)} \leq 2n \cdot e^{-t^2/8n}$$

If we take, for example $t = 4\sqrt{n \log n}$, we have

$$\mathbb{P}[||X|| > 4\sqrt{n \log n}] \leq 2n \cdot e^{-2 \log n} = \frac{2}{n}$$

which goes to zero. Unfortunately we get nothing if we choose t smaller than $\sqrt{8n \log n}$, so we are not capturing the correct result about the largest eigenvalue of Wigner matrices, which is known to be of the order of $\sqrt{n}$. We will return to this point in the next lecture.

Putting everything together, if A is the adjacency matrix of a graph sampled from $G_{n,\frac{1}{2}}$, we have

$$\mathbb{P}[||A - \mathbb{E}\, A|| > 2\sqrt{n \log n}] \leq \frac{2}{n}$$