Lecture 10: The Regularity Lemma

Instructor: Santosh Vempala *Lecture date: 10/23*

Scribed by: Mirabel Reid, Xinyuan Cao

1 Block Models

The motivation for this section is to define a random graph model that can approximate different properties of real-world graphs.

Consider the random graph model $G(n,p)$, where each edge is independently present with probability p. Denote this model $[p]$.

Suppose we want to model a different type of graph; for example, consider a graph $G = (V, E)$ where $V = J \cup C$, which connects a job j, to a candidate c if c is qualified for j. In this case, we will never connect a job to a job or a candidate to a candidate. So, to model this with a random graph, we might set $P(e_{j,c}) = p$, while $P(e_{j_1,j_2}) = P(e_{c_1,c_2}) = 0$. Denote this model with the following matrix: $\begin{bmatrix} 0 & p \\ p & 0 \end{bmatrix}$. This tells us that if one vertex is in group 0 and the other vertex is in group 1, the probability of an edge between them is p.

In general, suppose that the set of vertices are partitioned into k groups, $V = V_1 \cup V_2 \cup \ldots V_k$. Say that if $v \in V_i$ and $u \in V_j$, $P(e_{v,u}) = p_{i,j}$; that is, the edge probability is some number that depends on the groups that the endpoints belong to. In a similar way, we can denote this model with a $k \times k$ matrix:

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \cdots & & & \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{bmatrix}$$

This graph model is known as a *stochastic block model.*

2 The Regularity Lemma

The follow lemma shows that for any dense matrix of size n, there exists a block model that approximates it in a specific useful way. Crucially, the number of blocks *does not depend on n.*

Lemma 1 (Szemerédi's Regularity Lemma)**.** *Let G be a dense graph. Define, for any $X, Y \subset [n]$ $d(X,Y) = \frac{e(X,Y)}{|X||Y|} = \frac{|\{e_{u,v}: u \in X, v \in Y\}|}{|X||Y|}$. For any $\epsilon > 0$, $\exists k = k(\epsilon)$, and $\exists$ a partition $V = V_0 \cup V_1 \cup \cdots \cup V_k$ with the following sizes:*

$$|V_0| < \epsilon n, \quad |V_1| = |V_2| = \cdots = |V_k|$$

such that the following holds: For all but ϵk^2 pairs (i,j) with $1 \le i,j \le k$ ($i = j$ allowed),

$$\forall X \subset V_i, Y \subset V_j \text{ with } |X|, |Y| > \epsilon |V_i|, |d(X,Y) - d(V_i,V_j)| > \epsilon$$

The lemma has a lot of pieces, but essentially it makes the following promise: it is possible to divide V into a finite (in n) number of groups such that $\{V_i\}$ approximate the quantity $d(X,Y)$ for any subset of the groups. Preserving this quantity turns out to help for preserving many other graph properties as well - for example, it approximately preserves all cuts in the graph.

A downside of this lemma is that the constant k is a very large function of ϵ - it is proportional to a tower function of height $O(1/\epsilon)$. The following lemma weakens the result slightly, but gives a much better dependence on ϵ.

First, we define a useful metric for analyzing matrices: the *cut norm.*

Definition 2 (Cut Norm)**.** *For any* $A \in \mathbb{R}^n$*,*

$$\|A\|_\square = \max_{S,T \subset [n]} \left| \sum_{i \in S, j \in T} A_{ij} \right|$$

If $\|A - B\|_\square \le \epsilon n^2$*, we say that* B ϵ*-approximates* A *in cut norm.*

Definition 3 (Cut Matrix)**.** *A cut matrix is a matrix* $B \in \mathbb{R}^{m \times n}$ *where there exists a contiguous submatrix, with indices* $S \subset [m], T \subset [n]$*, such that* $B_{ij} = c_1, \forall i \in S, j \in T$*, and* B_{ij} *is zero everywhere else. We denote* $B = c(S,T,c_1)$*. That is,* B *is of the following form:*

$$B = c_1 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

where the 1 *is a rectangle of* 1*'s and the rest of the matrix has all zero entries.*

Lemma 4 (Weak Regularity Lemma)**.** $\forall A \in \mathbb{R}^{m \times n}, \forall \epsilon > 0$*,* $\exists$ *a set of cut matrices* $B_1, B_2, \dots, B_k$*, where* $k \le 1/\epsilon^2$ *such that:*

$$\|A - (B_1 + B_2 + \cdots + B_k)\|_\square \le \epsilon\sqrt{mn}\|A\|_F, \text{ where } \|A\|_F := \sqrt{\sum_{i,j} A_{ij}^2}$$

Proof. Suppose that $\|A\|_\square > \epsilon\sqrt{mn}\|A\|_F$. (Otherwise the proof is done.) Then $\exists S, T$ s.t.

$$|\sum_{i \in S, j \in T} A_{ij}| > \epsilon\sqrt{mn}\|A\|_F$$

Let $B_1 = c(S,T,c_1)$ be a cut matrix. Then

$$\begin{aligned} \|A - B_1\|_F^2 - \|A\|_F^2 &= \sum_{i \in S, j \in T} (A_{ij} - c_1)^2 - A_{ij}^2 \\ &= \sum_{i \in S, j \in T} (c_1^2 - 2A_{ij}c_1) \\ &= -2c_1 \sum_{i \in S, j \in T} A_{ij} + c_1^2|S| \cdot |T| \end{aligned}$$

Denote $A(S,T) := \sum_{i \in S, j \in T} A_{ij}$. Assume wlog that $A(S,T) > 0$. (Otherwise we can set c_1 to be $-c_1$ and A_{ij} to be $-A_{ij}$.) Set

$$c_1 = \frac{A(S,T)}{|S| \cdot |T|}.$$

Then we have

$$\begin{aligned} \|A - B_1\|_F^2 - \|A\|_F^2 &= -2\frac{A^2(S,T)}{|S| \cdot |T|} + \frac{A^2(S,T)}{|S| \cdot |T|} = -\frac{A^2(S,T)}{|S| \cdot |T|} \\ &\le -\frac{\epsilon^2 mn\|A\|_F^2}{|S| \cdot |T|} \le -\epsilon^2\|A\|_F^2 \end{aligned}$$

This implies that each step decrease the squared Frobenius norm of the remaining matrix by $\epsilon^2\|A\|_F^2$. So the number of steps (the number of cut matrices) is at most $1/\epsilon^2$. □

When A is the adjacency matrix of a graph, then $m = n$, $\|A\|_F^2$ is the number of edges and so the bound above says that the error in cut norm is at most ϵn^2, i.e., every cut is approximated to within ϵn^2 additive error.