Additional notes for applications of Szemerédi's Regularity Lemma

In this notes we complete the results explained in the lectures, with possibly more details related to the ε - δ notation. In particular, we introduce the notion of ε-regular partition and we state the celebrated *Szemerédi's Regularity Lemma*. We apply it in a Counting Lemma for Triangles and, as a second step, we use it for the so-called *Triangle Removal Lemma*. We then show one application studied at the lectures: Roth's Theorem.

Szemerédi's Regularity Lemma and consequences

The first main result we need is the decomposition of a (large) graph given by the Szemerédi's Regularity Lemma. Before this we need some preliminary definitions:

Definition 1 (ε-regular pair) *Let $G = (V, E)$ be a graph, and A, B subsets of V. We denote by $e(A, B)$ the number of edges between A and B. We write*

$$d(A, B) = \frac{e(A, B)}{|A||B|}.$$

Finally we say that the pair (A, B) is ε- regular *(or (X, Y) is an ε-*pair*) iff for all subsets $X \subseteq A$, $Y \subseteq B$ such that $|X| \geq \varepsilon|A|$, $|Y| \geq \varepsilon|B|$, we have that*

$$|d(X, Y) - d(A, B)| < \varepsilon.$$

This definition tells us that when dealing with ε-regular pairs we control (up to a little error term) the edge densities of big subsets (the condition cannot be assured for very small subsets of A and B, for instance single vertices). This idea is very powerful because once knowing a global density we can move to every local density associated to subsets.

This notion can be generalized in the following way:

Definition 2 (ε-regular partition) *Let $G = (V, E)$ be a graph, and let $V_0 \cup V_1 \cup \cdots \cup V_k$ a partition of V (namely, $V_i \cap V_j = \emptyset$ when $i \neq j$). We call each V_i a* cluster *of the graph. We say that this partition is ε-* regular *iff:*

1. *$|V_0| \leq \varepsilon|V|$ (V_0 it is called the* exceptional set *in the partition).*
2. *$|V_1| = \cdots = |V_k|$.*
3. *All but εk^2 of the pairs (V_i, V_j) with $i < j$ are ε-regular.*

Having an ε-regular partition in a graph would provide a lot of information about it: forgetting about the exceptional set (which is small, and roughly speaking, does not contribute to assure any structure to our big graph), we have lots of ε-regular pairs between big clusters (and this could be exploited as we will show later). Amazingly, every graph big enough admits such a decomposition:

Theorem 3 (Szemerédi's Regularity Lemma - SRL) *For every $\varepsilon > 0$ and every integer $m \geq 1$ there exists an integer $M := M(\varepsilon, m)$ such that every graph G with more than m vertices has an ε-regular partition $V_0 \cup V_1 \cup \cdots \cup V_k$ such that $m \leq k \leq M$.*

The important point of this result is that the size of the partition (namely, k) does *not* depend on the number of vertices of G. However, the dependence of M with respect to ε and m is *really* bad (M is huge with respect to m). Notice also that this result has sense only when dealing with *dense* graph (namely, graphs G where $C_1 n^2 \leq |E(G)| \leq C_2 n^2$, with $n = |V(G)|$: in the *sparse* case (namely, $|E(G)| = o(n^2)$ or equivalently, $|E(G)| < \gamma n^2$ for each choice of $\gamma > 0$) an ε-partition with a bounded number of clusters always exists (because the density between pairs is as small as we wish).

The prototype of results we use is the following: we start getting a counting lemma (or at least, an existence lemma) for the graph we are interested in (we will see now the case of triangles). And later, we show that, under given density conditions, by applying the SRL we can always apply the previous counting lemma to assure the existence of such substructure when G is big enough.

Lemma 4 (Counting Lemma for Triangles) *Let $G = (V, E)$ be a graph and let $X \cup Y \cup Z$ be a partition of V. Assume that $d(X, Y) = \alpha$, $d(X, Z) = \beta$, $d(Y, Z) = \gamma$. Let $\varepsilon > 0$ be such that $\min\{\alpha, \beta, \gamma\} \geq 2\varepsilon$. Assume also that the pairs (X, Y), (Y, Z) and (X, Z) are all ε-regular.*

Then, the number of triangles with vertices x, y, z, with $x \in X$, $y \in Y$ and $z \in Z$ is at least

$$(1 - 2\varepsilon)(\alpha - \varepsilon)(\beta - \varepsilon)(\gamma - \varepsilon)|X||Y||Z|.$$

Proof: We fix a vertex $x \in X$. Let $N(x)$ be the set of neighbours of x and $d_Y(x), d_Z(x)$ the number of neighbours of x in Y and Z, respectively.

We first start showing that we control the number of vertices in X with an *small* degree (we need to show that we do not have to much vertices of this type in order to be able later to apply the ε-regular condition). We start showing the following:

$$|\{x \in X : d_Y(x) < (\alpha - \varepsilon)|Y|\}| < \varepsilon|X|.$$

Assume the contrary: $|\{x \in X : d_Y(x) < (\alpha - \varepsilon)|Y|\}| \geq \varepsilon|X|$. For simplicity, write $\{x \in X : d_Y(x) < (\alpha - \varepsilon)|Y|\} = X'$. Then, $|X'| \geq \varepsilon|X|$. By definition of ε-pair, $|d(X', Y) - d(X, Y)| < \varepsilon$. As $d(X, Y) = \alpha$ and $d(X', Y) < \alpha - \varepsilon$ (see that each vertex x in X' contributes to at most $(\alpha - \varepsilon)|Y|$ edges) we have that

$$d(X', Y) - d(X, Y) < \alpha - \varepsilon - \alpha < -\varepsilon,$$

but this is a contradiction with the fact that $|d(X', Y) - d(X, Y)| < \varepsilon$. Of course, the same argument follows when dealing with Z instead of Y. So we have concluded that

$$|\{x \in X : d_Y(x) \geq (\alpha - \varepsilon)|Y| \,\text{and}\, d_Z(x) \geq (\alpha - \varepsilon)|Z|\}| \geq (1 - 2\varepsilon)|X|.$$

We will study now the number of triangles in which such x is involved. Then, $|N(x) \cap Y| = d_Y(x) \geq (\alpha - \varepsilon)|Y| \geq \varepsilon|Y|$ and $|N(x) \cap Z| = d_Z(x) \geq (\gamma - \varepsilon)|Z| \geq \varepsilon|Z|$ (Recall that $\min\{\alpha, \beta, \gamma\} \geq 2\varepsilon$), and by the ε-regularity condition,

$$|d(Y, Z) - d(N(x) \cap Y, N(x) \cap Z)| < \varepsilon \Rightarrow \gamma - \frac{e(N(x) \cap Y, N(x) \cap Z)}{|N(x) \cap Y||N(x) \cap Z|} < \varepsilon.$$

We conclude then that $e(N(x) \cap Y, N(x) \cap Z) > (\alpha - \varepsilon)(\gamma - \varepsilon)|Y||Z|$. Every edge in this set defines a triangle using the third vertex x, we conclude that the number of triangles is at least the number of triangles where vertex x satisfies the previous property, which is at least $(1 - 2\varepsilon)|X|e(N(x) \cap Y, N(x) \cap Z)$. □

Once having this counting result, we can study when we can assure the existence of many triangles in a big graph:

Theorem 5 (Triangle Removal Lemma) *For every $\varepsilon > 0$ there exists a $\delta := \delta(\varepsilon) > 0$ (such that $\delta \to 0$ when $\varepsilon \to 0$) such that for every graph G over n vertices and at most δn^3 triangles, it can made triangle free by removing at most εn^2 edges.*

Before going to the proof, let us see what this theorem says: pick a graph G (which is dense, otherwise we cannot have a number of triangles of order n^3). Then the theorem states that we can make this graph triangle-free (namely, without cycles of length 3) just by removing an small proportion of the total number of edges (see that the choice is all value of $\varepsilon > 0$). Of course, this result must be understood when n is large enough.

Proof: We prove the opposite implication: if we need to delete at least εn^2 edges from G in order to make the resulting graph triangle free, then we started from a graph with more than δn^3 triangles.

Pick $\varepsilon > 0$, and take $m = \lfloor \frac{4}{\varepsilon} \rfloor$. Now, consider a $\frac{\varepsilon}{4}$-regular partition of G with clusters $V_0 \cup V_1 \cup \cdots \cup V_k$. For convenience, write $c = |V_1| = \cdots = |V_k|$. Observe that $\lfloor \frac{4}{\varepsilon} \rfloor < k$, and that $kc < n$ (in the last inequality we are not summing the number of vertices in V_0, which satisfies that $|V_0| \leq \frac{\varepsilon}{4} n$).

We start removing some edges from G:

1. All edges which are incident with V_0 (internal or external edges): we have at most $|V_0|n = \frac{\varepsilon}{4} n^2$ of such edges.

2. All edges inside the clusters $V_1, \ldots, V_k$: we have at most $k\binom{c}{2} < kc^2 < \frac{n^2}{k} < \frac{\varepsilon}{4}n^2$ of such edges.

3. All edges defined by a non $\frac{\varepsilon}{4}$-regular pair (V_i, V_j): recall that we have at most $\frac{\varepsilon}{4}k^2$ of such pairs, hence the total number of edges here is less than $\frac{\varepsilon}{4}k^2c^2 < \frac{\varepsilon}{4}n^2$.

4. All edges lying between $\frac{\varepsilon}{4}$-regular pairs (V_i, V_j), with $d = d(V_i, V_j) < \frac{\varepsilon}{2}$. In this case we have at most $\binom{k}{2}$ such pairs (trivial bound), and the number of edges is then bounded by $\binom{k}{2}d(V_i, V_j)|V_i||V_j| < \frac{k^2}{2}\frac{\varepsilon}{2}c^2 < \frac{\varepsilon}{4}n^2$.

Resuming, adding the previous contributions we have deleted at most εn^2 edges in total. If at this point, the resulting graph is triangle-free, we are done. Otherwise, there are still triangles in the graph, and we need to delete more edges in order to make the graph triangle-free. Now, the edges that remain in the graph are defined by $\frac{\varepsilon}{4}$-regular pairs whose density is greater or bigger than $\frac{\varepsilon}{2}$. Pick 3 such pairs (call them V_i, V_j and V_k) and observe that the conditions needed in the Counting Lemma for Triangles are satisfied (takin $\varepsilon/4$ instead of ε):

- $d(V_i, V_j) = \alpha$, $d(V_j, V_k) = \beta\, d(V_i, V_k) = \gamma$, and $\min\{\alpha, \beta, \gamma\} \geq \frac{\varepsilon}{2}$
- Each pair is $\frac{\varepsilon}{4}$-regular (by assumption).

Then, by the Counting Lemma for Triangles, these three clusters define at least $(1 - \frac{\varepsilon}{2})\left(\frac{\varepsilon}{4}\right)^4 c^3$ triangles. This bound can be written in terms of n: as $n = |V_0| + kn$, $|V_0| \leq \frac{\varepsilon}{4}n$ and $k \leq M(m, \frac{\varepsilon}{4}) := M(\varepsilon)$ we have that

$$n = |V_0| + ck \Rightarrow c > \frac{1}{k}\left(1 - \frac{\varepsilon}{4}\right)n > \frac{1}{M(\varepsilon)}\left(1 - \frac{\varepsilon}{4}\right)n,$$

and hence, the number of triangles (defined for this triplet, and in fact, in the whole graph) is at least $(1 - \frac{\varepsilon}{2})\left(\frac{\varepsilon}{4}\right)^4 \frac{1}{M(\varepsilon)^3}\left(1 - \frac{\varepsilon}{4}\right)^3 n^3$. Now, choosing $\delta = (1 - \frac{\varepsilon}{2})\left(\frac{\varepsilon}{4}\right)^4 \frac{1}{M(\varepsilon)^3}\left(1 - \frac{\varepsilon}{4}\right)^3$ we have the result as claimed (namely, if we haven't deleted all the triangles by removing εn^2 edges, then the graph has at least δn^3 triangles). In particular, $\delta \to 0$ when $\varepsilon \to 0$. □

Something that is very interesting in the proof is that once we assure that we have a triangle, in fact we have many (at least δn^3). This phenomenon is known as *supersaturation*.

Application 1: Roth's Theorem

We start applying the Triangle Removal Lemma in a number theoretical context.

Definition 6 *A k-arithmetic progression (k-AP) in the integers is a set of the form* $\{a, a + d, \ldots, a + (k-1)d\}$.

Let us fix $k = 3$ and let us assume just *non-trivial* AP's ($d \neq 0$). A natural question is to give estimates for the function

$$r_3(n) = \max_{|A|:A\subseteq[n]}\{A \,\text{does not contain}\, 3 - \text{AP}'\text{s}\}.$$

The first result in this area is Roth's Theorem:

Theorem 7 (Roth's Theorem, weak version) $r_3(n) = o(n)$.

Proof: We will show that if for all $\varepsilon > 0$, and $S \subseteq [n]$ with $|S| > \varepsilon n$ (where n is large enough, we will precise it later), then S must contain a 3-AP. With this idea in mind, we construct a convenient graph and then we apply the Triangle Removal Lemma. For a set $S \subseteq [n]$, we define a graph $H(S) = (V, E)$ with vertex set equals to $\{(i, 1) : i \in [n]\} \cup \{(j, 2) : j \in [2n]\} \cup \{(k, 3) : k \in [3n]\}$ (hence, $|V| = 6n$) and edge set E defined by:

- $(i, 1)$ and $(j, 2)$ are adjacent iff $j - i \in S$.
- $(j, 2)$ and $(k, 3)$ are adjacent iff $k - j \in S$.
- $(i, 1)$ and $(k, 3)$ are adjacent iff $k - i \in 2 \cdot S = \{2s : s \in S\}$.

Observe now that if $(i,1),(j,2)$ and $(k,3)$ form a triangle in $H(S)$, then writing $j-i=a_1$, $k-j=a_2$ and $k-i=2a_3$ ($a_i\in S$), then $\{a_1,a_2,a_3\}$ defines a 3-AP (just check that x,y,z is a 3-AP iff $x+y=2z$, which in this case is satisfied: $a_1+a_2=(j-i)+(k-j)=k-i=2a_3$). Additionally, the triplets $(i,1)$, $(i+s,2)$, $(i+2s,3)$ with $s\in S$ are triangles associated to trivial 3-AP's s, $s+0$, $s+2\cdot 0$. We have then $|S|\cdot n$ such trivial triangles. Obviously these triangles do not share edges, so we need to delete at least $|S|n$ edges in order to make this graph triangle-free.

Assume now that $|S|>\varepsilon n$ (and also $|S|\leq n$). Then, the number of edges we need to remove for deleting all triangles is at least $\varepsilon n^2=\frac{\varepsilon}{36}(6n)^2=\frac{\varepsilon}{36}|V|^2$ (we need this in order to delete just the trivial ones). Then, by the Triangle Removal Lemma, there is a $\delta:=\delta(\frac{\varepsilon}{36})$ such that the graph $H(S)$ has at least $\delta|V|^3=\delta 6^3n^3$ triangles.

Then, the number of non-trivial triangles is at least $\delta 6^3n^3-n^2$. Hence, if choosing n such that

$$0<\delta 6^3n^3-n^2\Rightarrow n>\frac{1}{6^3\delta},$$

we assure the existence of some non-trivial triangle, and hence S must have some 3-AP.

□

The original proof of Roth was done using Fourier analytic tools, and provide the better estimate $r_3(n)=O(n(\log\log n)^{-1})$. The arguments behind his proof (decrease of energy argument + dichotomoy between structure and randomness) are very important for the development of the theory of additive combinatorics (for instance, the analytic proof of Szemerédi's Theorem's by Gowers).

We present now an upper bound for the function $r_3(n)$. It is based in the following easy observation: let $\mathbb{S}^d:=\{(x_1,\ldots,x_d)\in\mathbb{R}^d:x_1^2+\ldots x_d^2=1\}$. Then, there are not three points on $\mathbb{S}^d$ laying on the same line.

Theorem 8 (Behrend's construction, 1949) *$r_3(n)\geq n\exp(-c\sqrt{\log n})$, for a certain value $c>0$.*

Observe that $\log(\exp(-c\sqrt{n\log n}))=-c\sqrt{\log n}$, and hence $-\log\log n\geq -c\sqrt{\log n}$ for n large enough. In particular,

$$\exp(-c\sqrt{n\log n})\leq(\log n)^{-1}.$$

Proof: We consider integers m and M that we will specify later. We define $S(r)=\{\overrightarrow{x}=(x_1,\ldots,x_m)\in[M]^n:x_1^2+\cdots+x_m^2=r^2\}$. In particular, as each component of an element $\overrightarrow{x}$ of $S(r)$ belongs to $[M]$, we have that $m\leq r^2\leq mM^2$. So the total number of sets of the form $S(r)$ is at most mM^2-m (each square r^2 must belong to the interval $[m,mM^2]$, and we have at most mM^2-m elements there).

Observe also that $[M]^m\subseteq\bigcup S(r)$, because every choice of a vector $\overrightarrow{x}\in[M]^m$ defines a certain radius r, with $m\leq r^2\leq mM^2$. So, by the pigeonhole principle, there exists an r_0 such that

$$|S(r_0)|\geq\frac{|[M]^m|}{\sharp\text{ of sets }S(r)}\geq\frac{M^m}{mM^2-m}\leq\frac{M^{m-2}}{m}.$$

Observe that if we assume the contrary, then for all choice of r,

$$|S(r)|<\frac{|[M]^m|}{\sharp\text{ of sets }S(r)},$$

and consequently, $\sum|S(r)|\leq|S(r)|(mM^2-m)<mM^2$, which is a contradiction.

Now our objective is to project the vectors of $S(r_0)$ into a convenient interval. To do so, we consider the following operator: $P(\overrightarrow{x})=P(x_1,\ldots,x_m)=\frac{1}{2M+1}\sum_{i=1}^m x_i(2M+1)^i$. Then, by uniqueness of representation of an integer in given basis, all numbers in $P(S(r_0)):=\{P(\overrightarrow{x}):\overrightarrow{x}\in S(r_0)\}$ are different (in fact, the operator can be applied to every $S(r)$, and the conclusion would be the same). Additionally, we have that for different $\overrightarrow{x}$, $\overrightarrow{y}$ and $\overrightarrow{z}$ in $S(r_0)$, $P(\overrightarrow{x}),P(\overrightarrow{y})$ and $P(\overrightarrow{z})$ does *not* define a 3-AP. This is true because in the oposite case, looking digit by digit, be conclude that if $P(\overrightarrow{x})+P(\overrightarrow{y})=2P(\overrightarrow{z})$ then $\overrightarrow{x}+\overrightarrow{y}=2\overrightarrow{z}$. And this last statement is not possible because $\overrightarrow{x}$, $\overrightarrow{y}$ and $\overrightarrow{z}$ belong to the same $S(r_0)$. Finally, observe that $P(\overrightarrow{x})\leq(2M+1)^m$ for all $\overrightarrow{x}\in S(r_0)$.

After these observations, we can start looking at the parameters: we would like to put all this integers $P(\overrightarrow{x})$ with $\overrightarrow{x} \in S(r_0)$ in an interval, and we know that the largest possible value is bounded by $(2M+1)^m$. So let us choose $(2M+1)^m = n$. This means then that

$$M = \left\lceil \frac{n^{1/m} - 1}{2} \right\rceil.$$

We move now to order estimates. Hence, we are only concerned with the *orders*, and not the precise constants. The order of magnitude of M is equal to the one of $\left\lceil \frac{n^{1/m}}{2} \right\rceil$ (namely, we can forget about the 1). Now, we have that

$$|P(S(r_0))| = |S(r_0)| \geq \frac{M^{m-2}}{m} = \frac{(n^{1/m})^{m-2}}{m2^{m-2}} \geq \frac{n^{1-2/m}}{m2^m}$$

Now the next step is to optimize the choice of m. So, we want to choose m (in terms of n) which minimizes the value of $\frac{(n^{1-2/m})^{m-2}}{m2^m}$. In this case we find that

$$m := \frac{-1 + \sqrt{1 + 8\log(2)\log(n)}}{2\log(2)}$$

whose order of magnitude is

$$\frac{\sqrt{8\log(2)\log(n)}}{2\log(2)} = \frac{\sqrt{2\log(2)^2\log_2(n)}}{\log(2)} = \sqrt{2}\sqrt{\log_2(n)} = O(\sqrt{\log_2(n)}).$$

Hence, returning to the estimate for $|P(S(r_0))|$, and forgetting about the multiplicative constants, we have that:

$$|P(S(r_0))| \geq \frac{n^{1-2/m}}{m2^m} = \frac{n}{2^{\sqrt{\log_2(n)}}} \frac{n^{-2/\sqrt{\log_2(n)}}}{\sqrt{\log_2(n)}}.$$

Now observe the second fraction. Note that:

$$\log_2\left(n^{-2/\sqrt{\log_2(n)}}\right) = -\frac{2}{\sqrt{\log_2(n)}}\log_2(n) = -2\sqrt{\log_2(n)}$$

and consequently

$$n^{-2/\sqrt{\log_2(n)}} = 2^{-2\sqrt{\log_2(n)}}.$$

Finally,

$$|P(S(r_0))| \geq \frac{n^{1-2/m}}{m2^m} = \frac{n}{2^{3\sqrt{\log_2(n)}}} \frac{1}{\sqrt{\log_2(n)}} = n2^{-(3+o(1))\sqrt{\log_2(n)}},$$

and the result follows just writing all in basis e instead of basis 2.

□