Lecture 12: ARV Analysis

In which we begin the analysis of the ARV rounding algorithm

We want to prove

Lemma 1 (ARV Main Lemma) *Let d be a negative-type metric over a set V such that the points are contained in a unit ball and have constant average distance, that is,*

- *there is a vertex z such that $d(v,z) \leq 1$ for every $v \in V$*
- *$\sum_{u,v \in V} d(u,v) \geq c \cdot |V|^2$*

Then there are sets $S, T \subseteq V$ such that

- *$|S|, |T| \geq \Omega(|V|)$;*
- *for every $u \in S$ and every $v \in T$, $d(u,v) \geq 1/O(\sqrt{\log |V|})$*

where the multiplicative factors hidden in the $O(\cdot)$ and $\Omega(\cdot)$ notations depend only on c.

In this lecture, we will show how to reduce the ARV Main Lemma to a statement of the following form: if $\{\mathbf{x}_v\}_{v \in V}$ is a set of vectors such that the metric $d(\cdot,\cdot)$ in the ARV Main Lemma can be written as $d(u,v) = ||\mathbf{x}_u - \mathbf{x}_v||^2$, and $\mathbf{g}$ is a random Gaussian vectors, and if ℓ is such that with $\Omega(1)$ probability, there are $\Omega(n)$ disjoint pairs u, v such that $d(u,v) < \ell$ and $|\langle g, \mathbf{x}_u \rangle - \langle g, \mathbf{x}_v \rangle| \geq \Omega(1)$, then $\ell \geq \Omega(1/\sqrt{\log n})$. We will then prove such a statement in the next lecture.

1 Bottlenecks

Before beginning with the proof, it will be useful to see that certain variations of the ARV Main Lemma are false, and that we must use the assumptions of the lemma in a certain way in order to be able to prove it.

For example, consider the variation of the lemma in which $d(\cdot,\cdot)$ is an arbitrary semi-metric, rather than being of negative type. We have the following counterexample.

Fact 2 *For every n, there is a metric $d(\cdot,\cdot)$ over $V = \{1,\ldots,n\}$ such that*

- $d(i,j) \leq 1$ *for all* i,j
- $\sum_{i,j} d(i,j) \geq \Omega(n^2)$
- *For every subsets S,T of size $\Omega(n)$ we have*

$$\min_{i\in S, j\in T} d(i,j) \leq O\left(\frac{1}{\log n}\right)$$

We will not provide a full proof but here is a sketch: consider a family $G_n = ([n], E_n)$ of constant-degree graphs of constant edge expansion. (We will see later in the course that such a family exists.) Consider the shortest-path distance $d_{SP}(\cdot,\cdot)$ on $[n]$. We have:

- For every pair i,j, $d_{SP}(i,j) \leq O(\log n)$, because graphs of constant expansion have logarithmic diameter (another fact that we will prove later in the course)
- $\sum_{i,j} d_{SP}(i,j) \geq \Omega(n^2 \log n)$, because, if r is the degree of the graph, then every vertex has at most r^{t+1} other vertices at distance at most t from it, and so every vertex has at least $n/2$ other vertices at distance $\Omega(\log n)$ from itself.
- For every subsets S,T of size $\Omega(n)$ we have

$$\min_{i\in S, j\in T} d_{SP}(i,j) \leq O(1)$$

  Because, if the edge expansion is $\Omega(1)$ and the degree is $O(1)$, then for every set A of $\leq n/2$, there are $\Omega(|A|)$ vertices outside A with neighbors in A, and so the number of vertices at distance at most t from S is at least $\min\{n/2, |S|\cdot 2^{\Omega(t)}\}$. If $|S| \geq \Omega(n)$, then there is a $t = O(1)$ such that more than $n/2$ vertices are at distance $\leq t$ from S, and the same is true for T, meaning that S and T are at distance at most $2t = O(1)$ from each other.

If divide $d_{SP}(\cdot,\cdot)$ by the diameter of G, which is $O(\log n)$, we obtain a metric that satisfies the conditions of the Fact above.

This means that we cannot only use the property of $d(\cdot,\cdot)$ being a semi-metric, but we have to use the fact that it is of negative type, and we need to use in the proof the vectors $\mathbf{x}_v$ such that $d(u,v) = ||\mathbf{x}_v - \mathbf{x}_u||^2$.

Fact 2 is tight: using Bourgain's theorem, or an earlier technique of Leighton and Rao, if $d(\cdot,\cdot)$ is a semi-metric over $[n]$ such that $\max_{i,j} d(i,j) \leq 1$ and $\sum_{i,j} d(i,j) \geq \Omega(1)$, then we can find sets S, T of size $\Omega(n)$ such that $\min_{i\in S, j\in T} d(i,j) \geq \Omega(1/\log n)$.

Fact 3 *For every n, there are vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ such that*

- $||\mathbf{x}_i - \mathbf{x}_j||^2 \leq 1$ *for all* i, j
- $\sum_{i,j} ||\mathbf{x}_i - \mathbf{x}_j||^2 \geq \Omega(n^2)$
- *For every subsets S, T of size $\Omega(n)$ we have*

$$\min_{i\in S, j\in T} ||\mathbf{x}_i - \mathbf{x}_j||^2 \leq O\left(\frac{\log\log n}{\log n}\right)$$

Here we will not even provide a sketch, but the idea is to use an ϵ-net of the sphere of radius $1/2$ in dimension $O(\log n/\log\log n)$, with $\epsilon = o(1)$, and the isoperimeteric inequality for the sphere.

This means that we need to use the fact that our vectors satisfy the triangle inequalities $||\mathbf{x}_i - \mathbf{x}_j||^2 \leq ||\mathbf{x}_i - \mathbf{x}_k||^2 + ||\mathbf{x}_k - \mathbf{x}_j||^2$. It is also worth noting that for all vectors, including those of Fact 3, we have

$$||\mathbf{x}_i - \mathbf{x}_j||^2 \leq 2||\mathbf{x}_i - \mathbf{x}_k||^2 + 2||\mathbf{x}_k - \mathbf{x}_j||^2$$

so any argument that proves the ARV Main Lemma will need to use the triangle inequalities in a way that breaks down if we substitute them with the above "factor-of-2-triangle-inequalities".

Fact 3 is also tight, up to the factor of $\log\log n$, as we will see later in this lecture.

Finally, we note that the ARV Main Lemma is tight, which means that every step of its proof will have to involve statements that are tight up to constant factors.

Fact 4 *For every n that is a power of two, there is a negative-type metric $d(\cdot,\cdot)$ over a set V of size n such that*

- $d(i,j) \leq 1$ *for all* i, j
- $\sum_{i,j} d(i,j) \geq \Omega(n^2)$

- *For every subsets S, T of size $\Omega(n)$ we have*

$$\min_{i \in S, j \in T} d(i,j) \leq O\left(\frac{1}{\sqrt{\log n}}\right)$$

Let $n = 2^t$ and $V = \{0,1\}^t$. The Hamming distance $d_H(\cdot,\cdot)$ is a negative-type metric over $\{0,1\}^t$ (let $\mathbf{x}_v$ be v itself, and notice that $d_H(u,v) = ||u-v||^2$) , and it satisfies

- $d(i,j) \leq t$ for all i, j
- $\sum_{i,j} d(i,j) \geq \Omega(t \cdot n^2)$
- For every subsets S, T of size $\Omega(n)$ we have

$$\min_{i \in S, j \in T} d(i,j) \leq O(\sqrt{t})$$

  which follows from isoperimetric results on the hypercube that we will not prove

Fact 4 follows by dividing the above metric by t.

2 Gaussian Projections

The tool of *Gaussian projections* is widely used to analyze semidefinite programs. Given vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^m$ which are solutions to a semidefinite program of interest, we pick a random Gaussian vector $\mathbf{g} \sim \mathbb{R}^m$, and we consider the projections $Y_1, \ldots, Y_m$, where $Y_i := \langle \mathbf{x}_i, \mathbf{g} \rangle$. The vector $\mathbf{g} = (g_1, \ldots, g_m)$ is sampled so that the coordinates g_i are independent standard normal distributions.

We see that each Y_i has a Gaussian distribution with expectation 0 and variance $||\mathbf{x}_i||^2$, and each difference $Y_i - Y_j$ has a gaussian distribution with expectation 0 and variance $||\mathbf{x}_i - \mathbf{x}_j||^2 = d(i,j)$.

From standard bounds on Gaussian random variables,

$$\mathbb{P}[|Y_i - Y_j| \leq \delta\sqrt{d(i,j)}] \leq \frac{2}{\sqrt{2\pi}}\delta < \delta \tag{1}$$

$$\mathbb{P}[|Y_i - Y_j| \geq t\sqrt{d(i,j)}] \leq \frac{2}{\sqrt{2\pi}} e^{-t^2/2} < e^{-t^2/2} \tag{2}$$

And, setting $t = \sqrt{5 \log n}$ in (2), we get

$$\mathbb{P}[\ \forall i, j.\ |Y_i - Y_j|^2 \leq 5 \log n \cdot d(i,j)] \geq 1 - o(1) \tag{3}$$

Our first result is that, with $\Omega(1)$ probability, there are $\Omega(n^2)$ pairs i, j such that $|Y_i - Y_j| \geq \Omega(1)$.

Lemma 5 *There are constants c_1, σ that depend only on c such that with probability at least 90%, if we let L be the $c_1 n$ indices i with smallest Y_i, and R be the $c_1 n$ indices i with largest Y_i, we have*

$$\forall i \in L. \forall j \in R \;\; |Y_i - Y_j| \geq \sigma$$

PROOF: A standard Markov argument shows that if $d(i,j) \leq 1$ for all pairs i, j, and $\sum_{i,j} d(i,j) \geq cn^2$, then there are at least $cn^2/2$ pairs at distance at least $c/2$. We argue that, with probability at least 90%, $\Omega(n^2)$ of those pairs are such that $|Y_i - Y_j| \geq \Omega(1)$, which implies the conclusion.

Let F be the set of "far" pairs i, j such that $d(i,j) \geq c/2$.

By setting $\delta = \frac{1}{20}$ in (1), we have for each $(i,j) \in F$

$$\mathbb{P}[|Y_i - Y_j| \leq \sqrt{c}/20\sqrt{2}] < \frac{1}{20}$$

so, by linearity of expectation,

$$\mathbb{E}[\; |\{(i,j) \in F.\; |Y_i - Y_j| \leq \sqrt{c}/20\sqrt{2}] < \frac{|F|}{20}$$

and by Markov inequality

$$\mathbb{P}\left[\left| \left\{ (i,j) \in F.\; |Y_i - Y_j| \leq \frac{\sqrt{c}}{20\sqrt{2}} \right\} \right| > \frac{|F|}{2} \right] < .1$$

so, with probability $\geq 90\%$, there are at least $|F|/2 \geq cn^2/4$ pairs (i,j) such that $|Y_i - Y_j| \geq \frac{\sqrt{c}}{20\sqrt{2}}$.

If L and R are defined as above, and $\sigma = \min_{i \in L,\; j \in R} Y_j - Y_i$, then the number of pairs i, j at distance $> \sigma$ is at most

$$(1 - (1 - 2c_1)^2) \cdot n^2 \leq 4c_1 n^2$$

and the lemma follows if we set $c_1 = c/16$ and $\sigma = \sqrt{c}/20\sqrt{2}$. □

Note that, with $90\% - o(1)$ probability, we have sets L, R, both of size $\geq c_1 n$, such that

$$\forall i, j \in V. \;\; |Y_i - Y_j|^2 \leq 5 \log n \cdot d(i,j)$$
$$\forall i \in L,\; j \in R,\; Y_j - Y_i \geq \sigma$$

so that

$$\forall i \in L,\; j \in R,\; d(i,j) \geq \frac{\sigma^2}{5 \log n} \geq \frac{1}{O(\log n)}$$

Since we have not used the triangle inequality, the above bound is almost best possible, given Fact 3 .

3 The Algorithm to Refine L and R

Consider the following algorithm, given $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^m$ satisfying the assumptions of the Main Lemma, and a parameter ℓ,

- Pick a random gaussian vector $\mathbf{g} \sim \mathbb{R}^m$
- Define $Y_i := \langle \mathbf{x}_i, \mathbf{g} \rangle$ for $i = 1, \ldots, n$
- Let L be the $c_1 n$ indices i for which Y_i is smallest
- Let R be the $c_1 n$ indices i for which Y_i is largest
- while there is an $i \in L$ and $j \in L$ such that $d(i,j) < \ell$
  - remove i from L and j from R
- return L, R

Where c_1, σ are the constants (that depend only on c) of Lemma 5. We will prove

Lemma 6 *There is a constant c_2 (dependent only on c) such that, if we set $\ell \leq \frac{c_2}{\sqrt{\log n}}$, there is at least a 80% probability that the algorithm removes at most $\frac{c_1 n}{2}$ pairs (i,j) in the 'while' loop.*

Once we establish the above lemma, we have completed our proof of the ARV Main Lemma, because, with $70\% - o(1)$ probability, the output of the algorithm is a pair of sets L, R of size $\geq \frac{c_1 n}{2}$ such that for each $i \in L$ and $j \in R$ we have $d(i,j) \geq \frac{c_2}{\sqrt{\log n}}$.

We will prove the contrapositive, that is, if the algorithm has probability at least 20% of removing at least $\frac{c_1 n}{2}$ pairs (i,j) in the 'while' loop, then $\ell \geq c_2/\sqrt{\log n}$.

Call M the set of pairs (i,j) removed by the algorithm (like $Y_1, \ldots, Y_n$, L and R, M is a random variable determined by $\mathbf{g}$). If the algorithm has probability at least 20% of removing at least $\frac{c_1 n}{2}$ pairs (i,j) in the 'while' loop, then there is a probability at least 10% that the above happens, and that $\min_{i \in L, j \in R} |Y_i - Y_j| \geq \sigma$. This means that with probability at least 10% there are $\frac{c_i n}{2}$ disjoint pairs (i,j) such that $|Y_i - Y_j| \geq \sigma$ and $d(i,j) \leq \ell$.

By the above observation, the following lemma implies Lemma 6 and hence the ARV Main Lemma.

Lemma 7 *Let $d(\cdot,\cdot)$ be a negative-type metric over a set $V = \{1, \ldots, n\}$, let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^m$ be vectors such that $d(i,j) = ||\mathbf{x}_i - \mathbf{x}_j||^2$, let $\mathbf{g} \sim \mathbb{R}^m$ be a random vector with a Gaussian distribution, and let $Y_i := \langle \mathbf{g}, \mathbf{x}_i \rangle$.*

Suppose that, for constants c', σ and a parameter ℓ, we have that there is a $\geq 10\%$ probability that there are at least $c'n$ pairs (i, j) such that $d(i, j) \leq \ell$ and $|Y_i - Y_j| \geq \sigma$. Then there is a constant c_2, that depends only on c' and σ, such that

$$\ell \geq \frac{c_2}{\sqrt{\log n}}$$

We will prove Lemma 7 in the next lecture.