The Arora Rao Vazirani algorithm for Sparsest Cut

Recall: $G(V,E)$: undirected graph on $n$ vertices, $m$ edges.

For all $S \subseteq V$, $\Phi(S) = \frac{|E(S, \bar{S})|}{|S| \cdot |\bar{S}|}$

$$\Phi(G) = \min_{S \subseteq V} \Phi(S) \quad \rightarrow \text{Sparsity of the cut}$$

Quadratic program

$$\min \frac{1}{4} \sum_{i,j \in E} (x_i - x_j)^2$$

subject to

$$\frac{1}{4} \sum_{1 \leq i \leq n} (x_i - x_j)^2 = \beta.$$
For illustration today, we will assume that \( p = 2|S| (n-|S|) \cdot = 0.1 n^2 \)

This is equivalent to assuming that the optimal set is roughly balanced, i.e. \( |S| = n - |S| = \frac{1}{2} n \). This also happens to be the hardest case & the general case can be reduced to it.
Vector Program:

\[
\min \frac{1}{4} \sum_{f i j \in E} ||v_i - v_j||_2^2
\]

\[
||v_i||_2 = 1 \forall i
\]

\[
\frac{1}{4} \sum_{1 \leq i,j \leq n} ||v_i - v_j||_2^2 = \beta - 0.1 \frac{1}{n^2}
\]

\[
\forall i,j,k: \quad ||v_i - v_k||_2^2 \leq ||v_i - v_j||_2^2 + ||v_j - v_k||_2^2
\]

**Squared Triangle Inequality.**

**Caution:** Arbitrary vectors \(v_i, v_j, v_k\) satisfy the triangle inequality:

\[
||v_i - v_k||_2 \leq ||v_i - v_j||_2 + ||v_j - v_k||_2
\]
This does not imply the squared triangle inequality.

Why is this SDP feasible?

Let \( S^* \subseteq V \) be the sparsest cut such that \( 2|S^*| \cdot (n-|S^*|) = \beta \).

Let \( V_i = \begin{cases} 1 & \text{if } i \in S^* \\ 0 & \text{if } i \not\in S^* \end{cases} \)

Then:
\[
\frac{1}{4} \sum_{1 \leq i, j \leq n} \| V_i - V_j \|_2^2 = 2 \cdot |S^*| \cdot (n-|S^*|) = \beta.
\]
\[
\frac{1}{\mathcal{E}} \sum_{i \neq j} \|v_i - v_j\|_2^2 = \|E(S_+, S_-)\| = \Phi(G) \beta.
\]

\[|v_i|_2^2 = 1 \text{ if } i \checkmark\]

**What about squared triangle inequality?**

**Obs:** for any \(a, b, c \in \{\pm 1\}\),

\[(a - b)^2 \leq (a - c)^2 + (c - b)^2\]

trivial if \(a = b\).

if not one of \(a \neq c\) or \(b \neq c\). \(\Box\)

So, the vectors also satisfy the squared triangle inequality.
MAIN THEOREM

Thm: Let $G$ be a graph that admits a sparsest cut of size $2\sqrt{s} \cdot \left| (n-15s) \right| = \beta$. Let $V_1 \ldots V_n$ be unit vectors such that

$$\frac{1}{4} \sum_{i,j \in E} ||V_i - V_j||_2^2 = C \quad \text{and} \quad \{i,j\} \in E$$

satisfying the constraint system $\mathcal{Q}$.

Then, there's a randomized rounding algorithm that outputs a cut $S$ such that

$$\frac{E(S, \overline{S})}{\left| S \right| (n-|S|)} \leq O(\sqrt{\log n}) \cdot \left( \frac{C}{\beta} \right)$$
For each $i, j \leq n$, let
\[ d(c_i, j) = \frac{1}{4} ||v_i - v_j||^2 \]
Then, $d(c_i, j) \leq d(c_i, k) + d(c_k, j) + ij, k$.

Thus, $d$ is a "metric".

In fact, it's a "Squared Euclidean" metric in a special kind of metric often also called "metric of negative type".

So, one can view SDP a refinement of Leighton Rao LP that corresponds to $d$ being an arbitrary metric.
How might you want to round? Suppose you try to do “GW-style” rounding.

Let \( g \in \mathcal{N}(0,1)^n \); std. gaussian vector.

Let \( x_i = \text{sign}(\langle g, v_i \rangle). + i \)

**Lemma [Gaussian Rounding]**

\[
\Pr [ \text{sign}(\langle v_i, g \rangle) \neq \text{sign}(\langle v_j, g \rangle) ] = \Theta(\sqrt{d(c_{ij})}).
\]
Proof: familiar 2D analysis. 

$\frac{1}{4}\|V_i - V_j\|^2 = 8$ (say).

Then $\frac{1}{2} - \frac{1}{2} \langle V_i, V_j \rangle = 0$

or $\langle V_i, V_j \rangle = 1 - 2.8$.

Thus

\[ P_V [\text{sign}(\langle g, V_i \rangle) \neq \text{sign}(\langle g, V_j \rangle)] \]

\[= \frac{\arccos(1-2.8)}{\pi} \]

\[ \sim \sqrt{2.8} \frac{\pi}{\pi} \sim \Theta(\|V_i - V_j\|_2). \]

\[= \Theta(\sqrt{d(i,j)}) \]

$d(i,j)$ is a # bet'n 0 & 1

and can be, say $\frac{1}{\log n}$.
\[
\text{then } \sqrt{d(i,j)} = \frac{1}{\log^2 n} \gg \left( \frac{1}{\log n} \right)^{\frac{1}{\log \log n}}.
\]

So, GW rounding separates short edges (i.e. \(d(i,j) \sim \frac{1}{\text{poly}(\log n)}\)) with too large a probability.

**Idea 2: (Region Growing)**

**Def:** For any \(A \subseteq V\), & \(i \in V\)

\[d(i, A) = \min_{j \in A} d(i, j)\]

“distance of \(i\) from nearest point in \(A\)."
Then, triangle inequality for distance implies that
\[ d(i, A) \leq d(i, j) + d(j, A). \]

Lemma (Region Growing)

Let \( A \subseteq V \) be any set of vertices.
Choose \( \gamma \in [0,1] \) uniformly at random, and set
\[ S = \{ k \mid d(k, A) \leq \gamma \}. \]
Then, for every \( i,j \),
\[ P_\gamma [x_i \neq x_j] \leq d(i, j). \]
Proof: Suppose \( d(c_i, A) > d(c_j, A) \).
\[
d(c_i, A) - d(c_j, A) \leq d(c_{ij})
\]

Triangle Inequality

So \( P_x [x_i \neq x_j] \)
\[
= P_y [d(c_i, A) \geq y \geq d(c_j, A)]
\]
\[
= |d(c_i, A) - d(c_j, A)|
\]
\[
\leq d(c_{ij}).
\]

So \( E \sum_{f_{ij} \in E} \mathbb{1}(x_i \neq x_j) \)
\[
\leq \sum_{f_{ij} \in E} d(c_{ij})
\]
\[
\leq C \quad \rightarrow \quad \text{SDP opt!}
\]
Why aren't we done?

For all we know, $S = V!$

But we need $2|S| \sqrt{n - |S|} \geq \frac{\beta}{O(\sqrt{\log n})}$

Why would $S$ be too large (i.e. the cut be too imbalanced?) because somehow most vertices i ended up near the set $A$ (i.e. $d(S,A)$ was small) and thus were put in $S$. 
In order to avoid this, we will need that a significant fraction of vertices be "separated" or far from the set $A$ we choose in region growing.

**Key ARV Lemma**: Such a set $A$ exists.

Let's formalize this.
Theorem [ARV Structure Theorem] [Arora-Rao-Vazirani ’04 & Lee ’05]

Let \( V_1, \ldots, V_n \) be unit vectors. Let \( d(c(i,j)) = \frac{1}{4} ||V_i - V_j||^2 \) satisfy the triangle inequality. Further suppose \( \sum_{i \neq j \in [n]} d(c(i,j)) \geq 0.1 n^2 \) for \( \{i,j\} \in [n] \) “balanced case”. Then, there exist sets \( A, B \subseteq V \) such that:

1. \( |A|, |B| = \Omega(n) \).
2. \( \min_{i \in A, j \in B} d(c(i,j)) \geq \Delta \) for \( \Delta = \Theta\left(\frac{1}{\sqrt{\log n}}\right) \).

Further, can find \( A, B \) as above in polynomial time.
Observe: Immediately done if structure theorem is true!
Use A to "region grow."

Then, expected \(|E(S; S)| \leq C\).

while \(\mathbb{E}|S|(n-|S|) = \Omega\left(\frac{n^2}{\sqrt{n \log n}}\right)\).

\[
\frac{\mathbb{E}|E(S; S)|}{\mathbb{E}|S|(n-|S|)} = \frac{\beta}{\Omega(\sqrt{n \log n})}
\]

So \(\frac{\mathbb{E}|E(S; S)|}{\mathbb{E}|S|(n-|S|)} \leq O(\sqrt{n \log n}) \frac{\text{SDP-Opt}}{\text{OPT}} \leq O(\sqrt{n \log n}) \Phi(6).\)
Will now focus on proving structure theorem.

**OBSERVE**: Structure theorem has NOTHING to do with graph. It's a statement about existence of "well separated" sets in a "\(l_2\)-squared" metric space where avg. distances are large.
Let's now proceed to proving struct thm.

Let's now define a directed graph $H$ in terms of $y_i$'s.

H has vertices $i \in \{n\}$.

$E(H) = \{ (i,j) \in \{n\}^2 \mid d_{ciij} \leq \Delta \}$

**Def (Vertex Separator)**

$A \cap B \subseteq \{n\}$, $(A, B)$ is a vertex separator in $H$ if no edge of $H$ goes from $A$ to $B$

That is, $E(H) \cap A \times B = \emptyset$.

We say that a vertex separator $(A, B)$ is good if $|A| \cdot |B| = 2(n^2)$. 
Then struct theorem is same as saying that \( H \) has a good vertex separator for \( \Delta = \sqrt{\frac{\log n}{m}} \).

Randomized Algorithm to find vertex separators in \( H \).

Let \( y_i = \langle g, V_i \rangle \neq 1 \).

Then, \( E y_i^2 = 1 \), \( d(i,j) = \frac{1}{4} E (y_i - y_j)^2 \).

1. Choose subsets \( A^0, B^0 \) s.t.
   \[
   A^0 = \{ i \mid y_i \leq -1 \} \quad B^0 = \{ j \mid y_j > 1 \}.
   \]

2. Find maximal matching of \( E \cap (A^0 \times B^0) \) greedily by processing edges \( [n] \times [n] \) in alphabetical order.

3. Output \( A = A^0 \cup V(M) \), \( B = B^0 \cup V(M) \).
Note: 1) $A^0, B^0, A, B$ are all random variables as they are functions of the gaussian vector $g$.

2) edges of matching $M$ are directed. In particular,

$$(i,j) \in E(M) \implies y_j - y_i \geq 2.$$ 

Prop: $(A, B)$ form a vertex sepia.

Proof: Consider any edge in $E(H) \cap A^0 \times B^0$. If it exists in $E(H) \cap A \times B$, can remove it, add it to $M$ and thus $M$ is not maximal.
Obs: \( M(y) = \text{reverse of } M(y) \).

Processing order of edges in greedy algo is independent of \( y \).

\[ \Pr[M] = \Pr[\text{rev}(M)] \]

Lemma (Vertex Sep or Large Matching)

There's a const. \( c' \) s.t. if

\[ |E(A \cap B)| < c' n^2 \text{ then } |E(M)| > c'n. \]

Proof:
Fix any \( i, j \), then, a constant absolute

\[ \Pr[y_i = -1, y_j = 1] \geq c \cdot d(i, j). \]

(Similar proofs as GW analysis.

Exercise!)
Thus, \[ E(A^0|B^0) > 0.1 \cdot c \cdot n^2 \]
\[ |A| \cdot |B| > |A^0| \cdot |B^0| - n \cdot |\text{IM}| \]
\[ E(A|B) \geq E(A^0|B^0) - n \cdot |\text{IM}| \]
\[ \geq 0.1 \cdot c \cdot n^2 \]
\[ c \]
So if \[ E(A|B) < 0.05c \cdot n^2 \]
them \[ |\text{EM}| > 0.05c \cdot n \]

We will analyze \[ |\text{EM}| \] and prove it is small.
Lemma (Main Tech Lemma)

For some abs. const. $c > 0$,

$$\frac{c}{\Delta} \left( \frac{E|MI|}{n} \right)^3 \leq E \max_{i,j} y_j - y_i \leq \sqrt{2 \log n}.$$ 

In words, we will analyze the expectation of maximum of

$$\{ y_j - y_i \mid 1 \leq i, j \leq n \}.$$

By standard results this is at most $\sqrt{2 \log n}$.

OTOH, we’ll show that if $T$ a large matching then expected maximum of $y_j - y_i$ is large.
Rearrange:

\[ E|M| \leq n \cdot (\sqrt{2 \log n} \cdot \Delta) \]

If \( \Delta > O(\sqrt{\log n}) \), RHS < \( C'n \). So done!

Fact (Proof on Raza)

Suppose \( Z_1, \ldots, Z_t \) are jointly Gaussian with \( \mathbb{E}Z_i = 0 \) & \( \mathbb{E}Z_i^2 = 1 \) for all \( i \).

Then \( \mathbb{E} \max_i Z_i \leq 2 \sqrt{\log 2t} \)

Apply this to \( \{Y_j - Y_i\} \) to get RHS.
Max y_j - y_i : max of a bunch of jointly "gaussian process" of gaussian random variables

Proof of Main Tech Lemma

Idea of the proof:
"large matching implies that there's an j, i s.t. Y_j - Y_i" is
large.

Why? matching edge $i \rightarrow j$

$\Rightarrow Y_j - Y_i \geq 2$

Chain together edges and show that we can form long paths. Then end points of the path must be very far!

We'll run this argument "in average".
**Def**

* \( H^k(i) \) = Set of vertices reachable from \( i \) from at most \( k \)-steps in \( H \).

* \( Z_i^{(k)} = \max_{j \in H^k(i)} y_j - y_i \).

* \( \Psi(k) = \sum_{i=1}^{n} E \max_{j} Z_i^{(k)} \).

Then, notice that

\[ \Psi(k) \leq \frac{E}{n} \max_{i,j} y_j - y_i \]

**Goal**: Prove \( \Psi(k) \) is large if \( E \| M \| \) is large.
Lemma (Chaining Lemma)

For some $C \geq 0$,

$$\Psi(k+1) \geq \Psi(k) + 4 \cdot E|\tilde{M}|$$

If $H$ is large on avg.

$$\frac{y^2}{2}$$

large

$$- C \cdot n \cdot \max_{i \in \text{nn}} (E(G_i - y_i))^2$$

$$j \in H^k + \{i\}$$

Error/Noise/Variance Term

Proof of Main Tech Lemma assuming Chaining Lemma

\[ E(y_i - y_j)^2 = d(i,j) \leq k \cdot \Delta \]

Squared Triangle Inequality

So \( \Psi(k+1) \geq \Psi(k) + E|\tilde{M}| - Cn \sqrt{k \cdot \Delta} \)
So for every \( k \leq k_0 = \frac{\frac{1}{4c^2} \left( \frac{E(M)}{n} \right)^2}{\Delta} \)

\[ \Psi(k+1) \geq \Psi(k) + \left( \frac{E(M)}{2} \right). \]

Thus, \( \frac{\Psi(k_0)}{n} \geq \frac{1}{2} k_0 \left( \frac{E(M)}{n} \right) \)

\( \leq \frac{1}{8c^2} \left( \frac{E(M)}{n} \right)^{\frac{3}{2}}. \)
Fact (Borell): \( Z_1, \ldots, Z_t \) mean 0, jointly gaussian.

\[ \text{Var} \left[ \max Z_1, \ldots, Z_t \right] \]

\[ \leq \text{O}(1) \cdot \max \{ \text{Var} (Z_1), \ldots, \text{Var} (Z_t) \} \].

Borell's Inequality, can be proved via "Lipschitz Concentration" / Sudakov Tsirelson '74

---

Proof of Chaining Lemma [uses Borell]

\[ Z_i^{k+1} \geq Z_i^k + X_j - X_i - \epsilon \]

for every \((i,j) \in E(C,H)\).
This is because for any \( r \in T_f \)
there's a \( \leq k+1 \) length path from \( i \).

"Fluff edges "

let \( N \subseteq [n] \times [n] \): arbitrary matching of vertices not in \( M \). Thus,

\[
\begin{align*}
    \forall (i,j) \in M: & \quad z_i^{k+1} > z_j + 2 \\
    \forall (i,j) \in N: & \quad \frac{1}{2} z_i^{k+1} + \frac{1}{2} z_j^{k+1} > \frac{1}{2} z_i^k + \frac{1}{2} z_j^k
\end{align*}
\]

\( \text{total of } \frac{n}{2} \text{ inequalities} \).
Add the inequalities, take exp.

\[ \sum_{i=1}^{n} z_i \cdot L_i \geq \sum_{j=1}^{n} z_j \cdot R_j + 2|M| \]

where

\[ L_i = \begin{cases} 1 & \text{if } i \text{ has an outgoing edge in } M \\ \frac{1}{2} & \text{if } i \text{ is not matched in } M \\ 0 & \text{if } i \text{ has an incoming edge in } M \end{cases} \]

\[ R_i = \begin{cases} 1 & \text{if } i \text{ has an incoming edge in } M \\ \frac{1}{2} & \text{if } i \text{ is not matched in } M \\ 0 & \text{if } i \text{ has an outgoing edge in } M \end{cases} \]
Recall obs. \( Pr[i \text{ has an incoming edge in } M] = Pr[i \text{ has an outgoing edge in } M] \)

Thus, \( EL_i = \frac{1}{2} \)

\( ER_i = \frac{1}{2} \).

Error incurred if we replace \( Ez_{i+1}^k \cdot Li \) by \( Ez_{i+1}^k \cdot EL_i \)

\[
|Ez_{i+1}^k \cdot Li - Ez_{i+1}^k \cdot EL_i|
\]

\[
= |E | Ez_{i+1}^k - Ez_{i+1}^k | \cdot | Li - EL_i |
\]

\[
\leq \sqrt{E | Ez_{i+1}^k - Ez_{i+1}^k |^2}
\]

\[
\leq \sqrt{E | Ez_{i+1}^k - Ez_{i+1}^k |^2}
\]

\[
O(1) \cdot \max_j \sqrt{E (y_j - y)^2} \leq 1 \quad \text{since } Li \text{ is in } [0, 1]
Similarly,
\[ |E_z^j \cdot R_j - E_z^j \cdot R_j| \]
\[ \leq O(c) \cdot \max_{j \in H^k(i)} \sqrt{E(y_j - y_i)^2} \]

Thus,
\[ \sum_{i=1}^{n} E_z^i \geq \sum_{j=1}^{n} E_z^j \]
\[ + 4 \cdot M \]
\[ - O(n) \cdot \max_{j \in H^k(i)} \sqrt{E(y_j - y_i)^2} \]

(error occurred)