## 1. Introduction

In the last lecture, we ended with a discussion of the application of SDP's to the Balanced Separator problem. Here, we present a rounding algorithm for the SDP and proceed to prove its approximation bound.

## 2. SDP for Balanced Separator

Remember from last time that in order to qualify as a SDP, the program must come in the format

$$
\begin{gathered}
\min C \bullet X \\
A \bullet X \geq b \\
X \succeq 0
\end{gathered}
$$

We can see that the format of our SDP for solving Balanced Separator (that we discussed last time) is in that format:

$$
\begin{array}{cc}
\min \sum_{(i, j) \in E} \frac{1}{4}\left\|v_{i}-v_{j}\right\|^{2} \\
\forall i, j, k & \left\|v_{i}-v_{j}\right\|^{2}+\left\|v_{j}-v_{k}\right\|^{2} \geq\left\|v_{i}-v_{k}\right\|^{2} \\
& \sum_{i, j \in V}\left\|v_{i}-v_{j}\right\|^{2} \geq 4 c(1-c)
\end{array}
$$

Theorem 2.1. There is a rounding algorithm for this SDP which finds a $\frac{c}{100}$-balanced cut ( $S, S^{c}$ ) such that

$$
\left|E\left(S, S^{c}\right)\right|=\left|\left\{(i, j) \in E \mid i \in S, j \in S^{c}\right\}\right| \leq O(\sqrt{\log n} \boldsymbol{S D P - O P T})
$$

This is called a pseudo-approximation, since the set output by the algorithm is $\frac{c}{100}$-balanced, but it is being compared to the best c-balanced cut. [ARV09]

To prove this theorem, we will have to find sets of vertices that are both large and far apart. To state this idea formally, we will state the following lemma:

Lemma 2.2. It suffices to find disjoint $L, R \in V$ such that they are
(1) Large in $\Omega(n)$, or more formally stated, $|L||R| \geq \alpha(n)$.
(2) $\Delta$-separated for $\Delta=\Omega\left(\frac{1}{\log n}\right)$

Let us define what $\Delta$-separated means so that we can formalize this notion of "farness". Recall that a distance from a vertex $i$ to a set $S$ is defined as $d(i, S)=\min _{j \in S} d(i, j)$ (distance to the closest point in the set) and a distance from a set $S$ to set $T$ is defined as $d(S, T)=\min _{i \in S, j \in T} d(i, j)$ (distance between the two closest points from each of the sets).

Definition 2.3. $L, R$ are $\Delta$-separated if $d(i, j) \geq \Delta$ for $i \in L$ and $j \in R$.

As we discussed in the last lecture, the metric we are using for $d(i, j)$ is $\left\|v_{i}-v_{j}\right\|^{2}$. Let us step back and ponder for a second on what sorts of answers the SDP will try to find. We can see that the SDP tries to minimize the average length of edges because of the $\min \sum \frac{1}{4}\left\|v_{i}-v_{j}\right\|^{2}$ condition. We realize that it is intuitively reasonable to think that if the edges are short on average, at least one of them must be short.


Let us formalize this notion and begin our proof of our lemma:
Proof. Pick a $r$ uniformly at random from $[0, \Delta]$. Let

$$
L_{r}=\{i \in V \mid d(i, L) \leq r\}
$$

Hence, $\left(L_{r}, L_{r}^{c}\right)$ is a cut. We want to show that the expected size of this cut is small. Consider a $(i, j) \in E$. Without loss of generality, let $d(i, L) \leq d(j, L)$. Let us consider if this particular edge is cut. We have

$$
\operatorname{Pr}_{r}[(i, j) \text { is cut }]=\frac{1}{\Delta}(d(j, L)-d(i, L)) \leq \frac{1}{\Delta} d(i, j)
$$

We have that this probability is equal to the difference between the distances over delta because our randomly chosen $r$ must fall within this area for the edge to be a cut. We can reduce this term to $d(i, j)$ through the careful use of triangle inequality on norms. Now, we can combine these probabilities to get an average number of edges expected to be in the cut.

$$
\underset{r}{\mathbf{E}}\left[\left|E\left(L_{r}, L_{r}^{c}\right)\right|\right]=\sum_{(i, j) \in E} \operatorname{Pr}_{r}[(i, j) \text { is cut }] \leq \frac{1}{\Delta} \sum_{i, j} d(i, j)
$$

We can see that the quantity $\sum_{i, j} d(i, j)$ will be the minimum, due to the $\min \sum \frac{1}{4}\left\|v_{i}-v_{j}\right\|^{2}$ objective in the SDP. Hence, we have

$$
\underset{r}{\mathbf{E}}\left[\left|E\left(L_{r}, L_{r}^{c}\right)\right|\right]=\sum_{(i, j) \in E} \operatorname{Pr}_{r}[(i, j) \text { is cut }] \leq \frac{1}{\Delta} \sum_{i, j} d(i, j) \leq O\left(\frac{\text { SDP-OPT }}{\Delta}\right)
$$

Technically, we still have to argue about the sizes of $L_{r}$ and $L_{r}^{c}$, but we can easily follow a similar probabilistic argument as above and show that $\left|L_{r}\right|\left|L_{r}^{c}\right| \geq \alpha(n)$. The details are left as an exercise to the reader.

This lemma is surprising in the sense that we were able to show that we could produce a sufficient cut without dealing with the particular details on how the graph was constructed or its properties. We were able to argue this using probabilistic methods on what amounts to a group of vectors on a unit sphere.

## 3. Algorithm for Pseudo-Approximation of Balanced Separator

Now we will move on to state an algorithm on how to find the $L$ and $R$ to satisfy the lemma above.

### 3.1. Algorithm for finding $L$ and $R$.

(1) $g \sim N(0,1)^{n \times 1}$ (we project onto a random Gaussian vector)
(2) $z_{i} \leftarrow v_{i}^{\top} g$ (we project our $v_{i}$ 's onto a line using a dimensional reduction system similar to discussed earlier in the course)
(3) Find the median $m$ of $\left\{z_{i}\right\}$
(4) Let $L=\left\{i \in V \mid z_{i} \leq m\right\}$ and $R=\{i \in V \mid z>m+\sigma\}$. Our $\sigma$ is a separator constant that we will get to pick later to fit our needs.
There are a number of claims we can make about the $L$ and $R$ generated from this algorithm.
Proposition 3.1. L is $\frac{1}{2}$-large
This is not too hard to see from how $L$ and $R$ were constructed.
Proposition 3.2. $L$ and $R$ are $\frac{\sigma^{2}}{8 \log n}$-separated with probability at least $1-\frac{2}{n^{2}}$.
Proof. Let us assume that $L$ and $R$ are not $\frac{\sigma^{2}}{8 \log n}$-separated. So there must exist a $i \in L$ and $j \in R$ such that $d(i, j) \leq \frac{\sigma^{2}}{8 \log n}$.

Let us consider any $(i, j)$ such that $d(i, j) \leq \frac{\sigma^{2}}{8 \log n}$. We also know that if $(i, j) \in L \times R$, then $\left|z_{i}-z_{j}\right| \geq \sigma$ by the separation condition in our construction.

Before we move on, let's get a refresher on what $\left|z_{i}-z_{j}\right|$ (distance on the one-dimensional line) and $d(i, j)$ (distance in the original space) essentially are. From the construction, we have:

$$
\begin{aligned}
\left|z_{i}-z_{j}\right| & =\left|g^{\top}\left(v_{i}-v_{j}\right)\right| \\
d(i, j) & =\left\|v_{i}-v_{j}\right\|^{2}
\end{aligned}
$$

so by the definition of normal distributions, we can also rewrite this as

$$
z_{i}-z_{j} \sim N\left(0,\left\|v_{i}-v_{j}\right\|^{2}\right) \sim\left\|v_{i}-v_{j}\right\| N(0,1)
$$

(We will let $h \sim N(0,1)$ as a shorthand from now on.)
Essentially, we are going to show that the chances of two things that start out farther in the original space being put closer in the projection is low by concentration. Then, we can take a union bound on all the edges to find the probability of this not happening to any of the edges.

$$
\begin{aligned}
\operatorname{Pr}\left[i \in L, j \in R \text { where } d(i, j) \leq \frac{\sigma^{2}}{8 \log n}\right] & \\
& \leq \operatorname{Pr}\left[\left|z_{i}-z_{j}\right| \geq \sigma \text { where }\left\|v_{i}-v_{j}\right\|^{2} \leq \frac{\sigma^{2}}{8 \log n}\right] \\
& =\operatorname{Pr}\left[\left\|v_{i}-v_{j}|\||h|| \geq \sigma \text { where }\right\| v_{i}-v_{j} \left\lvert\, \| \leq \frac{\sigma}{\sqrt{8 \log n}}\right.\right] \\
& =\operatorname{Pr}[|h| \geq \sqrt{8 \log n}] \geq \frac{2}{n^{4}}
\end{aligned}
$$

The last inequality rises out of the fact that on a Gaussian $h \sim N(0,1)$ has the property that $\operatorname{Pr}[|h| \geq t] \leq 2 e^{\frac{-t^{2}}{2}}$.

We can now take the union bound to finish this proof.

$$
\operatorname{Pr}\left[\exists i \in L, j \in R \text { such that } d(i, j) \leq \frac{\sigma^{2}}{8 \log n}\right] \leq n^{2} \frac{2}{n^{4}}=\frac{2}{n^{2}}
$$

Proposition 3.3. $R$ is $\frac{c^{\prime}}{16}$-large with probability at least $\frac{\frac{c}{}^{\prime}}{16}$ if $\sigma=\frac{\frac{c}{}_{\prime}^{4}}{4}$.
Here, we can use the averaging argument in the proof. Recall that this lemma was given out as a homework assingment.

Lemma 3.4. If $X \in[0, B]$ and $\mathbf{E} X=\epsilon B$ then $\operatorname{Pr}\left[X \geq \frac{\epsilon B}{2}\right] \geq \frac{\epsilon}{2}$.
The proof of this lemma is simple and by contradiction, where we argue that if this weren't the case, then we must have an expectation that is smaller than $\epsilon B$.

Let us return to proving the proposition above.
Proof. We will rely on the properties of Gaussians for this proof.
(1) We recognize that $\left\|v_{i}-v_{j}\right\| \leq 2$ because $v_{i}$ and $v_{j}$ are vectors on the unit sphere.
(2) $\sum_{(i, j) \in E}\left\|v_{i}-v_{j}\right\|^{2} \geq 4 c^{\prime} n^{2}$ and hence, by (1), we have

$$
2 \sum_{(i, j) \in E}\left\|v_{i}-v_{j}\right\| \geq \sum_{(i, j) \in E}\left\|v_{i}-v_{j}\right\|^{2} \geq 4 c^{\prime} n^{2} \Longrightarrow \sum_{(i, j) \in E}\left\|v_{i}-v_{j}\right\| \geq 2 c^{\prime} n^{2}
$$

(3) Thus, there exist at least $\frac{c^{\prime}}{2} n^{2}$ pairs where $\left\|v_{i}-v_{j}\right\| \geq c^{\prime}$ by the averaging argument, where we set

$$
\begin{gathered}
X=\left\|v_{i}-v_{j}\right\| \\
i, j \sim_{r}[n] \\
\mathbf{E} X \geq 2 c^{\prime} n^{2} \\
0 \leq X \leq 2
\end{gathered}
$$

Now, let us use the property of Gaussians where they are not "too concentrated", or more formally stated

$$
\begin{gathered}
z_{i}-z_{j} \sim\left\|v_{i}-v_{j}\right\| N(0,1) \text { where } h \sim N(0,1) \\
\operatorname{Pr}[|h| \leq t] \leq \frac{1}{\sqrt{2 \pi}} 2 t \leq t
\end{gathered}
$$

(4) For a fixed $i, j$ such that $\left\|v_{i}-v_{j}\right\| \geq c^{\prime}$

$$
\operatorname{Pr}\left[\left|z_{i}-z_{j}\right| \leq \frac{c^{\prime}}{2}\right] \leq \frac{1}{2} \Longrightarrow \operatorname{Pr}\left[\left|z_{i}-z_{j}\right| \geq \frac{c^{\prime}}{2}\right] \geq \frac{1}{2}
$$

$$
\begin{equation*}
\mathbf{E}\left[\text { number of }(i, j) \text { such that }\left|z_{i}-z_{j}\right| \geq \frac{c^{\prime}}{2}\right] \geq \frac{c^{\prime}}{4} n^{2} \tag{5}
\end{equation*}
$$

(6) By using the averaging argument where $X$ is now the number of pairs that are far apart in the projection, we have

$$
\operatorname{Pr}\left[\text { At least } \frac{c^{\prime}}{8} n^{2} \text { pairs }(i, j) \text { satisfy }\left|z_{i}-z_{j}\right| \geq \frac{c^{\prime}}{2}\right] \geq \frac{c^{\prime}}{8}
$$

(7) By eliminating the double-counting on pairs, we have

$$
\operatorname{Pr}\left[\text { At least } \frac{c^{\prime}}{8} n^{2} \text { vertices } i \text { satisfy }\left|z_{i}-m\right| \geq \frac{c^{\prime}}{4}\right] \geq \frac{c^{\prime}}{8}
$$

(8) To finish off, by construction, we have

$$
\operatorname{Pr}\left[\text { At least one of } R_{1}, R_{2} \text { is } \frac{c^{\prime}}{16} \text {-large }\right] \geq \frac{c^{\prime}}{8} \Longrightarrow \operatorname{Pr}\left[R_{1} \text { is } \frac{c^{\prime}}{16} \text {-large }\right] \geq \frac{c^{\prime}}{16}
$$

Let us go back up the stack in the proofs of the claims of the lemma in order to clarify exactly what we have shown.
(1) We have just shown that $R$ is sufficiently large with a certain probability. By triviality, $L$ too is sufficiently large.
(2) We have seen by the proof of Proposition 3.2 that $L$ an $R$ are well-separated by Gaussian concentration.
(3) This means that a breadth-first search is possible on all possible cuts, and we should be able to find one that is good.
It is interesting to note that the only constant guarantee we get here is the $\frac{c^{\prime}}{16}$-largeness. The other bound, the one on separation, is $\frac{\sigma^{2}}{8 \log n}$ and not a constant.

## 4. Summary

Let us go all the way back to the top and summarize what we have done to approximate the Balanced Separator problem.
(1) Solve the SDP so that we minimize the average length of the edge
(2) Find two sets of vertices that are sufficiently large and sufficiently far
(3) Take vertices within a $r$-radius for a cut.
(4) Follow the algorithm to project these vertices to a line using a Gaussian vector and then sort and divide around the median.
(5) Prove that the sets we have found through this method are well-separated in the original space. In particular, we can analyze one such pair. The probability that they end up closer on the projection is very small, namely $\frac{1}{n^{4}}$.

## References

[ARV09] Sanjeev Arora, Satish Rao, and Umesh Vazirani. Expander flows, geometric embeddings and graph partitioning. J. ACM, 56(2):5:1-5:37, April 2009.

