SOS Lecture 3: Sparsest cut and the ARV algorithm

Boaz Barak

July 1, 2014

Suggested reading Section 15.4 in the approximation algorithms book of Williamson and Shmoys (available online) for more details. See also Section 3 in the book chapter of Chlamtac and Tulsiani (available on `http://ttic.uchicago.edu/~madhurt/Papers/sdpchapter.pdf`) for a good shorter overview of the proof. Trevisan also discusses the ARV algorithm in his Lectures 11 and 14 (though the heart of the analysis is in Lecture 15 which has no online notes).

Setting $G = (V, E)$ is d-regular graph on n vertices. It has normalized adjacency matrix A ($A_{i,j} = 1/d$ if $(i,j) \in E$ and $A_{i,j} = 0$ otherwise). Let $L = I - A$ be its normalized *Laplacian matrix.*

Sparsest cut problem Find set $S \subseteq V$ that minimizes $\phi(S) = nE(S, \overline{S})/(d|S||\overline{S}|)$. Let $\phi(G) = \min_S \phi(S)$.

Last lecture we saw the following theorem:

Cheeger-Alon-Milman $\exists$ poly-time algorithm to find S with $\phi(S) = O(\sqrt{\phi(G)})$.

This lecture we will show the ideas behind the following result:

Arora-Rao-Vazirani $\exists$ poly-time algorithm to find S with $\phi(S) = O(\sqrt{\log n}\phi(G))$

The algorithm uses (a subset of) the degree 6 SOS program. As before, we will assume that we have a pseudo-distribution over sets T with $\phi(T) \leq \phi$ for some ϕ. In fact, to simplify matters we will again assume that the sets have size exactly $1/2$ and represent each set by a ± 1-valued vector x.

Thus we have a pseudo-distribution $\{x\}$ over $\{\pm 1\}$ (i.e., satisfying $\{x_i^2 = 1\}$) satisfying $\{\sum x_i = 0\}$ and

$$\mathbb{E}_{(i,j)\in G}\tilde{\mathbb{E}}(x_i - x_j)^2 \leq \phi \ .$$

For simplicity we assume that all edges and vertices contribute the same to the solution (using convexity arguments this can be shown to be without loss of generality) and so we assume that for every i, $\mathbb{E}x_i = 0$, and for every $(i,j) \in E$, $\tilde{\mathbb{E}}(x_i - x_j)^2 = \phi$. Note that this means that

$$\sum_{i,j} \tilde{\mathbb{E}}(x_i - x_j)^2 = 2n \sum \tilde{\mathbb{E}}x_i^2 - 2\tilde{\mathbb{E}}(\sum_i x_i)^2 = 2n^2$$

and hence the random variables are "well spread" in the sense that $\mathbb{E}_{i,j}\tilde{\mathbb{E}}(x_i - x_j)^2 \geq 1/10$.

Outline of the proof We will give the proof under the assumption that the distribution $\{x_i\}$ is an actual distribution. We later comment on what we actually used about this assumption. This is fairly common when working with the SOS algorithm.

Reducing to vertex separator Define the following graph H of n vertices– we say that $i \sim_H j$ if $\Pr[x_i = x_j] < \Delta$ for some parameter Δ that we will choose later to be roughly $1/\sqrt{\log n}$. The main lemma is that

Definition: An n-vertex graph H is *separable* if there are disjoint sets $L, R \subseteq V(H)$ such that $|L|, |R| \geq n/1000$ and $E(L,R) = \emptyset$.

ARV Main Lemma: Let H be a graph defined by ± 1-valued random variables $\{x_i\}_{i \in [n]}$ such that $i \sim_H j$ iff $\Pr[x_i = x_j] < \Delta$, where $\Delta \ll 1/\sqrt{\log n}$, $\mathbb{E} x_i = 0$ and $\mathbb{E}_{i,j}(x_i - x_j)^2 > 1/10$. Then H is separable.

From main lemma to algorithm: The main lemma easily implies an $O(\sqrt{\log n})$ approximation algorithm for the sparsest cut problem. If we start a BFS from L until we reach R, we will obtain a sparse cut (exercise).

Why is $\Delta \ll \sqrt{\log n}$ necessary? Here is an example showing that the condition $\Delta \ll 1/\sqrt{\log n}$ is necessary. Let H be the graph with $n = 2^\ell$ vertices that we identify with the vectors of the cube $\{\pm 1\}^\ell$. We can view those as correlated random variables $x_1, \ldots, x_n$ which are sampled by choosing $i \in [\ell]$ and then letting $x_\alpha = \alpha_i$ for every index $\alpha \in \{\pm 1\}^\ell$. Suppose that $\Delta = c/\sqrt{\log n} = c/\sqrt{\ell}$ for some $c \gg 1$. Then it can be shown that for every set L of measure $1/1000$, the set $\Gamma(L)$ of all vectors having Hamming distance at most $\Delta \ell = c\sqrt{\ell}$ has measure $1 - o(1)$ and hence the graph is not separable. (The best set L to take would be the set $\{\alpha : \sum \alpha_i \geq \sqrt{\log(1000)\ell}\}$ but then $\Gamma(L) = \{\alpha : \sum \alpha_i \in [-(c - \log(1000))\sqrt{\ell}, +(c + \log(1000))\sqrt{\ell}]\}$ which has measure $1 - o(1)$ for $c \gg 1$.

Formally we have the following theorem:

Expansion of the Boolean cube Let $L \subseteq \{\pm 1\}^n$ such that $|L| \geq \Omega(2^n)$ and let $c \gg 1$ then the set $\tilde{L} = \{x : \exists y \in L \text{ s.t. } \|x - y\|_1 \leq c\sqrt{n}\}$ has size $(1 - o(1))2^n$.

Why is $\Delta \ll \sqrt{\log n}$ sufficient? Of course the hard part (which we will not fully prove) is showing that there is in fact a separator as long as $\Delta \ll \sqrt{\log n}$. We provide some intuition:

Let $\{y\}$ be the Gaussian disrtribution that matches the first two moments of $\{x\}$. We sample from y and let L be $\{i : y_i < -10\}$ and $R = \{i : y_i > +10\}$. These two sets might have some edges between them, but we will see that they can be "pruned" to remove those.

First, note that if $\Pr[x_i - x_j] = \frac{1}{4}\mathbb{E}(x_i - x_i)^2 \leq \Delta$ then $y_i - y_j$ is a Gaussian random variable with mean zero and variance Δ. Hence the probability that $|y_i - y_j| \geq 20$ is $\exp(-(1/\Delta))$. If $\Delta \ll 1/\log n$ then this probability is $n^{-\omega(1)}$ and hence there would be no pair (i,j) such that $i \sim_H j$ but $i \in L$ and $j \in R$ and we would be done. For $\Delta \ll 1/\sqrt{\log n}$, if the average degree of H is at most $2^{O(\sqrt{\log n})}$ (which incidentally is the case in the cube example above) then we can still do a union bound over all edges (i,j) in H and reach the same conclusion.

Handling higher degrees For every choice of y, let M_y be directed graph containing all edges $i \to j$ such that $i \sim_H j$ but $y_j - y_i \geq 20$. If M_y has $o(n)$ edges then we can simply remove all the vertices touching an edge of M_y from L and R and get the vertex separator. More generally, if M_y has a *vertex cover* of $o(n)$ size then we can simply remove it and complete the proof.

To handle the other case, where the graph has no small vertex cover, we follow the old Swedish proverb

One man's left is another man's right

That is, we say that since a vertex is equally likely to be in L as it is to be in R, we may hope that it would be the case that we are lucky and M_y, rather than just being a collection of edges from L to R, actually also contains some edges in the other direction as well. If that is the case, maybe we could actually find longer *paths* in M_y of the form $i_1 \to i_2 \to \cdots \to i_t$ in the graph M_y. This means that

- $y_{i_1} - y_{i_t} \geq 10t$
- $\frac{1}{4}\mathbb{E}(x_{i_1} - x_{i_t})^2 = \Pr[x_{i_1} \neq x_{i_t}] \leq t\Delta$ (using the union bound).

But since $y_{i_1} - y_{i_j}$ is a Gaussian random variable with expectation zero and variance at most $t\Delta$, the probability that it is at least $10t$ is at most

$$\exp(-10^2\sqrt{t\Delta}) \tag{1}$$

which would be $o(n^{-2})$ if we can get $t\Delta = \log n$, in which case by the union bound, there would not exist such a pair, contradicting our assumption.

Making this argument into a reality To actually make this argument formal takes a lot of work, and we will not show the full proof, but here is the intuition. Interestingly, we will use the same expansion theorem in the cube that was used for the counterexample showing that we can't go beyond $\sqrt{\log n}$.

The main issue is the following - using simple averaging we can assume without loss of generality that for every vertex j, with some probability $\delta > 0$ there will exist a vertex i such that $i \sim_H j$ and $y_j - y_i > 10$ and with the same probability there will exist a vertex k such that $i \sim_H k$ and $y_k - y_i > 10$, but to show that we can chain these together and get the path $i \to j \to k$ we need these two events to happen *simultaneously*.

We can think of every random variable x_i as a vector $v_j \in \{\pm 1\}^\ell$ (where ℓ is the number of points in the sample space). To sample the Gaussian variables $y_1, \ldots, y_n$ we can choose a random vector $r \in \{\pm 1\}^\ell$ and let $y_i = \langle x_i, r\rangle/\sqrt{\ell}$. (This is not exactly a Gaussian but is very close to one when ℓ is large; to get exact Gaussian we choose r to be a standard Gaussian vector — it will not change the argument.) So, we can think of the following situation: we have the vector v_0 correspond to the vertex j, and the vectors $v_1, \ldots, v_D$ corresponding to the variables that are Δ-close to x_j (i.e., the neighbors of j in H). We know that with probability at least, say, $1/1000$, over the choice of r there exists i such that $\langle v_0 - v_i, r\rangle \geq 10\sqrt{\ell}$, which also implies that with the same probability there exists k such that $\langle v_k - v_0, r\rangle \geq 10\sqrt{\ell}$. We would like to show that these two events happen simultanesouly with some constant probability.

Let R be the set of r's such that the first event happens for it. Note that $|R| \geq 2^n/1000$. We now consider what happens if we choose a random $r \in R$ and then perturb it to a string r' by modifying a set $I' \subseteq [\ell]$ of $c\sqrt{\ell}$ coordinates for some large constant c. Let i be such that $\langle v_0 - v_i, r\rangle \geq 10\sqrt{\ell}$, and let I be the set of coordinates where v_i and v_0 differ. Since v_0 and v_i correspond to Δ-close variables, $|I| \leq \Delta\ell$, and so with very high probability $|I \cap I'| \leq O(\Delta|I'|) = O(\Delta\sqrt{\ell})$. In this case, the perturbation can change the dot product $\langle v_0 - v_i, r\rangle$ by at most $O(\Delta\sqrt{\ell})$ and hence (since $\Delta = o(1)$) we get that $\langle v_0 - v_i, r'\rangle \geq 9\sqrt{\ell}$.

But by the hypercube expansion theorem, if c is big enough then the set R' obtained by first picking a random $r \in R$ and then perturbing it by $c\sqrt{\ell}$ has measure at least 0.999. Hence we have shown that (slightly weaker versions of) the events actually happen with probability 0.99 and so we can argue that they happen simultaneously. As we continue doing this to get longer paths, the main difference would be that we would need to consider v_0 and v_i that are not of distance 1 but of distance t in the graph, and hence in the calculation above we would need $t\Delta < 1/(10c)$, and so we would only be able to get paths of length $t = O(1/\Delta)$. Plugging this to (**??**) we see that we get a contrdiction as long as $\Delta^{-2} > c \log n$ for some constant c, and hence we get an approximation algorithm of $O(1/\Delta) = O(\sqrt{\log n})$.