S^{3}CS 2014: Sums of Squares July 3–4, 2014

Topic 6: Sparsest Cut and the ARV Algorithm

Lecturer: Boaz Barak *Scribe: Tal Wagner*

In this lecture we revisit the (Uniform) Sparsest Cut problem, a thoroughly studied optimization problem introduced in Lecture 2. We present a breakthrough result of Arora, Rao and Vazirani [?], that achieves an efficient $O\left(\sqrt{\log n}\right)$-approximation algorithm. It is of interest to us since it can be interpreted as an instantiation of the degree-4 SoS algorithm.

We will not give the full proof of [?], since it is very long and involved, but rather focus on presenting the main ideas.

1 Problem Definition

For simplicity we define the problem over regular graphs, and remark that the results presented here extend to general graphs as well.

Definition 1.1 (sparsity)**.** Let $G(V, E)$ be an r-regular graph on n vertices. For a subset of vertices $S \subset V$ such that $S \neq \emptyset, V$, denote $\bar{S} = V \setminus S$, and denote by $\mathcal{E}(S, \bar{S})$ the number of edges crossing the cut $(S, \bar{S})$.

The sparsity of the cut $(S, \bar{S})$, denoted $\phi(S, \bar{S})$, is defined as

$$\phi(S, \bar{S}) = \frac{n\mathcal{E}(S, \bar{S})}{r|S||\bar{S}|}.$$

The sparsity of G, denoted $\phi(G)$, is defined as

$$\phi(G) = \min_{S \subset V : S \neq \emptyset, V} \phi(S, \bar{S}).$$

To get some intuition for this definition, observe that up to the multiplicative factor n/r (which can be thought of as normalization), $\phi(S, \bar{S})$ is the ratio of the number of edges that actually cross the cut to the number of edges that *could have* crossed it (had all the possible edges been present).

Definition 1.2. The Uniform Sparsest Cut problem is, given a d-regular graph $G(V, E)$ over n vertices, to find $S \subset V$ such that $\phi(S, \bar{S}) = \phi(G)$.

2 Main Theorem

The main result of this lecture is an $O\left(\sqrt{\log n}\right)$-approximation for Uniform Sparsest Cut.

Theorem 2.1 (Arora-Rao-Vazirani [?])**.** *There is a randomized polynomial time algorithm, that given an r-regular graph $G(V, E)$ on n vertices, finds with high probability $S \subset V$ such that $\phi(S, \bar{S}) = O(\sqrt{\log n}) \cdot \phi(G)$.*

The best multiplicative approximation factor prior to [?] was $O(\log n)$ due to Leighton and Rao [?]. The Cheeger-Alon-Milman inequality, discussed in Lecture 2, achieves a square-root approximation, i.e. finds $S \subset V$ such that $\phi(S, \bar{S}) = O(\sqrt{\phi(G)})$.[1]

On the other hand, while Uniform Sparsest Cut is known to be NP-hard [?], all currently known inapproximability results rely on stronger hardness assumptions.

[1] As explained in Lecture 2, the Cheeger-Alon-Milman inequality applies to a slight variation of Uniform Sparsest Cut of which optimum cannot exceed 1, and hence a square-root approximation makes sense, i.e. $\phi(G) \leq \sqrt{\phi(G)}$.

3 The ARV Algorithm

We will work under the simplifying assumption that $\phi(G)$ is attained by a bisection - that is, there is $S^* \subset V$ with $|S^*| = n/2$ such that $\phi(G) = \phi(S^*, \bar{S^*})$. Of course, this assumption is not required for the analysis of [?].

We identify our vertex set V with $[n] = \{1, \ldots, n\}$. Let $x^* \in \{\pm 1\}^n$ be the indicator vector of S^*, i.e. for all $i \in V$, $x_i^* = 1$ if $i \in S^*$ and $x_i^* = -1$ otherwise. Observe that we have

$$\mathcal{E}(S^*, \bar{S^*}) = \frac{1}{4} \sum_{\{i,j\} \in E} (x_i^* - x_j^*)^2.$$

Moreover since $(S^*, \bar{S^*})$ is a bisection we have $|S^*| = |\bar{S^*}| = \frac{n}{2}$, and since G is r-regular we have $|E| = \frac{nr}{2}$. Plugging these into Definition 1.1, we may write

$$\phi(G) = \frac{1}{2|E|} \sum_{\{i,j\} \in E} (x_i^* - x_j^*)^2,$$

and now we can formulate our problem in the SoS framework.

3.1 The SoS Program

Our algorithm (that will be fully described in the next section) runs the SoS algorithm to get a degree-4 pseudo-distribution $\{x\}$ satisfying the constraints:

- $x_i^2 = 1 \quad \forall\, i \in [n];$ (3.1)
- $\sum_{i=1}^{n} x_i = 0;$ (3.2)
- $\frac{1}{2|E|} \sum_{\{i,j\} \in E} (x_i - x_j)^2 \leq \phi.$ (3.3)

The constraints in eq. (3.1) ensure that the pseudo-distribution is over vectors in $\{\pm 1\}^n$. The constraint eq. (3.2) ensures that the pseudo-distribution is over bisections (same number of $+1$'s and -1's). Eq. (3.3) is the minimization of our objective function: we can perform a binary search in order to find the minimal ϕ for which these constraints are satisfiable (this is a standard reduction of optimization to feasibility). Note that $\phi \leq \phi(G)$, since we have x^* that attains $\phi(G)$.

Our goal is to show that we can obtain from $\{x\}$ a subset $S \subset V$ that meets our approximation requirement. We make one more assumption, that our pseudo-distribution is "well spread", in the sense that

$$\mathbb{E}_{i,j \in [n]} \left[\tilde{\mathbb{E}} \left[(x_i - x_j)^2 \right] \right] \geq \frac{1}{10}. \tag{3.4}$$

It turns out that in the complementary case, it is not difficult to obtain a constant-factor approximation (and this was used prior to ARV). Hence this assumption focuses us on the challenging case.

3.2 The Algorithm

To obtain an approximate solution from $\{x\}$, our approach is by reduction to a vertex separation problem on graphs.

Definition 3.1. Let H be a graph over n vertices. We say H is *separable* if there are disjoint subsets of vertices L, R, such that $|L|, |R| \geq \Omega(n)$ and there are no edges crossing between them.

The key structural result of ARV is the following.

Lemma 3.2 (ARV main lemma)**.** *Let $\{x\}$ be a pseudo-distribution obtained from the SoS program in Section 3.1 on the input graph $G(V,E)$.*

Put $\Delta = c/\sqrt{\log n}$ for a sufficiently small constant c. Define the graph $H(V,E_H)$ on the same vertex set as G, with $\{ij\} \in E_H$ iff $\tilde{\mathbb{E}}\left[(x_i - x_j)^2\right] < \Delta$.

Then H is separable, and the subsets L,R can be found with high probability in randomized polynomial time.

Using this lemma we can obtain a sparse cut using standard techniques.

Lemma 3.3. *Given subsets L,R as guaranteed by Lemma 3.2, we can efficiently find $S \subset V$ such that $\phi(S,\bar{S}) \le O(\phi/\Delta) = O(\sqrt{\log n})\phi$.*

Now we can fully state the algorithm.

ARV approximation algorithm for Uniform Sparsest Cut:

1. Solve the SoS program stated in Section 3.1.
2. Construct the graph H as described above.
3. Apply Lemma 3.2 to find disjoint subsets $L,R \subset V$ as in Definition 3.1.
4. Use Lemma 3.3 to obtain from L,R a subset $S \subset V$ such that $\phi(S,\bar{S}) \le O(\sqrt{\log n})\phi(G)$.

4 Analysis

We will perform the analysis under the assumption that $\{x\}$ is an actual distribution, and later verify that all arguments used hold remain intact when $\{x\}$ is a pseudo-distribution (of degree 4, in this case). This is fairly common when working with the SoS algorithm.

4.1 Preliminaries

Suppose that $\{x\}$ is an actual distribution over vectors in $\{\pm 1\}^n$. The entries of a vector drawn from $\{x\}$ are correlated random variables $x_1, \ldots, x_n$ taking values in $\{\pm 1\}$. We visualize $\{x\}$ as a $\{\pm 1\}$-matrix $A_{\{x\}}$ of size $\ell \times n$, with columns corresponding to $x_1, \ldots, x_n$ and rows corresponding to the points in the sample space of $\{x\}$ (so ℓ is the size of the sample space). The x_i's can now be thought of as (column) vectors in $\{\pm 1\}^\ell$.[2] We stress that $\{x\}$ is a distribution over vectors in $\{pm1\}^n$ describing bisections in G (and corresponding to rows of $A_{\{x\}}$), while its random coordinates x_i's are interpreted as vectors in $\{\pm 1\}^\ell$ (corresponding to columns of $A_{\{x\}}$). The reader is alerted to avoid confusion.

We go on to define a notion of distance between the x_i's. For $t = 1, \ldots, \ell$, let p_t denote the probability to sample from $\{x\}$ the value of the t^{th} row in $A_{\{x\}}$. For simplicity one may think of $p_1, \ldots, p_\ell$ as the uniform distribution; it does not change the analysis.

Definition 4.1. Define $d : [n] \times [n] \to \mathbb{R}_{\ge 0}$ by

$$d(i,j) = \sum_{t=1}^{\ell} p_t (x_i(t) - x_j(t))^2 = \mathbb{E}\left[(x_i - x_j)^2\right],$$

It straightforward to verify that d is a *distance function*, in the sense that it satisfies the following properties:

[2] We remark in the usual presentation of the ARV analysis, which is outside of the SoS framework, the vectors are not assumed to have entries in $\{\pm 1\}$.

1. $d(i,i) = 0$ for all i;
2. (Symmetry) $d(i,j) = d(j,i)$ for all i,j;
3. (Triangle inequality) $d(i,k) \le d(i,j) + d(j,k)$ for all i,j,k.

In fact, d is equivalent to both the Hamming distance and the $\|\cdot\|_1$-distance.[3]

4.2 Why is $\Delta \ll 1/\sqrt{\log n}$ Necessary?

Before turning to the main part of the analysis, let us show why our approach can only work up to $\Delta \ll 1/\sqrt{\log n}$ and not greater values, which would have given a better approximation factor. Note that we are showing this limitation even for actual distributions (rather than just pseudo-distributions).

We rely on the following theorem, stated here without proof.

Theorem 4.2 (expansion of the boolean hypercube)**.** *Let $c > 0$ be a sufficiently large constant. For $L \subset \{\pm 1\}^\ell$, denote*

$$\tilde{L} = \left\{ u \in \{\pm 1\}^\ell : \exists u \in L \;\; s.t. \;\; \|v - u\|_1 \le c\sqrt{\ell} \right\}.$$

If $|L| \ge \Omega(2^\ell)$, then $|\tilde{L}| \ge (1 - o(1))2^\ell$.

Put $\ell = \log n$, and let A be a $\{\pm 1\}$-matrix of dimensions $\ell \times n$, such that its n columns are all the $2^\ell = n$ possible $\{\pm 1\}$-vectors of length ℓ. (The order of the columns does not matter.) Consider the distribution $\{x\}$ defined by uniformly sampling a row from A. Note that for this distribution, A is exactly the matrix $A_{\{x\}}$ defined in the previous section.

Since $\{x\}$ is uniform over its support, it can be easily seen that

$$d(i,j) = \mathbb{E}\left[(x_i - x_j)^2\right] = \frac{\|x_i - x_j\|_1}{2\ell}.$$

Recall that we have $\Delta = c/\sqrt{\log n} = c/\sqrt{\ell}$. In the graph H defined in Lemma 3.2, a pair of vertices i,j is adjacent if $d(i,j) < \Delta$, or equivalently (by the above), if $\|x_i - x_j\|_1 < 2\Delta\ell = 2c\sqrt{\ell}$. If we take c to be a large constant (rather than a small constant as in Lemma 3.2), then we can apply Theorem 4.2. It tells us that for any choice of L such that $|L| \ge \Omega(n)$, the subset R of vertices that have no neighbors in L would have size no larger than $o(n)$. This means that H is not separable, and the conclusion of Lemma 3.2 does not hold.

4.3 Why is $\Delta \ll 1/\sqrt{\log n}$ Sufficient? Proof of Lemma 3.2

We now turn to the core of the analysis, of showing that if $\Delta = c/\sqrt{\log n}$ then H is separable (with constant probability, that can then be boosted).

Apply the Quadratic Sampling Lemma (see previous lecture for the formal statement and proof) to obtain a Gaussian distribution $\{y\}$ that matches the first two moments of $\{x\}$. This is a distribution over vectors in $\mathbb{R}^n$ (with correlated entries), such that y_i is associated with the vertex i in H. Sample $y \sim \{y\}$ and define

$$L = \{i : y_i < -10\} \;\;,\;\; R = \{i : y_i > 10\}.$$

By the SoS program constraints, for each i we have $\mathbb{E}\left[y_i\right] = \mathbb{E}\left[x_i\right] \in [-1,1]$ and $\mathbb{E}\left[y_i^2\right] = \mathbb{E}\left[x_i^2\right] = 1$. Hence by standard properties of Gaussian random variables, we see that with high probability, $|L|, |R| \ge \Omega(n)$. If there are no edges crossing between L and R, then we are done.

[3] The equivalence is up to normalization and weighting by $p_1, \ldots, p_\ell$, and up to a factor of 4 (for Hamming distance) or 2 (for $\|\cdot\|_1$-distance..

Otherwise, we wish to remove a small subset of vertices from L and R in a way that eliminates all the edges crossing between them, but retains their linear sizes. For convenience, we treat all edges crossing between L and R as oriented edges from L to R.

To analyze this approach, let us derive a simple bound on the probability of an edge to cross from L to R. Let $\{ij\}$ be some edge in H. By definition this means that $d(i,j) = \mathbb{E}\left[(x_i - x_j)^2\right]\Delta$. Moreover we know that $\mathbb{E}\left[y_i - y_j\right] = \mathbb{E}\left[x_i - x_j\right] =\in [-2,2]$. Hence $y_i - y_j$ is a Gaussian random variable with mean in $[-2,2]$ and variance bounded by Δ, and therefore,

$$\Pr\left[|y_i - y_j| > 20\right] = \exp(-1/\Delta). \tag{4.1}$$

This means that the edge $\{ij\}$ crosses from L to R with probability at most $\exp(-1/\Delta)$.

Now let us consider some examples of how we can avoid edges crossing from L to R.

- Suppose Δ is as small as $\Delta \leq 1/3 \log n$. Then the probability in eq. (4.1) is roughly $1/n^3$, which means, by a union bound over all edges in H (of which there are at most n^2), that with constant probability there are no edges crossing from L to R. In this case the proof is finished.
- Suppose H has at most $2^{O(\sqrt{\log n})}$ edges. Recalling our choice of Δ as $c/\sqrt{\log n}$ for a sufficiently small constant c, we see that we can apply the same union bound argument as above.
-

4.4 From Vertex Separation to Sparse Cut: Proof of Lemma 3.3

We now prove Lemma 3.3. The proof is standard and is outside the core analysis of ARV (which is Lemma 3.2); it is presented here for the sake of completeness.

Relying on the definition of d as a distance function between points in $[n]$, we can define a notion of distance between a point and a subset of points: For $B \subset [n]$ and $i \in [n]$, we let

$$d(B,i) = \min_{b \in B} d(b,i).$$

It is easy to verify that from the standard triangle inequality of d, we can obtain the following triangle inequality for subsets: For $B \subset [n]$ and $i,j \in [n]$,

$$d(B,i) \leq d(B,j) + d(j,i). \tag{4.2}$$

To obtain our sparse cut, we sample $\tau \in (0,\Delta)$ uniformly at random (recall that Δ was defined in Lemma 3.2) and let

$$S = \{i \in [n] : d(L,i) \leq \tau\}.$$

Clearly we have $L \subset S$. Observe that in Lemma 3.2, the graph H was defined such that i,j are neighbours in H iff $d(i,j) < \Delta$. By hypothesis of Lemma 3.3 there are no edges crossing between L and R, and therefore $R \subset \bar{S}$. Since $|L|,|R| \geq \Omega(n)$, we conclude that

$$|S| \cdot |\bar{S}| \geq \Omega(n^2). \tag{4.3}$$

We turn to counting the edges crossing the cut $(S,\bar{S})$. Let $\{ij\}$ be an edge in G and let χ_{ij} be the 0-1 random variable indicating whether the edge crosses $(S,\bar{S})$. Suppose w.l.o.g. that $d(L,i) \leq d(L,j)$. The edge crosses $(S,\bar{S})$ iff both $\tau \geq d(L,i)$ and $\tau < d(L,j)$ occur, and hence

$$\mathbb{E}_\tau\left[\chi_{ij}\right] = \Pr[\tau \in [d(L,i), d(L,j))] = \frac{d(L,j) - d(L,i)}{\Delta} \leq \frac{d(i,j)}{\Delta},$$

where the final inequality is by eq. (4.2). Summing over all edges,

$$\mathbb{E}_\tau\big[\mathcal{E}(S,\bar{S})\big] = \mathbb{E}_\tau\big[\sum_{\{i,j\}\in E} \chi_{ij}\big] = \sum_{\{i,j\}\in E} \mathbb{E}_\tau\big[\chi_{ij}\big] \le \frac{\sum_{\{i,j\}\in E} d(i,j)}{\Delta},$$

and now using Markov's inequality we can find with high probability a threshold τ such that

$$\mathcal{E}(S,\bar{S}) \le O(1)\cdot \frac{\sum_{\{i,j\}\in E} d(i,j)}{\Delta}. \tag{4.4}$$

Finally, note that by eq. (3.3) we have

$$\sum_{\{i,j\}\in E} d(i,j) = \sum_{\{i,j\}\in E} \mathbb{E}\big[(x_i - x_j)^2\big] \le 2|E|\phi = 2nr\phi.$$

Combining this with eq. (4.3) and eq. (4.4) we find that

$$\phi(S,\bar{S}) = \frac{n\mathcal{E}(S,\bar{S})}{r|S||\bar{S}|} \le O(1)\cdot\frac{\phi}{\Delta},$$

as required.

4.5 From Actual Distribution to Pseudo-Distribution

5 Temp Graveyard

5.1 Squared Triangle Inequality for $\{\pm 1\}$

The reason we need a degree-4 SoS program, is for $\{x\}$ to have the following property.

Lemma 5.1. *Let $\{x\}$ be a degree-4 pseudo-distribution satisfying the constraints $\{\forall i,\ \ x_i^2 = 1\}$. Then for all i,j,k,*

$$\tilde{\mathbb{E}}\big[(x_i - x_k)^2\big] \le \tilde{\mathbb{E}}\big[(x_i - x_j)^2\big] + \tilde{\mathbb{E}}\big[(x_j - x_k)^2\big]. \tag{5.1}$$

Proof. We first note that if $\{x\}$ was an actual distribution over $\{\pm 1\}^n$ then the lemma would be easy, since the inequality $(x_i - x_k)^2 \le (x_i - x_j)^2 + (x_j - x_k)^2$ would hold for every i,j,k (this is trivial to verify by case analysis) and hence would clearly hold in expectation. However, in order to prove the lemma for pseudo-expectation, we need to give an SOS proof.

By linearity of pseudo-expectation, eq. (5.1) is equivalent to

$$\tilde{\mathbb{E}}\big[(x_i - x_j)^2 + (x_j - x_k)^2 - (x_i - x_k)^2\big] \ge 0.$$

By rearranging, this becomes

$$\tilde{\mathbb{E}}\big[(x_j - x_i)(x_j - x_k)\big] \ge 0.$$

Denote $P(x) = (x_j - x_i)(x_j - x_k)$. We need to show that $\tilde{\mathbb{E}}\big[P(x)\big] \ge 0$, so by definition, we need to find a polynomial $Q(x)$ such that $P = Q^2$ (over $\{\pm 1\}$).

We put $Q = \frac{1}{2}P$. The fact that $P = (\frac{1}{2}P)^2$ can be verified either by explicitly expanding Q^2 and plugging $x_i^2 = x_j^2 = x_k^2 = 1$, or by just observing that $P(x) \in \{0,4\}$ over $\{\pm 1\}$, which renders $P = (\frac{1}{2}P)^2$ immediate to see.

Note that our SoS proof Q^2 of eq. (5.1) is a polynomial of degree 4, and this is why we need $\{x\}$ to be a pseudo-distribution of atleast this degree. □

References