Lecture 11: Spectral sparsification

We study a generalization of the cut sparsification problem, called spectral sparsification, and see how this problem can be solved nicely using random sampling and a matrix Chernoff bound.

Spectral sparsification

A graph $H$ is a $(1+\epsilon)$-cut approximator of $G$ if $(1-\epsilon) w_H(S,S) \leq w_G(S) \leq (1+\epsilon) w_H(S,S)$ for all $S \subseteq V$.

where $w_H(S,S)$ is the total weight of the edges crossing $S$.

Benczúr and Karger proved that for any $G$, there exists a $(1+\epsilon)$-cut approximator with $O\left(\frac{n\log n}{\epsilon^2}\right)$ edges.

Today we will prove a spectral generalization of this result.

Spectral approximator

We say a graph $H$ is a $(1+\epsilon)$-spectral approximator of $G$ if $(1-\epsilon)L_G \preceq L_H \preceq (1+\epsilon)L_G$, or equivalently

$$(1-\epsilon)x^TL_Gx \leq x^TL_Hx \leq (1+\epsilon)x^TL_Gx \quad \forall x \in \mathbb{R}^n$$

where $n$ is the number of vertices.

Claim. A $(1+\epsilon)$-spectral approximator is a $(1+\epsilon)$-cut approximator.

Proof. For $S \subseteq V$, let $x_S \in \mathbb{R}^n$ be the vector with $x_S(i) = 1$ if $i \in S$ and zero otherwise.

Then, $x_S^TL_Gx_S = \sum_{i,j \in E} w_{ij} (x_S(i)-x_S(j))^2 = w_H(S,S)$ and similarly $x_S^TL_Hx_S = w_H(S,S)$.

Since $H$ is a $(1+\epsilon)$-spectral approximator of $G$, we have

$$(1-\epsilon)x_S^TL_Gx_S \leq x_S^TL_Hx_S \leq (1+\epsilon)x_S^TL_Gx_S \quad \forall S \subseteq V \text{ and thus } (1-\epsilon)w_H(S,S) \leq w_H(S,S) \leq (1+\epsilon)w_H(S,S) \text{ for all $S$.}$$

The following theorem by Spielman and Srivastava thus generalizes the result of Benczúr and Karger.

Theorem. Any graph has a $(1+\epsilon)$-spectral approximator with $O(n\log n/\epsilon^2)$ edges.

Reduction

The spectral sparsification result can be reduced to the following purely linear algebraic result.

Theorem. Suppose $v_1,\ldots,v_m \in \mathbb{R}^n$ are given with $\sum_{i=1}^m v_i v_i^T = I_n$.

There exist scalars $s_1,\ldots,s_m$ with at most $O(n\log n/\epsilon^2)$ non-zero such that

$$(1-\epsilon)I_n \preceq \sum_{i=1}^m s_i v_i v_i^T \preceq (1+\epsilon)I_n.$$
We sketch the proof of the reduction of the spectral sparsification result to the above result.

The idea is to apply a linear transformation so that the Laplacian matrix becomes the identity matrix. Let $M$ be a positive semidefinite matrix with eigen-decomposition $M = \sum_{i} \lambda_i u_i u_i^T$. The pseudo-inverse of $M$ is defined as $M^+ = \sum_{i: \lambda_i > 0} \frac{1}{\lambda_i} u_i u_i^T$, and $M^{1/2} = \sum_{i: \lambda_i > 0} \sqrt{\lambda_i} u_i u_i^T$.

Given $L_G = \sum_{e \in E} L_e = \sum_{e \in E} b_e b_e^T$, we consider $I = L_G^{1/2} L_G L_G^{1/2} = \sum_{e \in E} (L_G^{1/2} b_e)(b_e L_G^{1/2}) = \sum_{e \in E} v_e v_e^T$, where we define $v_e = L_G^{1/2} b_e \forall e \in E$.

Apply the above theorem gives us $S_e$ with at most $O(\log n / \epsilon^2)$ non-zeros so that

$$(1 - \epsilon) I \preceq \sum_{e \in S_e} v_e v_e^T \preceq (1 + \epsilon) I.$$ 

Now, multiplying $L_G^{1/2}$ on the left and right gives us $(1 - \epsilon) L_G \leq \sum_{e \in S_e} b_e b_e^T \preceq (1 + \epsilon) L_G$, so by scaling the weight of each edge by a factor of $S_e$, we get our spectral sparsifier.

The above "proof" is not precise as we are dealing with the pseudo-inverse (not the inverse), but the missing details are rather routine and is not the important part of the proof, and so omitted.

**Sampling algorithm**

Now, our focus is to prove the linear algebraic result, by random sampling.

First, we get some intuition about the condition $\sum_{i=1}^n v_i v_i^T = I_n$, the isotropy condition in L04.

When $m = n$, then $v_1, \ldots, v_n$ must be an orthonormal basis.

When $m > n$, we can also think of it as an "overcomplete" basis, as we can write any $x \in \mathbb{R}^n$ as

$x = I_n x = \left( \sum_{i=1}^m v_i v_i^T \right) x = \sum_{i=1}^m <x, v_i>v_i$.

Similarly, for any unit vector $y \in \mathbb{R}^m$, we have $1 = y^T y = y^T I_n y = y^T \left( \sum_{i=1}^m v_i v_i^T \right) y = \sum_{i=1}^m y_i^2 v_i v_i^T y = \sum_{i=1}^m <v_i, y>^2$.

Intuitively, the vectors are "evenly spread out", so that the projection of any direction $y$ to these vectors are the same.

**Idea**: Given $\sum_{i=1}^n v_i v_i^T = I_n$, we would like to find a small submatrix of vectors $S \subseteq \{1, \ldots, n\}$ and some scaling factors so that $\sum_{i \in S} g_i v_i v_i^T \approx I_n$.

So, the subsets should still be "evenly spread out", with contributions in each direction about the same. As in the graph sparsification case, uniform sampling won't work. For example, if some $v_j$ has $\|v_j\|_2 = 1$, then we must include $v_j$ in the solution, as otherwise that direction will not be covered in
The solution and so it won’t be a spectral sparsifier. The analogy in the graph sparsification result is that a cut edge must be included in any sparsifier.

So, as in the graph sparsification case, we need to do non-uniform sampling (if we do random sampling). The idea is similar: for longer vectors, the sampling probability is higher; for shorter vectors, we can be more aggressive in setting the sampling probability to be smaller, and when we choose them, we reweight the vector so that it has the correct expected value.

More concretely, we sample each vector \( v_i \) with probability \( \|v_i\|_2^2 \), and if it is chosen, we set the scalar \( s_i = \frac{1}{\|v_i\|_2^2} \), so that \( E[s_i \langle v_i, v_i \rangle] = \frac{\langle v_i, v_i \rangle}{\|v_i\|_2^2} \), \( Pr(v_i \text{ is chosen}) = \frac{\langle v_i, v_i \rangle}{\|v_i\|_2^2} \|v_i\|_2^2 = \langle v_i, v_i \rangle \).

**Algorithm**

The actual algorithm is basically the same as described above, but we need to repeat this experiment \( C = \Theta(\log n) \) times and take the average, so that we can prove concentration.

1. Initially, \( F = \emptyset \), \( s = 0 \), \( C = \frac{6 \log n}{\epsilon^2} \).

2. For \( 1 \leq t \leq C \) do

   For each \( e \in E \), with probability \( p_e = \|v_e\|_2^2 \), update \( F \leftarrow F \cup \{e\} \) and \( s_e \leftarrow s_e + \frac{1}{Cp_e} \).

3. Return \( \sum_{i=1}^n s_i \langle v_i, v_i \rangle \) as our spectral approximator.

**Analysis**

There are two steps in the analysis.

One is to show that there are \( O(n\log n/\epsilon^2) \) non-zero scalars, i.e. \( |F| = O(n\log n/\epsilon^2) \).

Another is to show that the returned solution is a \((1\pm\epsilon)\)-spectral sparsifier.

We first bound the number of non-zero scalars.

**Claim** With probability at least 0.9, \( |F| = O(n\log n/\epsilon^2) \).

**Proof** The expected value is \( E[|F|] = \sum_{i=1}^n Pr(\text{vector } i \text{ is in } F) = \sum_{i=1}^n (1 - (1 - p_i) C) = \sum_{i=1}^n (1 - (1 - C p_e)) = C \sum_{i=1}^n p_i, \)

which can also be seen by a union bound.

Note that

\[
\sum_{i=1}^n p_i = \sum_{i=1}^n \|v_i\|_2^2 = \sum_{i=1}^n \langle v_i, v_i \rangle = \sum_{i=1}^n \text{tr}(v_i v_i^T) = \text{tr}(\sum_{i=1}^n v_i v_i^T) = \text{tr}(v v^T) = n,
\]

where

\[
\text{tr}(A) = \sum_{i,j} A_{ij}
\]

and we use the fact that \( \text{tr}(AB) = \text{tr}(BA) \) (or directly check that \( vv^T = \text{tr}(v v^T) \)).

Therefore, \( E[|F|] = C \sum_{i=1}^n p_i = C n = 6n\log n/\epsilon^2 \). The result follows from Markov’s inequality.
Matrix Chernoff bound

There is an elegant generalization of the Chernoff-Hoeffding bound to the matrix setting.

**Theorem (Tropp)** Let \( X_1, \ldots, X_k \) be independent, \( n \times n \) symmetric matrices with \( 0 \leq X_i \leq R I \).

Let \( \mu_{\min} \leq \sum_{i=1}^k E[X_i] \leq \mu_{\max} I \). For any \( \varepsilon \in (0,1) \),

\[
\begin{align*}
\Pr \left( \lambda_{\max} \left( \sum_{i=1}^k X_i \right) \geq (1+\varepsilon) \mu_{\max} \right) & \leq n e^{-\frac{\varepsilon^2 \mu_{\max}^2}{2R}} \\
\Pr \left( \lambda_{\min} \left( \sum_{i=1}^k X_i \right) \leq (1-\varepsilon) \mu_{\min} \right) & \leq n e^{-\frac{\varepsilon^2 \mu_{\min}^2}{2R}}
\end{align*}
\]

Note that it is almost an exact analog of the Chernoff-Hoeffding bound in the scalar case, by using the maximum eigenvalue and minimum eigenvalue to measure the "size" of a matrix.

It says that if we consider the sum of independent random matrices, where each matrix is not too "big/influential", then the sum is concentrated around the expectation in terms of the eigenvalues.

**Concentration**

The proof that our solution is a \((1+\varepsilon)\)-spectral sparsifier is a direct application of the matrix Chernoff bound.

The random variables are \( X_{i,t} = \begin{cases} \frac{v_i v_i^T}{c \pi_i} & \text{with probability } \pi_i = \|v_i\|_2^2, \text{ for vector } i \text{ in iteration } t. \\ 0 & \text{otherwise} \end{cases} \)

Note that the output of the algorithm is \( S = \sum_{t=1}^\infty \sum_{i=1}^k X_{i,t} \).

As discussed before, \( E[S] = \sum_{t=1}^\infty \sum_{i=1}^k E[X_{i,t}] = \sum_{t=1}^\infty \sum_{i=1}^k \frac{v_i v_i^T}{c \pi_i} \cdot \pi_i = \sum_{t=1}^\infty \sum_{i=1}^k \frac{v_i v_i^T}{c} = \sum_{i=1}^n \frac{v_i v_i^T}{c} = I \).

So, the expected value is correct, with \( \mu_{\max} = \mu_{\min} = 2 \) in this problem.

To apply the matrix Chernoff bound, we just need to find a bound for \( R \) so that \( X_{i,t} \leq R I \).

Note that \( X_{i,t} = \frac{v_i v_i^T}{c \pi_i} = \frac{v_i v_i^T}{\|v_i\|_2^2} = \frac{1}{c} \left( \frac{v_i}{\|v_i\|_2} \right) \left( \frac{v_i}{\|v_i\|_2} \right)^T \). This is a rank one matrix of a unit vector, and so the maximum eigenvalue is just \( \frac{1}{c} \) (with the only eigenvector being \( \frac{v_i}{\|v_i\|_2} \)). So, \( R = \frac{1}{c} \).

By Tropp's theorem, we get \( \Pr \left( \lambda_{\max} (S) \geq (1+\varepsilon) \right) \leq n e^{-\frac{\varepsilon^2 \mu_{\max}^2 \cdot C}{2}} = n e^{-\frac{\varepsilon^2 \cdot \log n}{2}} = \frac{1}{n^\varepsilon} \) as \( C = \log n / \varepsilon^2 \).

The lower tail follows similarly.

So, with probability at least \( 1 - \frac{1}{n} \), we have \( \lambda_{\max}(S) \geq (1+\varepsilon) \) and \( \lambda_{\min}(S) \leq (1-\varepsilon) \), and so \((1-\varepsilon)I \leq S \leq (1+\varepsilon)I\), proving that our solution \( S \) is a \((1+\varepsilon)\)-spectral sparsifier of \( I_n \).

By a union bound, we know that a \((1+\varepsilon)\)-spectral sparsifier with \( O(n \log n / \varepsilon^2) \) edges exist, and indeed the random sampling algorithm will succeed with high probability, proving the theorem.
Discussions

There are a few things to discuss about.

1. By considering this linear algebraic generalization of cut sparsification, we have a clean and arguably simple proof of the result of Benczur and Karger.

A subsequent amazing result by Battison, Spielman and Srivastava proves that every graph has a \((1+\epsilon)\)-spectral sparsifier with \(O(n/\epsilon^2)\) edges, which is best possible.

We don’t know of an alternative (combinatorial) proof to achieve the same bound even for cut approximator (a special case).

This linear algebraic perspective seems to be the correct way to look at the problem.

2. The sampling probability \(p_e\) is directly proportional to the effective resistance of the edge \(e\).

Recall that \(p_e = \|v(e)\|^2 = \|L^+_G v(e)\|^2 = \|b^e L^+_G b_e\|^2\). Let \(e = uv\).

Note that \(L^+_G b_e\) is a solution \(x\) to \(L_G x = b\), which is the potential vector \(\Phi\) of the electrical flow problem when one unit of electrical flow is sent from \(u\) to \(v\).

Then, \(b^e L^+_G b_e = b^e \overline{\Phi} = \Phi(v) - \Phi(u)\) is just the definition of \(\text{Res}(u,v)\).

So, the sampling algorithm works by sampling each edge with probability proportional to its effective resistance, a somewhat surprising application of this concept.

3. There is a nearly linear time algorithm to estimate the effective resistances of all edges.

The main tools are a near linear time algorithm to solve a Laplacian system of equations (another breakthrough result by Spielman and Teng), and also dimension reduction.

So, we have a near linear time (randomized) algorithm for constructing spectral sparsifiers.

4. The analysis of the random sampling algorithm is tight.

In a complete graph, the effective resistance of every edge is the same, as the graph is symmetric.

So, the random sampling algorithm on a complete graph is just the uniform sampling algorithm.

And by a “coupon collector” argument that it won’t work with \(O(n \log n / \epsilon^2)\) edges.

References

- Spielman and Srivastava. Graph sparsification by effective resistance. 2008.

- Lecture notes by Nick Harvey on Tropp's inequality and Spectral Sparsification.