CSC 2421H : Graphs Matrices and Optimization Lecture 8 : 12 Nov 2018

Graph Sparsifiers

Lecturer: Sushant Sachdeva *Scribe: Kevan Hollbach*

Recall the following theorem from last class.

Theorem 0.1. *Let $0 < \epsilon < 1$, and let $\mathbf{X}_1, \ldots, \mathbf{X}_t \in \mathbb{R}^{n \times n}$ be symmetric independant random matrices such that $\forall i : 0 \preceq \mathbf{X}_i \preceq R\mathbb{I}$. If $\mu_{\min}\mathbb{I} \preceq \mathbb{E}\sum \mathbf{X}_i \preceq \mu_{\max}\mathbb{I}$ then*

$$\Pr\left[\sum \mathbf{X}_i \succeq (1+\epsilon)\mu_{\max}\right] \leq n \exp\left(-\frac{\epsilon^2 \mu_{\max}}{3R}\right) \text{ and}$$

$$\Pr\left[\sum \mathbf{X}_i \preceq (1-\epsilon)\mu_{\min}\right] \leq n \exp\left(-\frac{\epsilon^2 \mu_{\min}}{2R}\right).$$

In particular, this holds for $\mu_{\min} = \lambda_{\min}(\mathbb{E}\sum \mathbf{X}_i)$ and $\mu_{\max} = \lambda_{\max}(\mathbb{E}\sum \mathbf{X}_i)$, the tightest bounds on the spectrum of the expected sum.

1 Spectral Graph Sparsifiers

We introduce the notion of spectral sparsifiers, which are a generalization of spectral expanders. Given a graph $G = (V, E)$, a graph $H = (V, E')$ is an ϵ-spectral sparsifier of G if

$$(1-\epsilon)\mathbf{L}_G \preceq \mathbf{L}_H \preceq (1+\epsilon)\mathbf{L}_G.$$

In this case we will write $H \approx_\epsilon G$ and say that H spectrally approximates G. We will fudge $(1-\epsilon)$ versus $\frac{1}{1+\epsilon}$ in the definition of sparsifier, since for small ϵ, these quantities are close. Note that the definition does not require that $E_H \subseteq E_G$, but in our constructions this will indeed be the case.

An immediate consequence of this definition is that if $H \approx_\epsilon G$, then $\forall S \subseteq V$,

$$(1-\epsilon)\, w_G(E_G(S, \overline{S})) \leq w_H(E_H(S, \overline{S})) \leq (1+\epsilon)\, w_G(E_G(S, \overline{S})),$$

as seen via the following quadratic form:

$$\mathbb{1}_S^\top \mathbf{L}_H \mathbb{1}_S = \sum_{(u,v)\in E_H} w_H(u,v)\,(\mathbb{1}_S(u) - \mathbb{1}_S(v))^2 = w_H(E_H(S, \overline{S})).$$

What this says is that for any cut of the vertices V into two sets, the weight of the edges crossing the cut is approximately the same in H and G.

Consider the randomized construction of an expander on an even number of vertices seen in the previous lecture.

EXPANDER1:

for $i \leftarrow 1, \ldots, t = \Theta(\frac{1}{\epsilon^2} \log n)$

Add an independent random matching of V to H, scaled by $\frac{n-1}{t}$

Scaling the edge weights of H by $\frac{n-1}{t}$ ensures that vertices of H have degree exactly $n-1$, and that H is in expectation equal to the clique,

$$\mathbb{E}\mathbf{L}_H = \mathbf{L}_{K_n}.$$

Indeed, as seen last class, H is concentrated about its mean, and with high probability, H spectrally approximates the clique. That is, with high probability,

$$\mathbf{L}_H \approx_\epsilon \mathbf{L}_{K_n}.$$

The following is an alternative randomized construction of a spectral expander.

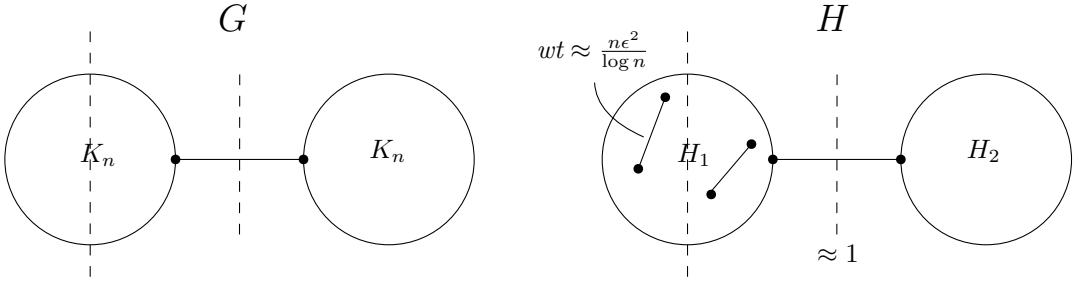
EXPANDER2:

for $i \leftarrow 1, \ldots, t = \Theta(\frac{1}{\epsilon^2} n \log n)$

Add 1 independent random edge to H, with scaling $\frac{n}{2}\frac{(n-1)}{t} \approx \frac{n\epsilon^2}{\log n}$

The above algorithm yields an ϵ-spectral expander with high probability, and a similar proof using matrix concentration bounds would show that taking $t \in \Omega(\frac{1}{\epsilon^2} n \log n)$ suffices. Notice however that H is unlikely to be a regular graph; its vertices will most likely have slightly different weighted degrees. It is worth noting that in both of the above algorithms, an edge (u, v) may be "added" to H more than once, depending on the random choices made. If an edge is added multiple times, the copies are combined into a single edge with weight equal to their sum. It is therefore possible that H is a weighted graph, even if the input graph G is "unweighted" (i.e. all edge weights are 1).

This second algorithm inspires our construction of spectral sparsifiers for graphs other than K_n. What if we took the same approach, but we restricted ourselves to choosing random edges from the edge set of some target graph G — would this then yield a sparsifier for G? We would of course have to scale the sampled edges according to their weight in G so that we output G in expectation. It turns out that this naïve approach would work, but in general we would have to take a very large number of samples, t, to get a sparsifier with high probability. This would somewhat defeat the purpose of getting a sparsifier, as in most applications we wish for a graph H that both spectrally approximates G and has much fewer edges (i.e. H is actually sparser than G); ideally something like $|E_H| \in O(n \log n)$ even when $|E_G| \in \Theta(n^2)$. The following example illustrates the issue with the naïve approach, and suggests how we might fix it.



In the above example, G is the (unweighted) dumbbell graph, two cliques connected by a single edge. Note that G is dense; $|E_G| \in \Theta(n^2)$. Any graph $H \approx_\epsilon G$ for which $E_H \subseteq E_G$ must have an edge between the two special vertices, with weight approximately 1. This is because, as noted before, any cut of H must have approximately the same weight as the corresponding cut in G. By similar reasoning, we can see that the two subgraphs H_1 and H_2 of H should more-or-less behave as ϵ-spectral expanders so as to mimic the structure of the corresponding subgraphs in G, each a copy of K_n. If we wish for H to be sparse and still have $H \approx_\epsilon G$ w.h.p., then in our randomized construction we must sample the edges of K_n each with very low probability, and scale the sampled edges' weights at approximately $\frac{n\epsilon^2}{\log n}$, as in EXPANDER2.

All this is to say that depending on the graph G, certain edges (e.g. the "special edge" connecting the cliques) must be sampled with higher probability; intuitively this is because they are more important to the structure of the graph. This suggests the idea of assigning "importance socres" to edges, and sampling them with probability proporional to their importance. We will also have to scale down the weights of the sampled edges by their importance scores so that, in expectation, we output the original graph G. This approach is due to Spielman and Srivastava [1], journal version [2]. For a detailed treatment of spectral sparsifiers, see [3]. (See also the versions on arXiv.)

2 Sparsifier Algorithm

SPARSIFIER(G, ϵ):

Construct a probabililty distribution $\{p_e\}_{e \in E_G}$ over the edges of G,
edge probabilities proportional to "importance scores" (defined later)
for $i \leftarrow 1, \ldots, t = \Theta(\frac{1}{\epsilon^2} n \log n)$
  Sample an independent edge e from $\{p_e\}$
  Scale it by $w'_e := \frac{w_e}{t\, p_e}$
  Add it to H

We use the notation e_i to mean the i^{th} edge sampled, a random variable. Our first observation is that the expected output of the algorithm is the original graph.

$$\mathbb{E}\mathbf{L}_H = \mathbb{E}\left[\sum_{i=1}^{t} w'_{e_i}\mathbf{L}_{e_i}\right] = t \cdot \mathbb{E}[w'_{e_1}\mathbf{L}_{e_1}] = t \sum_{e \in E_G} p_e w'_e \mathbf{L}_e = \sum_{e \in E_G} w_e \mathbf{L}_e = \mathbf{L}_G,$$

where $\mathbf{L}_{(u,v)} = (\mathbb{1}_u - \mathbb{1}_v)(\mathbb{1}_u - \mathbb{1}_v)^\top$. This is simply a consequence of the definition of w', and does not depend on the settings we give to t or $\{p_e\}$ (as long as $p_e > 0$, $\forall e \in E_G$).

What remains to be shown is that with high probability,

$$(1-\epsilon)\mathbf{L}_G \preceq \mathbf{L}_H \preceq (1+\epsilon)\mathbf{L}_G.$$

We will derive settings of t and $\{p_e\}$ that suffice, but in the meantime notice that we have a useful tool for proving statements such as the above: matrix Chernoff bounds, of the form

$$(1-\epsilon)\mu_{\min}\mathbb{I} \preceq \sum \mathbf{X}_i \preceq (1+\epsilon)\mu_{\max}\mathbb{I}.$$

Unfortunately however, it is not enough to apply Theorem 0.1 to the samples $\mathbf{X}_i = w'_{e_i}\mathbf{L}_{e_i}$.

First we must normalize our samples by the pseudoinverse of $\mathbf{L}_G$ — if $\psi_1, \ldots, \psi_n$ are orthonormal eigenvectors of $\mathbf{L}_G$ with corresponding eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$, then this is the matrix

$$\mathbf{L}_G^+ := \sum_{i\,:\,\lambda_i > 0} \frac{1}{\lambda_i} \psi_i \psi_i^\top = \sum_{i=2}^{n} \frac{1}{\lambda_i} \psi_i \psi_i^\top,$$

if we assume that G is connected. Consider the matrix

$$\mathbf{L}_G^{+/2} := (\mathbf{L}_G^+)^{\frac{1}{2}} = \sum_{i=2}^{n} \frac{1}{\sqrt{\lambda_i}} \psi_i \psi_i^\top$$

and, noting that it has the same kernel as $\mathbf{L}_G$ and $\mathbf{L}_H$, symmetrically multiply the statement of $\mathbf{L}_H \approx_\epsilon \mathbf{L}_G$ by it:

$$\begin{aligned} (1-\epsilon)\mathbf{L}_G \preceq & \mathbf{L}_H \preceq (1+\epsilon)\mathbf{L}_G \\ \iff (1-\epsilon)\Pi \preceq & \mathbf{L}_G^{+/2} \mathbf{L}_H \mathbf{L}_G^{+/2} \preceq (1+\epsilon)\Pi, \end{aligned}$$

where $\Pi := \mathbf{L}_G^{+/2} \mathbf{L}_G \mathbf{L}_G^{+/2} = \sum_{i=2}^{n} \psi_i \psi_i^\top = \mathbb{I} - \psi_1 \psi_1^\top = \mathbb{I} - \frac{1}{n} \mathbb{1}\mathbb{1}^\top$ is the matrix that performs projection onto the subspace orthogonal to the kernel of $\mathbf{L}_G$.

Now, to get a statement of the form $(1-\epsilon)\mathbb{I} \preceq \sum \mathbf{X}_i \preceq (1+\epsilon)\mathbb{I}$, we pad each of our t samples by $\frac{1}{t}\frac{1}{n}\mathbb{1}\mathbb{1}^T$, to fill in the kernel of Π.

$$\begin{aligned} \mathbf{X}_i &:= w'_{e_i} \mathbf{L}_G^{+/2} \mathbf{L}_{e_i} \mathbf{L}_G^{+/2} + \frac{1}{t}\frac{1}{n}\mathbb{1}\mathbb{1}^T \\ \sum_{i=1}^{t} \mathbf{X}_i &= \mathbf{L}_G^{+/2} \mathbf{L}_H \mathbf{L}_G^{+/2} + \frac{1}{n}\mathbb{1}\mathbb{1}^\top \\ \mathbb{E} \sum \mathbf{X}_i &= \mathbb{I} \end{aligned}$$

Noting that the samples are PSD, we can apply Chernoff to $\sum \mathbf{X}_i$ with $\mu_{\min} = \mu_{\max} = 1$ to get

$$\mathbf{Pr}[(1-\epsilon)\mathbb{I} \preceq \sum \mathbf{X}_i \preceq (1+\epsilon)\mathbb{I}] \geq 1 - 2n \exp(\frac{-\epsilon^2}{3R}).$$

It remains to define the distribution $\{p_e\}$, and to show that the samples are small in spectral norm. Our bound on the size of the samples, R, will depend on t, and it turns out that taking $t = \Theta(\frac{1}{\epsilon^2} n \log n)$ will suffice for the above Chernoff bound to hold with high probability.

We introduce the notion of spectral norm of a matrix,

$$\|\mathbf{A}\| := \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|},$$

and note that for symmetric $\mathbf{A} \succeq 0$,

$$\|\mathbf{A}\| = \lambda_{\max}(\mathbf{A}).$$

With this notation we can write

$$R = \max_i \max_{e_i \xleftarrow{\$} \{p_e\}} \|\mathbf{X}_i\| = \max_{e_1 \xleftarrow{\$} \{p_e\}} \|\mathbf{X}_1\|$$

(where the second max is over all choices of randomness), to represent the tightest bound for which

$$\forall i \,:\, 0 \preceq \mathbf{X}_i \preceq R\,\mathbb{I}.$$

That is,

$$\begin{aligned} R &:= \max_{e\in E_G} \left\| w'_e \mathbf{L}_G^{+/2}\mathbf{L}_e\mathbf{L}_G^{+/2} + \frac{1}{t}\frac{1}{n}\mathbb{1}\mathbb{1}^\top \right\| \\ &\geq \max\left\{ \left(\max_{e\in E_G} w'_e \left\| \mathbf{L}_G^{+/2}\mathbf{L}_e\mathbf{L}_G^{+/2} \right\| \right), \frac{1}{t}\left\| \frac{1}{n}\mathbb{1}\mathbb{1}^\top \right\| \right\} \\ &\geq \max_{e\in E_G} \frac{w_e}{t\,p_e} \left\| \mathbf{L}_G^{+/2}\mathbf{L}_e\mathbf{L}_G^{+/2} \right\| \end{aligned}$$

since $\mathbb{1}$ is in the kernel of $\mathbf{L}_G^{+/2}\mathbf{L}_e\mathbf{L}_G^{+/2}$, and therefore, $\forall e \in E_G$,

$$p_e \geq \frac{w_e}{tR}\left\| \mathbf{L}_G^{+/2}\mathbf{L}_e\mathbf{L}_G^{+/2} \right\|.$$

Let us set

$$p_e = \frac{w_e}{tR}\left\| \mathbf{L}_G^{+/2}\mathbf{L}_e\mathbf{L}_G^{+/2} \right\|,$$

and note that we wish to have $\sum_e p_e = 1$. This is equivalent to

$$\begin{aligned} tR &= \sum_e w_e \lambda_{\max}\left(\mathbf{L}_G^{+/2}\mathbf{L}_e\mathbf{L}_G^{+/2} \right) \\ &= \sum_{e=(u,v)} w_e \lambda_{\max}\left(\mathbf{L}_G^{+/2}(\mathbb{1}_u - \mathbb{1}_v)(\mathbb{1}_u - \mathbb{1}_v)^\top \mathbf{L}_G^{+/2} \right) \\ &= \sum_e w_e \operatorname{Tr}\left(\mathbf{L}_G^{+/2}(\mathbb{1}_u - \mathbb{1}_v)(\mathbb{1}_u - \mathbb{1}_v)^\top \mathbf{L}_G^{+/2} \right) \quad \text{(since the matrix is rank 1)} \\ &= \operatorname{Tr}\left(\mathbf{L}_G^{+/2}(\sum_e w_e \mathbf{L}_e)\mathbf{L}_G^{+/2} \right) \\ &= \operatorname{Tr}\left(\mathbf{L}_G^{+/2}\mathbf{L}_G\mathbf{L}_G^{+/2} \right) \\ &= \operatorname{Tr}(\Pi) = n-1. \end{aligned}$$

Therefore we define the distribution $\{p_e\}$ as, $\forall e$,

$$p_e := \frac{1}{n-1}\, w_e \left\| \mathbf{L}_G^{+/2}\mathbf{L}_e\mathbf{L}_G^{+/2} \right\|.$$

Picking $t := \frac{6}{\epsilon^2}(n-1)\log n$ gives us

$$\mathbf{Pr}[(1-\epsilon)\mathbb{I} \preceq \sum \mathbf{X}_i \preceq (1+\epsilon)\mathbb{I}] \geq 1 - \frac{2}{n}.$$

Thus with probability at least $1 - \frac{2}{n}$,

$$(1-\epsilon)\mathbb{I} \preceq \mathbf{L}_G^{+/2}\mathbf{L}_H\mathbf{L}_G^{+/2} + \frac{1}{n}\mathbb{1}\mathbb{1}^\top \preceq (1+\epsilon)\mathbb{I}.$$

Multiplying through symmetrically by $\mathbf{L}_G^{\frac{1}{2}}$, we get

$$(1-\epsilon)\mathbf{L}_G \preceq \mathbf{L}_H \preceq (1+\epsilon)\mathbf{L}_G.$$

That is, H is an ϵ-spectral sparsifier of G.

Theorem 2.1. *For any G, and any $0 < \epsilon < 1$, the graph H produced by* SPARSIFIER(G, ϵ) *is an ϵ-spectral sparsifier of G.*

2.1 An aside about probabilities / leverage scores

The scalars we used to define the probabilities $\{p_e\}$ have other applications as well, and have a special connection to electrical networks.

$$\begin{aligned}
& w_e \left\| \mathbf{L}_G^{+/2} \mathbf{L}_e \mathbf{L}_G^{+/2} \right\| \\
&= w_e \operatorname{Tr}\left(\mathbf{L}_G^{+/2} \mathbf{L}_e \mathbf{L}_G^{+/2} \right) \quad \text{(since the matrix is rank 1)} \\
&= w_e \operatorname{Tr}\left(\mathbf{L}_G^{+/2} (\mathbb{1}_u - \mathbb{1}_v)(\mathbb{1}_u - \mathbb{1}_v)^\top \mathbf{L}_G^{+/2} \right) \\
&= w_e \operatorname{Tr}\left((\mathbb{1}_u - \mathbb{1}_v)^\top \mathbf{L}_G^{+} (\mathbb{1}_u - \mathbb{1}_v) \right) \quad \text{(since trace is cyclic)} \\
&= w_e (\mathbb{1}_u - \mathbb{1}_v)^\top \mathbf{L}_G^{+} (\mathbb{1}_u - \mathbb{1}_v) \\
&= w_e R_{\mathit{eff}}(u, v),
\end{aligned}$$

where as seen in a previous lecture, $\mathbf{L}_G^{+}(\mathbb{1}_u - \mathbb{1}_v)$ is a vector of voltages and $R_{\mathit{eff}}(u,v) := (\mathbb{1}_u - \mathbb{1}_v)^\top \mathbf{L}_G^{+}(\mathbb{1}_u - \mathbb{1}_v)$ is the potential difference between u and v, if we were to send 1 unit of current from u to v.

The quantity $w_e R_{\mathit{eff}}(e)$ is often called the leverage score of edge e:

$$\textit{Lev-score}(e) := w_e R_{\mathit{eff}}(e).$$

2.2 Motivation for the next lecture

It is worth noting that a naïve, exact computation of $\mathbf{L}_G^{+}$ would take time $O(n^3)$, and thus calculating the probabilities $\{p_e\}$ would take time $O(n^3 + m) = O(n^3)$. This is often infeasible in practice, and we will see next lecture that we can bypass this $O(n^3)$ computation by instead computing approximate solutions to $Lx = b$.

References

[1] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pages 563–568, 2008.

[2] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM J. Comput.*, 40(6):1913–1926, 2011.

[3] Daniel A. Spielman and Shang-Hua Teng. Spectral sparsification of graphs. *SIAM J. Comput.*, 40(4):981–1025, 2011.