CS 270 Combinatorial Algorithms and Data Structures, Spring 2015

Lecture 26: Spectral sparsification

We study how to construct a spectral sparsifier by random sampling using effective resistance, and then discuss how to do a fast implementation and the proofs of matrix concentration results.

Spectral approximation

Recall that a graph H is an ε-cut approximator of G if $(1-\varepsilon)\,w(\delta_G(S)) \le w(\delta_H(S)) \le (1+\varepsilon)\,w(\delta_G(S))$, for all $S \subseteq V$, where $w(\delta_G(S))$ is the total weight of the edges crossing S.

We mentioned that for any graph G, there is a ε-cut approximator H with $O(n\log n/\varepsilon^2)$ edges.

Today we study a spectral generalization of this notion.

We say a graph H is an ε-spectral approximator of G if $(1-\varepsilon)L_G \preceq L_H \preceq (1+\varepsilon)L_G$, where L_G is the weighted Laplacian matrix of G.

Or equivalently, $(1-\varepsilon)\,x^T L_G x \le x^T L_H x \le (1+\varepsilon)\,x^T L_G x$ for all $x \in \mathbb{R}^n$, where n is the number of vertices.

Claim If H is an ε-spectral approximator of G, then H is an ε-cut approximator of G.

proof Let $S \subseteq V$ and $x_S \in \mathbb{R}^n$ be characteristic vector such that $x_S(i)=1$ if $i \in S$ and zero otherwise.

Since H is an ε-spectral approximator of G, we have $(1-\varepsilon)\,x_S^T L_G x_S \le x_S^T L_H x_S \le (1+\varepsilon)\,x_S^T L_G x_S$.

Note that $x^T L_G x = \sum_{ij \in E} w_{ij}(x_i - x_j)^2$, and thus $x_S^T L_G x_S = w(\delta_G(S))$.

Therefore, the spectral approximation implies that $(1-\varepsilon)\,w(\delta_G(S)) \le w(\delta_H(S)) \le (1+\varepsilon)\,w(\delta_G(S))$ $\forall S \subseteq V$. $\square$

The main theorem today is by Spielman and Srivastava, which is a generalization of Benczur and result about cut approximator

Theorem For any graph G and $\varepsilon > 0$, there is an ε-spectral approximator H with $O\left(\frac{n\log n}{\varepsilon^2}\right)$ edges.

Random Sampling

Like the proof of cut approximators, the proof of spectral approximators is also by random sampling.

Without loss of generality, we assume that G is unweighted.

Recall that $L_G = \sum_{ij \in E} L_{ij}$, where $L_{ij} = (\chi_i - \chi_j)(\chi_i - \chi_j)^T$ is the Laplacian matrix of edge ij.

So, L_G is a sum of m (simple) matrices.

We would like to construct a spectral approximator by picking a subset of edges and reweight them.

Sampling algorithm

The framework is very simple.

Suppose we have a probability distribution p over the edges of G and we want to pick k edges.

Initially, $w_e = 0$ for all edges $e \in E$.

For $1 \le i \le k$, pick a random edge e according to the probability distribution p.

Update $w_e = w_e + \frac{1}{kp_e}$.

Let H be the resulting weighted graph with at most k positive weight edges.

This is the algorithm.

We haven't specified what is k and what is p_e. It will turn out that p_e is proportional to the effective resistance and $k = O(n \log n / \varepsilon^2)$ would be enough.

Proof outline

First, observe that we set the weight in a way such that $E[L_H] = L_G$.

Let e_i be the i-th edge we picked and $Z_i = \frac{1}{kp_{e_i}} L_{e_i}$ be its weighted Laplacian.

Then $E[Z_i] = \sum_{e \in E} \frac{1}{kp_e} L_e \cdot \Pr(e \text{ is picked}) = \sum_{e \in E} \frac{1}{kp_e} L_e \cdot p_e = \sum_{e \in E} \frac{L_e}{k} = \frac{1}{k} L_G$.

Therefore, $E[L_H] = E[\sum_{i=1}^{k} Z_i] = \sum_{i=1}^{k} E[Z_i] = \sum_{i=1}^{k} \frac{L_G}{k} = L_G$.

To prove that H is a good spectral approximator, we would like to show that if k is large enough, then H is "concentrated" around its expectation.

There are different matrix concentration results and we use the following one by Ahlswede and Winter.

Theorem Let Z be a random $n \times n$ real symmetric PSD matrix. Suppose $Z \preccurlyeq R \cdot E[Z]$ for some $R \ge 1$.

Let $Z_1, Z_2, \ldots, Z_k$ be independent copies of Z. For any $\varepsilon \in (0,1)$, we have

$$\Pr\left[(1-\varepsilon) E[Z] \preccurlyeq \frac{1}{k} \sum_{i=1}^{k} Z_i \preccurlyeq (1+\varepsilon) E[Z] \right] \ge 1 - 2n \exp\left(\frac{-\varepsilon^2 k}{4R}\right).$$

We assume the theorem for now and will discuss the proof later.

For intuition, think of $E[Z] = I$ (we will eventually reduce to this case). Then, the theorem says that when we pick a random matrix with the expectation that "every direction is balanced" ($E[Z]=I$), if furthermore that "no outcome is very influential in some direction" ($Z \preccurlyeq R \cdot I$ for small R), then once we add them together "every direction is almost balanced".

This is in the same spirit as Chernoff bound in the scalar case.

In our case, $E[Z] = \frac{1}{k} L_G$ and $\sum_{i=1}^{k} Z_i = L_H$, and so the theorem is exactly what we want.

To bound k, it remains to set p_e in such a way that $Z_i = \frac{1}{k p_e} L_e \preceq \frac{R}{k} L_G$ for a small R.

That is, we need to choose p_e such that $L_e \preceq p_e R L_G$ for some small R.

Effective resistance: intuition

We can get some intuition about how to set p_e from what we did for cut sparsifiers.

If e is a cut edge, then we would like to set p_e to be large to make sure that the graph is not connected.

If e is inside a highly connected subgraph, then we can set p_e to be small.

It is proved that setting p_e proportional to the edge-connectivity of e (max-flow value) would work to construct a cut-sparsifier.

For spectral sparsification, we care about the relation between $v^T L_e v$ and $v^T L_G v$.

If there is a vector v such that $v^T L_e v \approx v^T L_G v$, then we should set p_e to be large to satisfy $L_e \preceq p_e \cdot R \cdot L_G$.

Informally speaking, these edges have large influence on the quadratic form, and we should keep them to preserve all quadratic form.

For example, if $e = ij$ is a cut edge and we don't choose it, then $v^T L_H v$ would be very different from $v^T L_G v$ for those v that set $v(i) = 1$ and $v(j) = 0$.

On the other hand, if $e = ij$ is an edge with many parallel edges, then whenever $v^T L_e v$ is large, $v^T L_G v$ is much larger, and so we can safely set p_e to be small in this case.

More generally, if $e = ij$ is an edge where there are many disjoint short paths between them, then whenever $v^T L_e v$ is large, then $v^T L_G v$ is also much larger and so can set p_e small.

The above discussion should lead to the conclusion that if the effective resistance between i and j is small, then we can set p_e to be small. It is because whenever $v^T L_e v = (v_i - v_j)^2$ is large, i.e. the voltage difference between v_i and v_j large, then (as effective resistance small) the energy of the flow becomes much larger, and thus $v^T L_e v \ll v^T L_G v$ and so it is okay to set p_e small.

(E.g. if $R_{eff}(i,j) = \frac{1}{10}$, if $v^T L_e v = (v_i - v_j)^2 = 1$, then $v^T L_G v = 10$.)

Also, it should be clear that sampling by edge-connectivity would not work, if edge $e = ij$ is connected by many edge-disjoint but long paths, then $v^T L_e v$ is roughly the same as $v^T L_G v$, and we should set p_e large.

So, we should set p_e to be proportional to the effective resistance, as edges of high effective resistance are important in preserving the energy of electric flow (the quadratic form).

The precise relation should be $L_e \preceq R_{eff}(e) \cdot L_G$. Is it a proof without equation?

Effective resistance: proof

The above intuition just came up when I am writing the notes, and so it is likely to have mistakes.

The proof that I know of is purely algebraic.

Recall that we need to choose p_e such that $L_e \preceq p_e R L_G$ for some small R.

In general, suppose we would like to find the smallest α such that $A \preceq \alpha B$ when $A, B \succeq 0$.

First, assume that B is invertible.

We need to check that $x^T A x \le \alpha x^T B x \iff \frac{x^T A x}{x^T B x} \le \alpha \iff \frac{y^T B^{-\frac{1}{2}} A B^{-\frac{1}{2}} y}{y^T y} \le \alpha$ where $y = B^{\frac{1}{2}} x$.

$$\iff \lambda_{max}(B^{-\frac{1}{2}} A B^{-\frac{1}{2}}) \le \alpha.$$

To bound $\lambda_{max}(B^{-\frac{1}{2}} A B^{-\frac{1}{2}})$, notice that since $A, B \succeq 0$, we have $B^{-\frac{1}{2}} A B^{-\frac{1}{2}} \succeq 0$, and thus

$\lambda_{max}(B^{-\frac{1}{2}} A B^{-\frac{1}{2}}) \le Tr(B^{-\frac{1}{2}} A B^{-\frac{1}{2}})$ as trace = sum of eigenvalues and all eigenvalues are nonnegative.

Now, for a moment, set $A = L_e$ and $B = L_G$,

then $\alpha \le Tr(L_G^{+/2} L_e L_G^{+/2}) = Tr(L_e L_G^{+}) = Tr((\chi_i - \chi_j)(\chi_i - \chi_j)^T L_G^{+}) = (\chi_i - \chi_j) L_G^{+} (\chi_i - \chi_j) = R_{eff}(i,j)$.

We get $L_{ij} \preceq R_{eff}(i,j) \cdot L_G$.

The proof seems not okay as L_G is not invertible, but it is actually okay.

The above proof is okay when nullspace$(B) \subseteq$ nullspace(A).

When $x \in$ nullspace$(B) \subseteq$ nullspace(A), then $x^T A x = x^T B x = 0$ and so the inequality holds trivially.

So, we only need to restrict our attention to $x \perp$ nullspace(A), and thus $x \perp$ nullspace(B).

Each such x can be written as $B^{+/2} y$ for some y, where B^{+} is the pseudo inverse of B, and $B^{+/2}$ is the square root of B^{+}.

Therefore, if we have bounded $\frac{y^T B^{+/2} A B^{+/2} y}{y^T y} \le \alpha$ for all y, we have bounded $\frac{x^T A x}{x^T B x} \le \alpha$ for those x.

and $B^{1/2}$ is the square root of B.

Therefore, if we have bounded $\frac{y^T B^{+/2} A B^{+/2} y}{y^T y} \leq \alpha$ for all y, we have bounded $\frac{x^T A x}{x^T B x} \leq \alpha$ for those x.

In our case, it is clear that nullspace $(L_G) \subseteq$ nullspace (L_e), and so we have the following.

<u>Lemma</u> $L_{ij} \preceq R_{eff}(i,j) \cdot L_G$.

Okay, recall that we want to choose p_e such that $L_e \preceq p_e \cdot R \cdot L_G$.

By the above lemma, we should set $p_e \sim R_{eff}(e)$.

We just need to compute $\sum_{e \in E} R_{eff}(e)$ and set $p_e = \frac{R_{eff}(e)}{\sum_e R_{eff}(e)}$ so that it is a probability distribution.

$$\sum_{ij \in E} R_{eff}(i,j) = \sum_{i \sim j} (\chi_i - \chi_j)^T L_G^+ (\chi_i - \chi_j) = \sum_{i \sim j} Tr\left((\chi_i - \chi_j)(\chi_i - \chi_j)^T L_G^+ \right) = \sum_{i \sim j} Tr\left(L_e L_G^+ \right)$$

$$= Tr\left(\left(\sum_{i \sim j} L_e \right) L_G^+ \right) = Tr\left(L_G L_G^+ \right).$$

Note that $L_G L_G^+ = \sum_{i \geq 2}^{n} u_i u_i^T$ where u_i are the eigenvectors of L_G, and thus there are $n-1$ eigenvalues of 1 and 1 eigenvalue of zero, and so $Tr(L_G L_G^+) =$ sum of eigenvalues $= n-1$.

<u>Lemma</u> $\sum_{e \in E} R_{eff}(e) = n-1$.

This is an important fact, as it says that there cannot be too many important edges.

Therefore, we can set $p_e = \frac{R_{eff}(e)}{n-1}$ and thus $L_e \preceq p_e \cdot (n-1) \cdot L_G$ and we have $R = n-1$.

Now, using Ahlswede-Winter, the failure probability that L_H is not an ε-approximator is at most

$$2n \exp\left(\frac{-\varepsilon^2 k}{4R} \right) = 2n \exp\left(-\frac{\varepsilon^2 k}{4(n-1)} \right).$$

Setting $k = O(n \log n / \varepsilon^2)$, this is inverse polynomial in n and we have proved the main theorem.

<u>Fast approximation</u>

To implement the algorithm, one needs to compute the effective resistance for every edge.

To compute effective resistance, one needs to solve $Lv = (\chi_s - \chi_t)$ and then get $v(s) - v(t)$.

There is a fast algorithm to compute $Lx = b$ in near linear time, which we will study next time.

Even with that, a direct implementation may still take $\tilde{O}(m^2)$ time.

There is a nice trick to get a good approximation much quicker, using the idea of dimension reduction.

First, we write $R_{eff}(i,j) = (\chi_i - \chi_j)^T L_G^+ (\chi_i - \chi_j) = (\chi_i - \chi_j)^T L_G^+ L_G L_G^+ (\chi_i - \chi_j) = (\chi_i - \chi_j)^T L_G^+ B^T B L_G^+ (\chi_i - \chi_j)$

First, we write $R_{eff}(i,j) = (\chi_i-\chi_j)^T L_G^+ (\chi_i-\chi_j) = (\chi_i-\chi_j)^T L_G^+ L_G L_G^+ (\chi_i-\chi_j) = (\chi_i-\chi_j)^T L_G^+ B^T B L_G^+ (\chi_i-\chi_j)$

$= \| B L_G^+ (\chi_i-\chi_j) \|_2^2$ where B is the $m\times n$ edge-vertex incidence matrix.

So, we care about the length of at most n^2 vectors in dimension n.

A well-known result shows that one can reduce the dimension to $O(\log n)$ without changing the lengths by much.

<u>Theorem</u> Given fixed vectors $u_1, u_2, \ldots, u_n \in \mathbb{R}^m$ and $\varepsilon > 0$, let $Q_{k\times m}$ be a random $\pm\frac{1}{\sqrt{k}}$ matrix with $k \geq 24\log n/\varepsilon^2$. Then, with probability $1-\frac{1}{n}$, we have for all pairs $i,j \leq n$

$$(1-\varepsilon)\|u_i-u_j\|_2^2 \leq \|Qu_i - Qu_j\|_2^2 \leq (1+\varepsilon)\|u_i-u_j\|_2^2$$

Unfortunately we won't do the proof. It is similar to the proof of estimating frequency moment in data streaming, and the argument is based on Chernoff-type bound.

This theorem is useful everywhere.

We are going to apply the dimension-reduction theorem for the vectors $BL_G^+\chi_i$.

For this, we will compute $Z = QBL_G^+$ efficiently and store this $O(\log n)\times m$ matrix.

Then, whenever we want to compute $\|QBL_G^+(\chi_i-\chi_j)\|_2^2 = \|Z(\chi_i-\chi_j)\|_2^2$, we just need to use two columns of Z, and can be done in $O(\log n)$ time since each column is of dimension $O(\log n)$, and so the total time after Z is computed is $\tilde{O}(m)$.

It remains to show how to compute Z in $\tilde{O}(m)$ time using a fast Laplacian solver.

First, we compute QB, which can be done in $O(km) = \tilde{O}(m)$ time since B has only $2m$ nonzeros.

Then, the i-th row of Z is just equal to the i-th row of QB times L_G^+.

Thus, it is of the form $L_G^+ y$ for some y, which can be solved by $L_G x = y$ in $\tilde{O}(m)$ time.

Therefore, the total time to compute Z is $\tilde{O}(m)$.

Spielman and Srivastava showed that these approximate effective resistances are enough for the purpose of constructing spectral sparsifiers, and we omit the details.

<u>Matrix Concentration</u> (optional)

We try to prove Ahlswede-Winter inequality.

The proof structure is similar to the proof of Chernoff bound, generalized to the matrix setting.

Let $X_1, \ldots, X_k$ be random $n\times n$ matrix, independent and symmetric.

Consider the partial sum $S_j = \sum_{i=1}^{j} X_i$.

We want to bound the probability that $S_k \preceq tI$.

Like Chernoff, this is equivalent to $e^{\lambda S_k} \preceq e^{\lambda t I}$, where we consider the matrix exponentials.

Suppose $e^{\lambda S_k} \not\preceq e^{\lambda t I}$. Then $Tr(e^{\lambda S_k}) \geq \lambda_{max}(e^{\lambda S_k}) \geq e^{\lambda t}$ where the first inequality holds as $e^{\lambda S_k} \succeq 0$.

So, $Pr(S_k \not\preceq tI) = Pr(e^{\lambda S_k} \not\preceq e^{\lambda t I}) \leq Pr(Tr(e^{\lambda S_k}) \geq e^{\lambda t}) \leq E[Tr(e^{\lambda S_k})]/e^{\lambda t}$ by Markov.

Therefore, we just need to bound $E[Tr(e^{\lambda S_k})]$ using that X_i are independent.

$$
\begin{aligned}
E[Tr(e^{\lambda S_k})] &= E[Tr(e^{\lambda X_k + \lambda S_{k-1}})] \\
&\leq E[Tr(e^{\lambda X_k} e^{\lambda S_{k-1}})] && (\text{Golden-Thompson } Tr(e^{A+B}) \leq Tr(e^A \cdot e^B)) \\
&= E_{X_1,\dots,X_{k-1}}[E_{X_k}[Tr(e^{\lambda X_k} e^{\lambda S_{k-1}})]] && (\text{indepence of } X_i) \\
&= E_{X_1,\dots,X_{k-1}}[Tr(E_{X_k}[e^{\lambda X_k} e^{\lambda S_{k-1}}])] && (\text{trace is linear}) \\
&= E_{X_1,\dots,X_{k-1}}[Tr(E_{X_k}[e^{\lambda X_k}] \cdot e^{\lambda S_{k-1}})] && (\text{independence of } X_i) \\
&\leq E_{X_1,\dots,X_{k-1}}[\|E_{X_k}[e^{\lambda X_k}]\| \cdot Tr(e^{\lambda S_{k-1}})] && (\text{if } A \succeq 0, \text{ then } Tr(A \cdot B) \leq \|B\| Tr(A)) \\
&= \|E_{X_k}[e^{\lambda X_k}]\| \cdot E_{X_1,\dots,X_{k-1}}[Tr(e^{\lambda S_{k-1}})].
\end{aligned}
$$

By induction, we get $E[Tr(e^{\lambda S_k})] \leq \prod_{i=1}^{k} \|E_{X_i}[e^{\lambda X_i}]\| \cdot Tr(e^{\lambda 0}) = n \cdot \prod_{i=1}^{k} \|E[e^{\lambda X_i}]\|$,

since $e^{\lambda 0} = I$ and $Tr[I] = n$.

So, we have $Pr(S_k \not\preceq tI) \leq d e^{-\lambda t} \prod_{i=1}^{k} \|E[e^{\lambda X_i}]\|$.

Apply the same argument to bound the probability that $S_k \not\succeq -tI$, and we get

$$Pr(\|S_k\| > t) \leq d e^{-\lambda t} \left(\prod_{i=1}^{k} \|E[e^{\lambda X_i}]\| + \prod_{i=1}^{k} \|E[e^{-\lambda X_i}]\| \right). \quad (*)$$

Now we prove Ahlswede-Winter in the special case when $E[Z] = I$, and later reduce the general case to this case.

<u>Theorem</u> Let Z be a random $n \times n$ real symmetric PSD matrix. Suppose $E[Z] = I$ and $\|Z\| \leq R$.

Let $Z_1, Z_2, \dots, Z_k$ be independent copies of Z. For any $\varepsilon \in (0,1)$, we have

$$Pr\left[(1-\varepsilon) E[Z] \preceq \frac{1}{k} \sum_{i=1}^{k} Z_i \preceq (1+\varepsilon) E[Z] \right] \geq 1 - 2n \exp\left(\frac{-\varepsilon^2 k}{4R}\right).$$

<u>proof</u> We set $X_i = (Z_i - E[Z_i])/R$ so that $E[X_i] = 0$ and $\|X_i\| \leq 1$.

Now, since $1 + x \leq e^x \;\; \forall x \in \mathbb{R}$ and $e^x \leq 1 + x + x^2 \;\; \forall x \in [-1, +1]$.

As all eigenvalues of X_i are in $[-1,+1]$, we have the inequalities (recall L17)

$$e^{\lambda X_i} \preccurlyeq I + \lambda X_i + \lambda^2 X_i^2 \quad \text{for} \quad \lambda \in [0,1]$$

It follows that $E[e^{\lambda X_i}] \preccurlyeq E[I + \lambda X_i + \lambda^2 X_i^2] = I + \lambda^2 E[X_i^2] \preccurlyeq e^{\lambda^2 E[X_i^2]}$.

Note that $E[X_i^2] = \frac{1}{R^2} E[(Z_i - E[Z_i])^2] = \frac{1}{R^2}(E[Z_i^2] - E[Z_i]^2)$

$$\preccurlyeq \frac{1}{R^2} E[Z_i^2] \preccurlyeq \frac{1}{R^2} E[\|Z_i\| Z_i] \preccurlyeq \frac{R}{R^2} E[Z_i] = \frac{1}{R} E[Z_i].$$

Therefore, $\|E[e^{\lambda X_i}]\| \leq \|e^{\lambda^2 E[X_i^2]}\| \leq e^{\lambda^2/R}$.

So, plugging into $(*)$, we have

$$\Pr\left(\left\| \sum_{i=1}^{k} \frac{1}{R}(Z_i - E[Z_i]) \right\| > t \right) \leq 2n \cdot e^{-\lambda t} \prod_{i=1}^{k} e^{\lambda^2/R} = 2n \cdot \exp\left(-\lambda t + \frac{k\lambda^2}{R}\right).$$

Setting $t = k\varepsilon/R$ and $\lambda = \varepsilon/2$, we have

$$\Pr\left(\left\| \frac{1}{R}\sum_{i=1}^{k} Z_i - \frac{k}{R} E[Z_i] \right\| > \frac{k\varepsilon}{R} \right) \leq 2n \exp(-k\varepsilon^2/4R). \quad \square$$

Finally, to reduce the general case to the special case, let $U := E[Z]$.

We apply the above theorem with $Z' = U^{+/2} E[Z] U^{+/2}$ and $Z_i' = U^{+/2} Z_i U^{+/2}$.

It is easy to check that it works when U is invertible, and it is also true for singular U using the pseudo-inverse.

References

The main theorem is in the paper "Graph sparsification by effective resistance" by Spielman and Srivastava. I follow the presentation of Nick Harvey (2012 course notes) and James Lee (2015 course notes), especially the former on matrix concentration.

Later, Batson, Spielman, and Srivastava proved that there is an ε-spectral sparsifier of $O(n/\varepsilon^2)$ edges, showing the surprising power of the spectral method.

Their algorithm is deterministic but takes $O(n^4)$ time.

Recently it was improved to $\tilde{O}(n^2)$ by using matrix multiplicative update by Allen-Zhu, Liao, Orrechia.

It is an interesting open problem whether it can be constructed in $\tilde{O}(m)$ time.