CS 798 - Algorithmic Spectral Graph Theory, Fall 2015, Waterloo

Lecture 17: 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 a $(1\pm\varepsilon)$-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 $(1\pm\varepsilon)$-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 a $(1\pm\varepsilon)$-spectral approximator of G if $(1-\varepsilon)L_G \preccurlyeq L_H \preccurlyeq (1+\varepsilon)L_G$, where L_G is the weighted Laplacian matrix of G.

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

Claim If H is a $(1\pm\varepsilon)$-spectral approximator of G, then H is a $(1\pm\varepsilon)$-cut approximator of G.

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

Since H is an ε-spectral approximator of G, we have $(1-\varepsilon)\chi_S^TL_G\chi_S \le \chi_S^TL_H\chi_S \le (1+\varepsilon)\chi_S^TL_G\chi_S$.

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

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 Karger result about cut approximator.

Theorem For any graph G and $\varepsilon>0$, there is a $(1\pm\varepsilon)$-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 $w_e = w_e + \frac{1}{k p_e}$.

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 $Z_i = \frac{1}{k p_{e_i}} L_{e_i}$ be its weighted Laplacian.

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

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 many of them together "every direction is almost balanced".

This is in the same spirit as Chernoff bound in the scalar case, with the absolute value replaced by the norm of the matrix (equal to maximum eigenvalue).

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 $Z_i = \frac{1}{k p_e} L_e \preccurlyeq \frac{R}{k} L_G$ for a small R.

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

Effective resistance

To bound $L_e \preccurlyeq \alpha L_G$, we bound $v^T L_e v / v^T L_G v$ for any $v \in \mathbb{R}^n$.

Note that $v^T L_e v = (v_i - v_j)^2$ for $e = ij$. Without loss we assume $v_i = 1$ and $v_j = 0$, and thus $v^T L_e v = 1$.

For the same v, by the result about effective conductance in L15, we have

$$v^T L_G v \geq \min_{v_i = 1, v_j = 0} \sum_{a \sim b} (v_a - v_b)^2 = C_{eff}(e) = \frac{1}{R_{eff}(e)}$$

Therefore, we have $v^T L_e v \leq R_{eff}(e) v^T L_G v \;\; \forall v \in \mathbb{R}^n$, and we conclude that $L_e \preccurlyeq R_{eff}(e) L_G$.

There is also an algebraic proof of this fact.

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

First, assume that B is invertible.

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

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

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

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

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

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

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

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 $x \perp \text{nullspace}(A)$, and thus $x \perp \text{nullspace}(B)$.

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 $\frac{y^T B^{+/2} A B^{+/2} y}{y^T y} \leq \alpha$ 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 $\text{nullspace}(L_G) \subseteq \text{nullspace}(L_e)$, 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 $v(s) - v(t)$.

In L13, we have seen a near-linear time algorithm to solve $Lx = b$ approximately.

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

$= \| B L_G^+ (\chi_i-\chi_j)\|_2^2$ 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 $u_1, u_2, \dots, u_n \in \mathbb{R}^m$ 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, which is based on some Chernoff-type argument.

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 $\hat{O}(m)$ 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 Chernoff bound</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.

Matrix exponential

First, we need the analog of exponential in the matrix setting.

Given a matrix A, we define $e^A := \sum_{i\geq 0} \frac{A^i}{i!}$ as the matrix exponential of A. Some properties of e^A:

- Regardless of A, $\exp(A)$ is positive semidefinite, because $\exp(A) = \exp(\frac{1}{2}A)\exp(\frac{1}{2}A)$ (exercise).
- Suppose A is real symmetric. Then $A = VDV^T$ as the eigendecomposition. Then $A^i = VD^iV^T$ since $VV^T = I$.

  Then $e^A = Ve^DV^T = \sum_{i=1}^{n} e^{\lambda_i} v_i v_i^T$ where λ_i is the i-th eigenvalue and v_i is the corresponding eigenvector.

  So, suppose we have an inequality that holds for all λ_i, then it also holds for the matrix. For example, if we know that $(1-\eta)^x \leq 1-\eta x$ for all $x \in [0,1]$, then we can conclude that $(1-\eta)^A \preceq (I-\eta A)$ for $0 \preceq A \preceq I$ as all eigenvalues of A are in $[0,1]$. To see it, first rewrite $(1-\eta)^x$ as $e^{-\eta' x}$ where $\eta' = -\ln(1-\eta)$ and $(1-\eta)^A = e^{-\eta' A}$. Then $I - \eta A - e^{-\eta' A} = VV^T - \eta VDV^T - Ve^{-\eta' D}V^T = V(I - \eta D - e^{-\eta' D})V^T$, and this is PSD iff $1 - \eta x - e^{-\eta' x} \geq 0$ holds for all eigenvalues of A as $I - \eta D - e^{-\eta' D}$ is a diagonal matrix. So, the matrix exponential of a symmetric matrix behaves like the exponential of a number.

Chernoff-type argument

Now, we follow the same pattern of proving Chernoff bounds.

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 $S_k \preceq aI$.

Like Chernoff, this is equivalent to $e^{tS_k} \preceq e^{taI}$, where we consider the matrix exponentials.

Suppose $e^{tS_k} \npreceq e^{taI}$. Then $Tr(e^{tS_k}) \geq \lambda_{max}(e^{tS_k}) \geq e^{ta}$ where the first inequality holds as $e^{tS_k} \succeq 0$,

and so trace = sum of eigenvalues $\geq$ max eigenvalue (as eigenvalues are non-negative).

So, $Pr(S_k \npreceq aI) = Pr(e^{tS_k} \npreceq e^{taI}) \leq Pr(Tr(e^{tS_k}) \geq e^{ta}) \leq E[Tr(e^{tS_k})]/e^{ta}$ by Markov.

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

$$
\begin{aligned}
E[Tr(e^{tS_k})] &= E[Tr(e^{tX_k + tS_{k-1}})] \\
&\leq E[Tr(e^{tX_k} e^{tS_{k-1}})] && (\text{Golden-Thompson } Tr(e^{A+B}) \leq Tr(e^A \cdot e^B)\text{, see reference}) \\
&= E_{X_1,\ldots,X_{k-1}}[E_{X_k}[Tr(e^{tX_k} e^{tS_{k-1}})]] && (\text{independence of } X_i\text{: product distribution}) \\
&= E_{X_1,\ldots,X_{k-1}}[Tr(E_{X_k}[e^{tX_k} e^{tS_{k-1}}])] && (\text{trace is linear}) \\
&= E_{X_1,\ldots,X_{k-1}}[Tr(E_{X_k}[e^{tX_k}] \cdot e^{tS_{k-1}})] && (\text{independence of } X_i)
\end{aligned}
$$

$\le E_{X_1,\dots,X_{k-1}}\left[\, \| E_{X_k}[e^{tX_k}] \| \cdot Tr(e^{tS_{k-1}}) \right]$ (if $A \succeq 0$, then $Tr(A\cdot B) \le \|B\| Tr(A)$)

$= \| E_{X_k}[e^{tX_k}] \| \cdot E_{X_1,\dots,X_{k-1}}\left[Tr(e^{tS_{k-1}}) \right].$ exercise using eigen-decomposition

By induction, we get $E[\,Tr(e^{tS_k})] \le \prod_{i=1}^{k} \| E_{X_i}[e^{tX_i}] \| \cdot Tr(e^{t0}) = n \cdot \prod_{i=1}^{k} \| E[e^{tX_i}] \|,$

since $e^{t0} = I$ and $Tr[I] = n$.

So, we have $\Pr(S_k \not\preceq aI) \le d e^{-ta} \prod_{i=1}^{k} \|E[e^{tX_i}]\|.$

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

$$\Pr(\|S_k\| > a) \le d e^{-ta} \left(\prod_{i=1}^{k} \| E[e^{tX_i}] \| + \prod_{i=1}^{k} \| E[e^{-tX_i}] \| \right). \quad (*)$$

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

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

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

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

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

We would like to bound $E[e^{tX_i}]$ and apply $(*)$.

We have $1 + x \le e^x \ \forall x \in \mathbb{R}$ and $e^x \le 1 + x + x^2 \ \forall x \in [-1,+1]$ in the scalar world.

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

$e^{tX_i} \preceq I + tX_i + t^2 X_i^2$ for $t \in [0,1]$

It follows that $E[e^{tX_i}] \preceq E[I + tX_i + t^2 X_i^2] = I + t^2 E[X_i^2] \preceq e^{t^2 E[X_i^2]}$.

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

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

Therefore, $\| E[e^{tX_i}] \| \le \| e^{t^2 E[X_i^2]} \| \le e^{t^2/R}$.

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

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

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

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

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.

We will discuss this reduction again in the next lecture.

References

- Graph sparsification by effective resistance, by Spielman and Srivastava, 2012.
- Course notes on "sparse approximations", by Nick Harvey, 2012 (for matrix Chernoff bound).
- Notes by Harvey (http://www.cs.ubc.ca/~nickhar/Cargese2.pdf) on Tropp's inequality which is very similar to Chernoff.
- The Golden-Thompson inequality — historical aspects and random matrix applications, by Forrester, Thompson.