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 $(\pm \varepsilon)$-cut approximator of $G$ if

$$\left(1-\varepsilon\right) w(B_G(S)) \leq w(B_H(S)) \leq \left(1+\varepsilon\right) w(B_G(S)),$$

for all $S \subseteq V$, where $w(B_G(S))$ is the total weight of the edges crossing $S$.

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

Today we study a spectral generalization of this notion.

We say a graph $H$ is a $(\pm \varepsilon)$-spectral approximator of $G$ if $\left(1-\varepsilon\right) \mathbf{L}_G \succeq \mathbf{L}_H \succeq \left(1+\varepsilon\right) \mathbf{L}_G$, where

$\mathbf{L}_G$ is the weighted Laplacian matrix of $G$.

Or equivalently, $\left(1-\varepsilon\right) x^T \mathbf{L}_G x \leq x^T \mathbf{L}_H x \leq \left(1+\varepsilon\right) x^T \mathbf{L}_G x$ for all $x \in \mathbb{R}^n$, where $n$ is the number of vertices.

Claim. If $H$ is a $(\pm \varepsilon)$-spectral approximator of $G$, then $H$ is a $(\pm \varepsilon)$-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 $\varepsilon$-spectral approximator of $G$, we have $\left(1-\varepsilon\right) x_S^T \mathbf{L}_G x_S \leq x_S^T \mathbf{L}_H x_S \leq \left(1+\varepsilon\right) x_S^T \mathbf{L}_G x_S$.

Note that $x_S^T \mathbf{L}_G x_S = \sum_{ij \in E} w_{ij} (x_i - x_j)^2$, and thus $x_S^T \mathbf{L}_G x_S \leq w(B_G(S))$.

Therefore, the spectral approximation implies that $\left(1-\varepsilon\right) w(B_G(S)) \leq w(B_H(S)) \leq \left(1+\varepsilon\right) w(B_G(S))$ for all $S \subseteq V$.

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 $(\pm \varepsilon)$-spectral approximator $H$ with $O\left(n \log n / \varepsilon^2\right)$ 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 $\mathbf{L}_G = \sum_{ij \in E} \mathbf{L}_{ij}$, where $\mathbf{L}_{ij} = (x_i - x_j)(x_i - x_j)^T$ is the Laplacian matrix of edge $ij$.

So, $\mathbf{L}_G$ 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 $1 \leq i \leq k$, 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 $k = O(n \log n / \epsilon^2)$ 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 $Z_i = \frac{1}{k} p_e r_e$ be its weighted Laplacian.

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

Therefore, $E(L_H) = \sum_{i=1}^{k} E[Z_i] = \sum_{i=1}^{k} \frac{1}{k} L e = \frac{1}{k} L G$.

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 Altschuler and Witten.

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

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

$$Pr \left[ \left( (1 - e) E[Z] \right)^{ 1 \leq i \leq k } \leq (1 + e) E[Z] \right] \geq 1 - 2n \exp \left( -\frac{k e^2}{4k} \right).$$

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" and furthermore that "no outcome is very influential in some direction" ( $Z \leq R I$ 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, $E[Z] = \frac{1}{k}Lg$ and $\sum_{i=1}^{n} Z_i = LH$, and so the theorem is exactly what we want.

To bound $L$, it remains to set $p_e$ in such a way that $z_i = \frac{1}{R_p} Le \leq \frac{R}{L} Lg$ for a small $R$. That is, we need to choose $p_e$ such that $Le \leq p_e R Lg$ for some small $R$.

**Effective resistance**

To bound $Le \leq xLg$, we bound $v^T Le v/v^T Lg v$ for any $v \in \mathbb{R}^n$.

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

For the same $v$, by the result about effective conductance in (15), we have

$$v^T Le v \geq \min_{v \in \mathbb{R}^n} \sum_{a \in b} (v_a - v_b)^2 = \text{Con}(e) = \frac{1}{\text{Reff}(e)}$$

Therefore, we have $v^T Le v \leq \text{Reff}(e) v^T Lg v \forall v \in \mathbb{R}^n$, and we conclude that $Le \leq \text{Reff}(e) Lg$.

There is also an algebraic proof of this fact:

In general, suppose we would like to find the smallest $\alpha$ such that $A \leq B = 0$ when $A, B \in \mathbb{R}^n$.

First, assume that $B$ is invertible.

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

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

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

$$\lambda_{\max}(B^{-1} A^{-1}) \leq \text{tr}(B^{-1} A^{-1})$$ as trace is sum of eigenvalues and all eigenvalues are nonnegative.

Now, for a moment, set $A = Lg$ and $B = Lg$,

then $\alpha \leq \text{tr}(Lg^T Lg) = \text{tr}(Lg Lg^T) = \text{tr}((x_i, x_j) (x_i, x_j)^T Lg^T) = (x_i, x_j)^T Lg^T (x_i, x_j) = \text{Reff}(e)$.

We get $L_{ij} \leq \text{Reff}(e) Lg$.

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

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

When $x \in$ nullspace $(B) \leq$ nullspace $(A)$, then $x^T Ax = x^T Bx = 0$ and so the inequality holds trivially.
So, we only need to restrict our attention to \( \mathbf{x} \perp \) nullspace(\( \mathbf{A} \)), and thus \( \mathbf{x} \perp \) nullspace(\( \mathbf{B} \)).

Each such \( \mathbf{x} \) can be written as \( B^+y \) for some \( y \), where \( B^+ \) is the pseudo inverse of \( B \), and \( B^{1/2} \) is the square root of \( B^+ \).

Therefore, if we have bounded \( \mathbf{y}^T \mathbf{B}^{-1/2} \mathbf{A} \mathbf{B}^{1/2} \mathbf{y} \leq \alpha \) for all \( y \), we have bounded \( \mathbf{x}^T \mathbf{A} \mathbf{x} \leq \alpha \) for those \( x \).

In our case, it is clear that \text{nullspace}(\( \mathbf{L}_G \)) \( \subseteq \) \text{nullspace}(\( \mathbf{L}_e \)), and so we have the following.

**Lemma** \( L_{ij} \leq \text{Reff}(i,j) - \mathbf{L}_G \).

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 \( p_e \propto \text{Reff}(e) \).

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

\[
\sum_{i,j} \text{Reff}(i,j) = \sum_{i,j} \text{Tr} \left( \mathbf{L}_G \mathbf{x}_i \mathbf{x}_j^T \right) = \sum_{i,j} \text{Tr} \left( \mathbf{L}_G \mathbf{L}_G^+ \right) = \sum_{i,j} \text{Tr} \left( \mathbf{L}_G \mathbf{L}_G^+ \right).
\]

Note that \( \mathbf{L}_G \mathbf{L}_G^+ = \sum_{i,j} u_i u_j^T \) where \( u_i \) are the eigenvectors of \( \mathbf{L}_G \), and thus there are \( n-1 \) eigenvalues of 1 and 1 eigenvalue of zero, and so \( \text{Tr}(\mathbf{L}_G \mathbf{L}_G^+) = \text{sum of eigenvalues} = n-1 \).

**Lemma** \( \sum_{e \in E} \text{Reff}(e) = n-1 \).

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

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

Now, using Amini-Winter, the failure probability that \( L_H \) is not an \( \varepsilon \)-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 \( k = O(\varepsilon \log n / \varepsilon^2) \), this is inverse polynomial in \( n \) and we have proved the main theorem.

**Fast approximation**

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

To compute effective resistance, one needs to solve \( \mathbf{L} \mathbf{v} = (\mathbf{x}_3 - \mathbf{x}_4) \) and then get \( \mathbf{u}(s) \cdot \mathbf{v}(t) \).

In L13, we have seen a near-linear time algorithm to solve \( \mathbf{L} \mathbf{x} = \mathbf{b} \) 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 $\text{Rep}_{ij} = (X_i - X_j)^T L_g^+ (X_i - X_j) = (X_i - X_j)^T L_g^+ L_g + (X_i - X_j) = (X_i - X_j)^T L_g^+ B B^T L_g^+ (X_i - X_j)$

$= \| B L_g^+ (X_i - X_j) \|_2^2$ 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.

**Theorem** Given fixed vectors $u_1, u_2, \ldots, u_n \in \mathbb{R}^m$ and $\epsilon > 0$, let $Q_{km}$ be a random $\pm \frac{1}{\sqrt{k}}$ matrix with $k \geq 2\log n / \epsilon^2$. Then, with probability $1 - \frac{1}{m}$, we have for all pairs $i, j \in [n]$, $(1 - \epsilon) \| u_i - u_j \|_2^2 \leq \| Q u_i - Q u_j \|_2^2 \leq (1 + \epsilon) \| u_i - u_j \|_2^2$

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 $B L_g^+ X_i$.

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

Then, whenever we want to compute $\| Q B L_g^+ (X_i - X_j) \|_2^2 = \| Z (X_i - X_j) \|_2^2$, 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 $Q B$, which can be done in $O(mn) = \tilde{O}(m)$ time since $B$ has only $2m$ nonzeroes.

Then, the $i$-th row of $Z$ is just equal to the $i$-th row of $Q B$ 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.

---

**Matrix Chernoff bound** (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 $e^A := \sum_{n=0}^{\infty} \frac{A^n}{n!}$ 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 $A = V D V^T$ as the eigendecomposition. Then $A^2 = V D^2 V^T$ since $V V^T = I$.
  $e^A = V e^D V^T = \sum_{i=1}^{n} e^{\lambda_i} v_i v_i^T$ where $\lambda_i$ is the $i$-th eigenvalue and $v_i$ is the corresponding eigenvector.
So, suppose we have an inequality that holds for all $\lambda_i$, then it also holds for the matrix. For example, if we know that $(1-\eta)^x \leq 1-\eta \eta$ for all $x \in [0,1]$, then we can conclude that $(1-\eta)^A \leq (I-\eta A)$ for $0 \leq A \leq I$ as all eigenvalues of $A$ are in $[0,1]$. To see it, first rewrite $(1-\eta)^x$ as $e^{-\eta x}$ where $\eta \equiv -\lambda_n (1-\eta)$ and $(1-\eta)^A = e^{-\eta A}$. Then $I-\eta A = e^{-\eta A} - 1 = e^{-\eta A} - 1 = V (I-\eta D - \eta B) V^T$, and this is PSD iff $1-\eta x \leq \sum_{i=1}^{n} e^{-\eta x} \lambda_i$ holds for all eigenvalues of $A$ as $I-\eta A = e^{-\eta A}$ 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 x n matrix, independent and symmetric.

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

We want to bound the probability that $S_k \leq a I$.

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

Suppose $e^{t S_k} \leq e^{t a I}$. Then $\text{Tr}(e^{t S_k}) = \lambda_{\text{max}}(e^{t S_k}) \geq \lambda^a$ where the first inequality holds as $e^{t S_k} \geq 0$, and so trace $= \text{sum of eigenvalues} \geq \text{max eigenvalue}$ (as eigenvalues are non-negative).

So, $\Pr(S_k \leq a I) = \Pr(e^{t S_k} \leq e^{t a I}) \leq \Pr(\text{Tr}(e^{t S_k}) \geq e^{t a}) \leq e^{\frac{t}{2} \text{Var} \text{Tr}(X)}$ by Markov.

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

$E \left[ \text{Tr}(e^{t S_k}) \right] = E \left[ \text{Tr}(e^{t X_k + t S_{k-1}}) \right] \leq E \left[ \text{Tr}(e^{t X_k} e^{t S_{k-1}}) \right] \leq \text{(Golden-Thompson) Tr}(e^{t A + B}) \leq \text{Tr}(e^{A} e^{B})$, see reference.

$= E_{X_1, \ldots, X_{k-1}} \left[ E_{X_k} \left[ \text{Tr}(e^{t X_k} e^{t S_{k-1}}) \right] \right]$ (independence of $X_i$; product distribution).

$= E_{X_1, \ldots, X_{k-1}} \left[ \text{Tr} \left( E_{X_k} [e^{t X_k} e^{t S_{k-1}}] \right) \right]$ (trace is linear).

$= E_{X_1, \ldots, X_{k-1}} \left[ \text{Tr} \left( E_{X_k} [e^{t X_k}] e^{t S_{k-1}} \right) \right]$ (independence of $X_k$)
By induction, we get
\[ E[\text{Tr}(e^{tX_k})] \leq \prod_{i=1}^{k} E[e^{tX_i}]. \]
so that
\[ \text{Tr}(e^{tX}) = n \prod_{i=1}^{k} E[e^{tX_i}]. \]

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 \( \|Z\| \leq R \).

Let \( Z_1, Z_2, \ldots, Z_k \) be independent copies of \( Z \). For any \( \epsilon \in (0,1) \), we have
\[ \Pr\left( \epsilon \leq \text{Tr}(Z_i) \leq (1+\epsilon)I \right) \geq 1-2n \exp\left(-\frac{\epsilon^2 k}{4R}\right). \]

**Proof.** We set \( X_i = (Z_i - E[Z_i])/R \) so that \( E[X_i] = 0 \) and \( \|X_i\| \leq 1 \).

We would like to bound \( E[e^{tX_i}] \) and apply (1).

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

As all eigenvalues of \( X_i \) are in \([-1,1] \), we have the inequalities
\[ e^{tX_i} \leq I + tX_i + t^2X_i^2 \quad \text{for} \quad t \in [0,1]. \]

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

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

Therefore, \( \|E[e^{tX_i}]\| \leq \|e^{t^2 E[X_i^2]}\| \leq e^{t^2/R}. \)

So, plugging into (1), we have
\[ \Pr\left( \|Z_i - E[Z_i]\| > a \right) \leq 2n \exp\left(-\frac{\epsilon^2 k}{4R}\right). \]

Setting \( a = \frac{k\epsilon}{R} \) and \( t = \epsilon/2 \), we have
\[ \Pr\left( \|Z_i - E[Z_i]\| > \frac{k\epsilon}{R} \right) \leq 2n \exp\left(-\frac{k\epsilon^2}{4R}\right). \]

Finally, to reduce the general case to the special case, let \( U = E[Z] \).
We apply the above theorem with \( Z' = U^T E[Z] U^{1/2} \) and \( Z_i' = U^T Z_i U^{1/2} \). 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.