Advanced Graph Algorithms and Optimization Spring 2020

Random Matrix Concentration and Spectral Graph Sparsification

Rasmus Kyng, Scribe: Hongjie Chen *Lecture 8 — Wednesday, April 8th*

1 Matrix Concentration

Last time, we saw the Bernstein matrix concentration bound (Tropp 2011), i.e.,

Theorem 1.1. *Suppose $\boldsymbol{X}_1, \dots, \boldsymbol{X}_k \in \mathbb{R}^{n \times n}$ are independent, symmetric matrix-valued random variables. Assume each $\boldsymbol{X}_i$ is zero-mean, i.e. $\mathbb{E}\left[\boldsymbol{X}_i\right] = \boldsymbol{0}_{n \times n}$, and that $\|\boldsymbol{X}_i\| \leq R$ always. Let $\boldsymbol{X} = \sum_i \boldsymbol{X}_i$, and $\sigma^2 = \|\,Var\left[\boldsymbol{X}\right]\| = \|\sum_i \mathbb{E}\left[\boldsymbol{X}_i^2\right]\|$, then for $t > 0$*

$$\Pr[\|\boldsymbol{X}\| \geq t] \leq 2n \exp\left(\frac{-t^2}{2Rt + 4\sigma^2}\right).$$

In this section, we'll prove this theorem. But let's collect some useful tools for the proof first.

Definition 1.2 (trace)**.** The trace of a square matrix $\boldsymbol{A}$ is defined as

$$\operatorname{Tr}\left(\boldsymbol{A}\right) := \sum_i \boldsymbol{A}(i,i)$$

Claim 1.3 (cyclic property of trace)**.** $\operatorname{Tr}\left(\boldsymbol{A}\boldsymbol{B}\right) = \operatorname{Tr}\left(\boldsymbol{B}\boldsymbol{A}\right)$

Let S^n denote the set of all $n \times n$ real symmetric matrices, S_+^n the set of all $n \times n$ positive semidefinite matrices, and S_{++}^n the set of all $n \times n$ positive definite matrices. Their relation is clear, $S_{++}^n \subset S_+^n \subset S^n$. For any $\boldsymbol{A} \in S^n$ with eigenvalues $\lambda_1(\boldsymbol{A}) \leq \cdots \leq \lambda_n(\boldsymbol{A})$, by spectral decomposition theorem, $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$ where $\boldsymbol{\Lambda} = \operatorname{diag}_i\{\lambda_i(\boldsymbol{A})\}$ and $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{I}$, we'll use this property without specifying in the sequel.

Claim 1.4. *Given a symmetric and real matrix $\boldsymbol{A}$, $\operatorname{Tr}\left(\boldsymbol{A}\right) = \sum_i \lambda_i$, where $\{\lambda_i\}$ are eigenvalues of A.*

Proof.

$$\operatorname{Tr}\left(\boldsymbol{A}\right) = \operatorname{Tr}\left(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top\right) = \operatorname{Tr}\left(\boldsymbol{\Lambda}\underbrace{\boldsymbol{V}^\top\boldsymbol{V}}_{\boldsymbol{I}}\right) = \operatorname{Tr}\left(\boldsymbol{\Lambda}\right) = \sum_i \lambda_i.$$

□

1.1 Matrix Functions

Definition 1.5 (Matrix function)**.** Given a real-valued function $f : \mathbb{R} \to \mathbb{R}$, we extend it to a matrix function $f : S^n \to S^n$. For $\boldsymbol{A} \in S^n$ with spectral decomposition $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, let

$$f(\boldsymbol{A}) = \boldsymbol{V} \operatorname*{diag}_i \left\{f(\lambda_i)\right\} \boldsymbol{V}^\top.$$

c

Example. Recall that every PSD matrix $\boldsymbol{A}$ has a square root $\boldsymbol{A}^{1/2}$. If $f(x) = x^{1/2}$ for $x \in \mathbb{R}_+$, then $f(\boldsymbol{A}) = \boldsymbol{A}^{1/2}$ for $\boldsymbol{A} \in S^n_+$.

Example. If $f(x) = \exp(x)$ for $x \in \mathbb{R}$, then $f(\boldsymbol{A}) = \exp(\boldsymbol{A}) = \boldsymbol{V} \exp(\boldsymbol{\Lambda}) \boldsymbol{V}^\top$ for $\boldsymbol{A} \in S^n$. Note that $\exp(\boldsymbol{A})$ is positive definite for any $\boldsymbol{A} \in S^n$.

1.2 Monotonicity and Operator Monotonicity

Cosider a function $f : \mathcal{D} \to \mathcal{C}$. If we have a partial order $\leq_{\mathcal{D}}$ defined on $\mathcal{D}$ and a partial order $\leq_{\mathcal{C}}$ defined on $\mathcal{C}$, then we say that the function is monotone increasing (resp. decreasing) w.r.t. this pair of orderings if for all $d_1, d_2 \in \mathcal{D}$ s.t. $d_1 \leq_{\mathcal{D}} d_2$ we have $f(d_1) \leq_{\mathcal{C}} f(d_2)$ (resp. decreasing if $f(d_2) \leq_{\mathcal{C}} f(d_1)$).

Let's introduce some terminonology for important special cases of this idea. We say that a function $f : \mathcal{S} \to \mathbb{R}$, where $\mathcal{S} \subseteq S^n$, is monotone increasing if $\boldsymbol{A} \preceq \boldsymbol{B}$ implies $f(\boldsymbol{A}) \leq f(\boldsymbol{B})$.

Meanwhile, a function $f : \mathcal{S} \to \mathcal{T}$ where $\mathcal{S}, \mathcal{T} \subseteq S^n$ is said to be operator monotone increasing if $\boldsymbol{A} \preceq \boldsymbol{B}$ implies $f(\boldsymbol{A}) \preceq f(\boldsymbol{B})$.

Lemma 1.6. *Let $T \subseteq \mathbb{R}$. If the scalar function $f : T \to \mathbb{R}$ is monotone increasing, the matrix function $\boldsymbol{X} \mapsto \operatorname{Tr}(f(\boldsymbol{X}))$ is monotone increasing.*

Proof. From previous lectures, we know if $\boldsymbol{A} \preceq \boldsymbol{B}$ then $\lambda_i(\boldsymbol{A}) \leq \lambda_i(B)$ for all i. As $x \mapsto f(x)$ is monotone, then $\lambda_i(f(\boldsymbol{A})) \leq \lambda_i(f(B))$ for all i. By Claim 1.4, $\operatorname{Tr}(f(\boldsymbol{A})) \leq \operatorname{Tr}(f(\boldsymbol{B}))$. □

From this, and the fact that $x \mapsto \exp(x)$ is a monotone function on the reals, we get the following corollary.

Corollary 1.7. *If $\boldsymbol{A} \preceq \boldsymbol{B}$, then $\operatorname{Tr}(\exp(\boldsymbol{A})) \leq \operatorname{Tr}(\exp(\boldsymbol{B}))$, i.e. $\boldsymbol{X} \mapsto \operatorname{Tr}(\exp(\boldsymbol{X}))$ is monotone increasing.*

Lemma 1.8. *If $\boldsymbol{0} \prec \boldsymbol{A} \preceq \boldsymbol{B}$, then $\boldsymbol{B}^{-1} \preceq \boldsymbol{A}^{-1}$, i.e. $\boldsymbol{X} \mapsto \boldsymbol{X}^{-1}$ is operator monotone decreasing on S^n_{++}.*

You will prove the above lemma in this week's exercises.

Lemma 1.9. *If $0 \prec \boldsymbol{A} \preceq \boldsymbol{B}$, then $\log(\boldsymbol{A}) \preceq \log(\boldsymbol{B})$.*

To prove this lemma, we first recall an integral representation of the logarithm.

Lemma 1.10.

$$\log a = \int_0^\infty \left(\frac{1}{1+t} - \frac{1}{a+t} \right) \mathrm{d}t$$

Proof.

$$\begin{aligned}
\int_0^\infty \left(\frac{1}{1+t} - \frac{1}{a+t} \right) \mathrm{d}t &= \lim_{T \to \infty} \int_0^T \left(\frac{1}{1+t} - \frac{1}{a+t} \right) \mathrm{d}t \\
&= \lim_{T \to \infty} \left[\log(1+t) - \log(a+t) \right]_0^T
\end{aligned}$$

$$
\begin{aligned}
&= \log(a) + \lim_{T\to\infty} \log\left(\frac{1+T}{a+T}\right) \\
&= \log(a)
\end{aligned}
$$

□

Proof sketch of Lemma 1.9. Because all the matrices involved are diagonalized by the same orthogonal transformation, we can conclude from Lemma 1.10 that for a matrix $\boldsymbol{A} \succ \mathbf{0}$,

$$
\log(\boldsymbol{A}) = \int_0^\infty \left(\frac{1}{1+t}\boldsymbol{I} - (t\boldsymbol{I} + \boldsymbol{A})^{-1}\right) \mathrm{d}t
$$

This integration can be expressesd as the limit of a sum with positive coefficients, and from this we can show that is the integrand (the term inside the integration symbol) is operator monotone increasing in $\boldsymbol{A}$ by Lemma 1.8, the result of the integral, i.e. $\log(\boldsymbol{A})$ must also be operator monotone increasing. □

Lemma 1.11. *Let $T \subset \mathbb{R}$. If the scalar function $f : T \to \mathbb{R}$ is monotone, the matrix function $\boldsymbol{X} \mapsto \operatorname{Tr}(f(\boldsymbol{X}))$ is monotone.*

Remark 1.12. It is not always true that when $f : \mathbb{R} \to \mathbb{R}$ is monotone, $f : S^n \to S^n$ is operator monotone. For example, $\boldsymbol{X} \mapsto \boldsymbol{X}^2$ and $\boldsymbol{X} \mapsto \exp(\boldsymbol{X})$ are *not* operator monotone.

1.3 Some Useful Facts

Lemma 1.13. $\exp(\boldsymbol{A}) \preceq \boldsymbol{I} + \boldsymbol{A} + \boldsymbol{A}^2$ *for* $\|\boldsymbol{A}\| \leq 1$.

Proof.

$$
\begin{aligned}
\boldsymbol{I} + \boldsymbol{A} + \boldsymbol{A}^2 - \exp(\boldsymbol{A}) &= \boldsymbol{V}\boldsymbol{I}\boldsymbol{V}^\top + \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top + \boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}^\top - \boldsymbol{V}\exp(\boldsymbol{\Lambda})\boldsymbol{V}^\top \\
&= \boldsymbol{V}\left(\boldsymbol{I} + \boldsymbol{\Lambda} + \boldsymbol{\Lambda}^2 - \exp(\boldsymbol{\Lambda})\right)\boldsymbol{V}^\top \\
&= \boldsymbol{V} \operatorname*{diag}_i\{1 + \lambda_i + \lambda_i^2 - \exp(\lambda_i)\}\boldsymbol{V}^\top
\end{aligned}
$$

Recall $\exp(x) \leq 1 + x + x^2$ for all $|x| \leq 1$. Since $\|A\| \leq 1$ i.e. $|\lambda_i| \leq 1$ for all i, thus $1 + \lambda_i + \lambda_i^2 - \exp(\lambda_i) \geq 0$ for all i, meaning $\boldsymbol{I} + \boldsymbol{A} + \boldsymbol{A}^2 - \exp(\boldsymbol{A}) \succeq 0$. □

Lemma 1.14. $\log(\boldsymbol{I} + \boldsymbol{A}) \preceq \boldsymbol{A}$ *for* $A \succ -\boldsymbol{I}$.

Proof.

$$
\begin{aligned}
\boldsymbol{A} - \log(\boldsymbol{I} + \boldsymbol{A}) &= \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top - \boldsymbol{V}\log(\boldsymbol{\Lambda} + \boldsymbol{I})\boldsymbol{V}^\top \\
&= \boldsymbol{V}\left(\boldsymbol{\Lambda} - \log(\boldsymbol{\Lambda} + \boldsymbol{I})\right)\boldsymbol{V}^\top \\
&= \boldsymbol{V} \operatorname*{diag}_i\{\lambda_i - \log(1 + \lambda_i)\}\boldsymbol{V}^\top
\end{aligned}
$$

Recall $x \geq \log(1+x)$ for all $x > -1$. Since $\|A\| \succ -\boldsymbol{I}$ i.e. $\lambda_i > -1$ for all i, thus $\lambda_i - \log(1+\lambda_i) \geq 0$ for all i, meaning $\boldsymbol{A} - \log(\boldsymbol{I} + \boldsymbol{A}) \succeq 0$. □

Theorem 1.15 (Lieb)**.** *Let $f : S^n_{++} \to \Re$ be a matrix function given by*

$$f(\boldsymbol{A}) = \operatorname{Tr}\left(\exp\left(\boldsymbol{H} + \log(\boldsymbol{A})\right)\right)$$

for some $\boldsymbol{H} \in S^n$. Then $-f$ is convex (i.e. f is concave).

The Lieb's theorem will be crucial in our proof of Theorem 1.1, but it is also highly non-trivial and we will omit its proof here. The interested reader can find a proof in Chapter 8 of [T$^+$15].

Lemma 1.16 (Jensen's inequality)**.** $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$ *when f is convex;* $\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$ *when f is concave.*

1.4 Proof of Matrix Bernstein Concentration Bound

Now, we are ready to prove the Bernstein matrix concentration bound.

Proof of Theorem 1.1. For any $\boldsymbol{A} \in S^n$, its spectral norm $\|\boldsymbol{A}\| = \max\{|\lambda_n(\boldsymbol{A})|, |\lambda_1(\boldsymbol{A})|\} = \max\{\lambda_n(\boldsymbol{A}), -\lambda_1(\boldsymbol{A})\}$. Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of $\boldsymbol{X}$. Then,

$$\Pr[\|\boldsymbol{X}\| \geq t] = \Pr\left[(\lambda_n \geq t) \bigvee (-\lambda_1 \geq t)\right] \leq \Pr[\lambda_n \geq t] + \Pr[-\lambda_1 \geq t].$$

Let $\boldsymbol{Y} := \sum_i -\boldsymbol{X}_i$. It's easy to see that $-\lambda_n \leq \cdots \leq -\lambda_1$ are eigenvalues of $\boldsymbol{Y}$, implying $\lambda_n(\boldsymbol{Y}) = -\lambda_1(\boldsymbol{X})$. Since $\mathbb{E}[-\boldsymbol{X}_i] = \mathbb{E}[\boldsymbol{X}_i] = 0$ and $\|-\boldsymbol{X}_i\| = \|\boldsymbol{X}_i\| \leq R$ for all i, if we can bound $\Pr[\lambda_n(\boldsymbol{X}) \geq t]$, then applying to $\boldsymbol{Y}$, we can bound $\Pr[\lambda_n(\boldsymbol{Y}) \geq t]$. As

$$\Pr[-\lambda_1(\boldsymbol{X}) \geq t] = \Pr[\lambda_n(\boldsymbol{Y}) \geq t],$$

it suffices to bound $\Pr[\lambda_n \geq t]$.

For any $\theta > 0$, $\lambda_n \geq t \iff \exp(\theta\lambda_n) \geq \exp(\theta t)$ and $\operatorname{Tr}(\exp(\theta\boldsymbol{X})) = \sum_i \exp(\theta\lambda_i)$ by Claim 1.4, thus $\lambda_n \geq t \Rightarrow \operatorname{Tr}(\exp(\theta\boldsymbol{X})) \geq \exp(\theta t)$. Then, using Markov's inequality,

$$\begin{aligned}\Pr[\lambda_n \geq t] &\leq \Pr[\operatorname{Tr}(\exp(\theta\boldsymbol{X})) \geq \exp(\theta t)] \\ &\leq \exp(-\theta t)\,\mathbb{E}[\operatorname{Tr}(\exp(\theta\boldsymbol{X}))]\end{aligned}$$

For two independent random variables $\boldsymbol{U}$ and $\boldsymbol{V}$, we have

$$\mathop{\mathbb{E}}_{\boldsymbol{U},\boldsymbol{V}} f(\boldsymbol{U},\boldsymbol{V}) = \mathop{\mathbb{E}}_{\boldsymbol{U}} \mathop{\mathbb{E}}_{\boldsymbol{V}} [f(\boldsymbol{U},\boldsymbol{V}) | \boldsymbol{U}] = \mathop{\mathbb{E}}_{\boldsymbol{U}} \mathop{\mathbb{E}}_{\boldsymbol{V}} [f(\boldsymbol{U},\boldsymbol{V})].$$

Define $\boldsymbol{X}_{<i} = \sum_{j<i} \boldsymbol{X}_j$. Let $0 < \theta \leq 1/R$,

$$\begin{aligned}\mathbb{E}\operatorname{Tr}(\exp(\theta\boldsymbol{X})) &= \mathop{\mathbb{E}}_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{k-1}} \mathop{\mathbb{E}}_{\boldsymbol{X}_k} \operatorname{Tr}\exp\left(\underbrace{\theta\boldsymbol{X}_{<k}}_{\boldsymbol{H}} + \underbrace{\theta\boldsymbol{X}_k}_{=\log\exp(\theta\boldsymbol{X}_k)}\right), \quad \{\boldsymbol{X}_i\} \text{ are independent} \\ &\leq \mathop{\mathbb{E}}_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{k-1}} \operatorname{Tr}\exp\left(\theta\boldsymbol{X}_{<k} + \log\mathbb{E}\exp(\theta\boldsymbol{X}_k)\right), \quad \text{by 1.15 and 1.16} \\ &\leq \mathop{\mathbb{E}}_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{k-1}} \operatorname{Tr}\exp\left(\theta\boldsymbol{X}_{<k} + \log\mathbb{E}\left[\boldsymbol{I} + \theta\boldsymbol{X}_k + \theta^2\boldsymbol{X}_k^2\right]\right), \quad \text{by 1.13, 1.7, and 1.9}\end{aligned}$$

$$
\begin{aligned}
&\leq \mathop{\mathbb{E}}_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{k-1}} \operatorname{Tr}\exp\left(\theta \boldsymbol{X}_{<k} + \theta^2 \mathbb{E}\, \boldsymbol{X}_k^2\right), \quad \text{by 1.14 and 1.7} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{k-2}} \mathop{\mathbb{E}}_{\boldsymbol{X}_{k-1}} \operatorname{Tr}\exp\left(\underbrace{\theta^2 \mathbb{E}\, \boldsymbol{X}_k^2 + \theta \boldsymbol{X}_{<k-1}}_{\boldsymbol{H}} + \theta \boldsymbol{X}_{k-1}\right), \\
&\ \vdots \\
&\leq \operatorname{Tr}\exp\left(\theta^2 \sum_i \mathbb{E}\left[\boldsymbol{X}_i^2\right]\right), \\
&\leq \operatorname{Tr}\exp\left(\theta^2\sigma^2 \boldsymbol{I}\right), \quad \text{by 1.7 and } \sum_i \mathbb{E}\left[\boldsymbol{X}_i^2\right] \preceq \sigma^2 \boldsymbol{I} \\
&= n \cdot \exp(\theta^2\sigma^2).
\end{aligned}
$$

Then,

$$
\Pr[\lambda_n \geq t] \leq n \cdot \exp(-\theta t + \theta^2\sigma^2),
$$

and

$$
\Pr[\|\boldsymbol{X}\| \geq t] \leq 2n \cdot \exp(-\theta t + \theta^2\sigma^2).
$$

Similar to the proof of Bernstein concentration bound for one-dimension random variable, minimize the RHS over $0 < \theta \leq 1/R$ yields

$$
\Pr[\|\boldsymbol{X}\| \geq t] \leq 2n \cdot \exp\left(\frac{-t^2}{2Rt + 4\sigma^2}\right).
$$

□

2 Spectral Graph Sparsification

In this section, we will see that for any dense graph, we can find another sparser graph whose graph Laplacian is approximately the same as measured by their quadratic forms. This turns out to be a very useful tool for designing algorithms.

Definition 2.1. Given $\boldsymbol{A}, \boldsymbol{B} \in S_+^n$ and $\epsilon > 0$, we say

$$
\boldsymbol{A} \approx_\epsilon \boldsymbol{B} \text{ if and only if } \frac{1}{1+\epsilon}\boldsymbol{A} \leq \boldsymbol{B} \leq (1+\epsilon)\boldsymbol{A}.
$$

Suppose we start with a connected graph $G = (V, E, \boldsymbol{w})$, where as usual we say that $|V| = n$ and $|E| = m$. We want to produce another graph $\tilde{G} = (V, \tilde{E}, \tilde{\boldsymbol{w}})$ s.t $\left|\tilde{E}\right| \ll |E|$ and at the same time $\boldsymbol{L}_G \approx_\epsilon \boldsymbol{L}_{\tilde{G}}$. We call $\tilde{G}$ a *spectral sparsifier* of G. Our construction will also ensure that $\tilde{E} \subseteq E$, although this is not important in most applications. Figure 1 shows an example of a graph G and spectral sparsifier $\tilde{G}$.

We are going to construct $\tilde{G}$ by sampling some of the edges of G according to a suitable probability distribution and scaling up their weight to make up for the fact that we pick fewer of them.

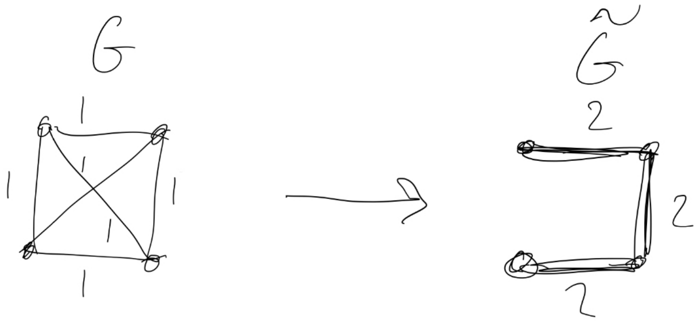

Figure 1: A graph G and a spectral sparsifier $\tilde{G}$, satisfisying $\boldsymbol{L}_G \approx_\epsilon \boldsymbol{L}_{\tilde{G}}$ for $\epsilon = 2.42$.

To get a better understanding for the notion of approximation given in 2.1 means, let's observe a simple consequence of it.

Given a vertex subset $T \subseteq V$, we say that $(T, V \setminus T)$ is a *cut* in G and that the value of the cut is

$$c_G(T) = \sum_{e \in E \cap (T \times V \setminus T)} \boldsymbol{w}(e).$$

Figure 2 shows the $c_G(T)$ in a graph G.

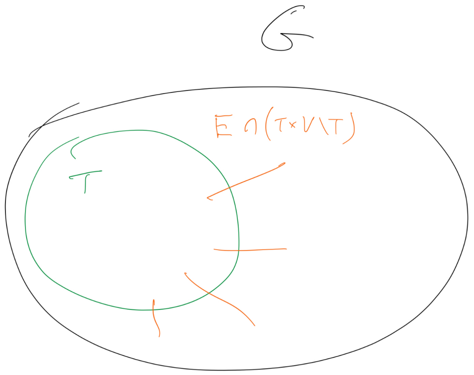

Figure 2: The cut $c_G(T)$ in G.

Theorem 2.2. *If $\boldsymbol{L}_G \approx_\epsilon \boldsymbol{L}_{\tilde{G}}$, then for all $T \subseteq V$,*

$$\frac{1}{1+\epsilon} c_G(T) \leq c_{\tilde{G}}(T) \leq (1+\epsilon) c_G(T).$$

Proof. Let $\mathbf{1}_T \in \mathbb{R}^V$ be the indicator of the set T, i.e. $\mathbf{1}_T(u) = 1$ for $u \in V$ and $\mathbf{1}_T(u) = 0$ otherwise. We can see that $\mathbf{1}_T^\top \boldsymbol{L}_G \mathbf{1}_T = c_G(T)$, and hence the theorem follows by comparing the quadratic forms. □

But how well can we spectrally approximate a graph with a sparse graph? The next theorem gives us a nearly optimal answer to this question.

Theorem 2.3 (Spectral Graph Approximation by Sampling, (Spielman-Srivastava 2008))**.** *Consider a connected graph $G = (V, E, \boldsymbol{w})$, with $n = |V|$. For any $0 < \epsilon < 1$ and $0 < \delta < 1$, there exist sampling probabilities p_e for each edge $e \in E$ s.t. if we include each edge e in $\tilde{E}$ independently with probabilty p_e and set its weight $\tilde{\boldsymbol{w}}(e) = \frac{1}{p_e}\boldsymbol{w}(e)$, then with probability at least $1 - \delta$ the graph $\tilde{G} = (V, \tilde{E}, \tilde{\boldsymbol{w}})$ satisfies*

$$\boldsymbol{L}_G \approx_\epsilon \boldsymbol{L}_{\tilde{G}} \text{ and } \left|\tilde{E}\right| \leq O(n\epsilon^{-2}\log(n/\delta)).$$

The original proof can be found in [SS11].

Remark 2.4. For convenience, we will abbreviate $\boldsymbol{L}_G$ as $\boldsymbol{L}$ and $\boldsymbol{L}_{\tilde{G}}$ as $\tilde{\boldsymbol{L}}$ in the rest of this section.

We are going to analyze a sampling procedure by turning our goal into a problem of matrix concentration. Recall that

Fact 2.5. *$\boldsymbol{A} \preceq \boldsymbol{B}$ implies $\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}^\top \preceq \boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^\top$ for any $\boldsymbol{C} \in \mathbb{R}^{n \times n}$.*

By letting $\boldsymbol{C} = \boldsymbol{L}^{+/2}$, we can see that

$$\boldsymbol{L} \approx_\epsilon \tilde{\boldsymbol{L}} \text{ implies } \boldsymbol{\Pi}_{\boldsymbol{L}} \approx_\epsilon \boldsymbol{L}^{+/2}\tilde{\boldsymbol{L}}\boldsymbol{L}^{+/2}, \tag{1}$$

where $\boldsymbol{\Pi}_{\boldsymbol{L}} = \boldsymbol{L}^{+/2}\boldsymbol{L}\boldsymbol{L}^{+/2}$ is the orthogonal projection to the complement of the kernel of $\boldsymbol{L}$.

Definition 2.6. Given a matrix $\boldsymbol{A}$, we define $\boldsymbol{\Pi}_{\boldsymbol{A}}$ to be the orthogonal projection to the complement of the kernel of $\boldsymbol{A}$, i.e. $\boldsymbol{\Pi}_{\boldsymbol{A}}\boldsymbol{v} = \boldsymbol{0}$ for $\boldsymbol{v} \in \ker(\boldsymbol{A})$ and $\boldsymbol{\Pi}_{\boldsymbol{A}}\boldsymbol{v} = \boldsymbol{v}$ for $\boldsymbol{v} \in \ker(\boldsymbol{A})^\perp$. Recall that $\ker(\boldsymbol{A})^\perp = \operatorname{im}(\boldsymbol{A}^\top)$.

Claim 2.7. *For a matrix $\boldsymbol{A} \in S^n$ with spectral decomposition $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top = \sum_i \lambda_i \boldsymbol{v}_i\boldsymbol{v}_i^\top$ s.t. $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}$, we have $\boldsymbol{\Pi}_{\boldsymbol{A}} = \sum_{i:\lambda_i \neq 0} \boldsymbol{v}_i\boldsymbol{v}_i^\top$, and $\boldsymbol{\Pi}_{\boldsymbol{A}} = \boldsymbol{A}^{+/2}\boldsymbol{A}\boldsymbol{A}^{+/2} = \boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+\boldsymbol{A}$.*

From the definition, we can see that $\boldsymbol{\Pi}_{\boldsymbol{L}} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top$.

Now that we understand the projection $\boldsymbol{\Pi}_{\boldsymbol{L}}$, it is not hard to show the following claim.

Claim 2.8.

1. *$\boldsymbol{\Pi}_{\boldsymbol{L}} \approx_\epsilon \boldsymbol{L}^{+/2}\tilde{\boldsymbol{L}}\boldsymbol{L}^{+/2}$ implies $\boldsymbol{L} \approx_\epsilon \tilde{\boldsymbol{L}}$.*
2. *For $\epsilon \leq 1$, we have that $\left\|\boldsymbol{\Pi}_{\boldsymbol{L}} - \boldsymbol{L}^{+/2}\tilde{\boldsymbol{L}}\boldsymbol{L}^{+/2}\right\| \leq \epsilon/2$ implies $\boldsymbol{\Pi}_{\boldsymbol{L}} \approx_\epsilon \boldsymbol{L}^{+/2}\tilde{\boldsymbol{L}}\boldsymbol{L}^{+/2}$.*

Really, the only idea needed here is that when comparing quadratic forms in matrices with the same kernel, we necessarily can't have the quadratic forms disagree on vectors in the kernel. Simple! But we are going to write it out carefully, since we're still getting used to these types of calculations.

Proof of Claim 2.8. To prove Part 2, we assume $\boldsymbol{\Pi}_{\boldsymbol{L}} \approx_\epsilon \boldsymbol{L}^{+/2}\tilde{\boldsymbol{L}}\boldsymbol{L}^{+/2}$. Recall that G is a connected graph, so $\ker(\boldsymbol{L}) = \operatorname{span}\{\boldsymbol{1}\}$, while $\tilde{\boldsymbol{L}}$ is the Laplacian of a graph which may or may not be connected, so $\ker(\tilde{\boldsymbol{L}}) \supseteq \ker(\boldsymbol{L})$, and equivalently $\operatorname{im}(\tilde{\boldsymbol{L}}) \subseteq \operatorname{im}(\boldsymbol{L})$. Now, for any $\boldsymbol{v} \in \ker(\boldsymbol{L})$ we have $\boldsymbol{v}^\top\tilde{\boldsymbol{L}}\boldsymbol{v} = 0 = \boldsymbol{v}^\top\boldsymbol{L}\boldsymbol{v}$. For any $\boldsymbol{v} \in \ker(\boldsymbol{L})^\perp$ we have $\boldsymbol{v} = \boldsymbol{L}^{+/2}\boldsymbol{z}$ for some $\boldsymbol{z}$, as $\ker(\boldsymbol{L})^\perp = \operatorname{im}(\boldsymbol{L}) = \operatorname{im}(\boldsymbol{L}^{+/2})$. Hence

$$\boldsymbol{v}^\top\tilde{\boldsymbol{L}}\boldsymbol{v} = \boldsymbol{z}^\top\boldsymbol{L}^{+/2}\tilde{\boldsymbol{L}}\boldsymbol{L}^{+/2}\boldsymbol{z} \geq \frac{1}{1+\epsilon}\boldsymbol{z}^\top\boldsymbol{L}^{+/2}\boldsymbol{L}\boldsymbol{L}^{+/2}\boldsymbol{z} = \frac{1}{1+\epsilon}\boldsymbol{v}^\top\boldsymbol{L}\boldsymbol{v}$$

and similarly

$$\boldsymbol{v}^\top \tilde{\boldsymbol{L}} \boldsymbol{v} = \boldsymbol{z}^\top \boldsymbol{L}^{+/2} \tilde{\boldsymbol{L}} \boldsymbol{L}^{+/2} \boldsymbol{z} \leq (1+\epsilon) \boldsymbol{z}^\top \boldsymbol{L}^{+/2} \boldsymbol{L} \boldsymbol{L}^{+/2} \boldsymbol{z} = (1+\epsilon) \boldsymbol{v}^\top \boldsymbol{L} \boldsymbol{v}.$$

Thus we have established $\boldsymbol{L} \approx_\epsilon \tilde{\boldsymbol{L}}$.

To prove Part 2, we assume $\left\| \boldsymbol{\Pi}_{\boldsymbol{L}} - \boldsymbol{L}^{+/2} \tilde{\boldsymbol{L}} \boldsymbol{L}^{+/2} \right\| \leq \epsilon/2$. This is equivalent to

$$-\frac{\epsilon}{2} \boldsymbol{I} \preceq \boldsymbol{L}^{+/2} \tilde{\boldsymbol{L}} \boldsymbol{L}^{+/2} - \boldsymbol{\Pi}_{\boldsymbol{L}} \preceq \frac{\epsilon}{2} \boldsymbol{I}$$

But since

$$\boldsymbol{1}^\top (\boldsymbol{L}^{+/2} \tilde{\boldsymbol{L}} \boldsymbol{L}^{+/2} - \boldsymbol{\Pi}_{\boldsymbol{L}}) \boldsymbol{1} = 0,$$

we can in fact sharpen this to

$$-\frac{\epsilon}{2} \boldsymbol{\Pi}_{\boldsymbol{L}} \preceq \boldsymbol{L}^{+/2} \tilde{\boldsymbol{L}} \boldsymbol{L}^{+/2} - \boldsymbol{\Pi}_{\boldsymbol{L}} \preceq \frac{\epsilon}{2} \boldsymbol{\Pi}_{\boldsymbol{L}}.$$

Rearranging, we then conclude

$$(1 - \frac{\epsilon}{2}) \boldsymbol{\Pi}_{\boldsymbol{L}} \preceq \boldsymbol{L}^{+/2} \tilde{\boldsymbol{L}} \boldsymbol{L}^{+/2} \preceq (1 + \frac{\epsilon}{2}) \boldsymbol{\Pi}_{\boldsymbol{L}}.$$

Finally, we note that $1/(1+\epsilon) \leq (1 - \frac{\epsilon}{2})$ to reach our conclusion, $\boldsymbol{\Pi}_{\boldsymbol{L}} \approx_\epsilon \boldsymbol{L}^{+/2} \tilde{\boldsymbol{L}} \boldsymbol{L}^{+/2}$. □

We now have most of the tools to prove Theorem 2.3, but to help us, we are going to establish one small piece of helpful notation: We define a matrix function $\Phi : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$ by

$$\Phi(\boldsymbol{A}) = \boldsymbol{L}^{+/2} \boldsymbol{A} \boldsymbol{L}^{+/2}.$$

We sometimes call this a "normalizing map", because it transforms a matrix to the space where spectral norm bounds can be translated into relative error guarantees compare to the $\boldsymbol{L}$ quadratic form.

Proof of Theorem 2.3. By Claim 2.8, it suffices to show

$$\left\| \boldsymbol{\Pi}_{\boldsymbol{L}} - \boldsymbol{L}^{+/2} \tilde{\boldsymbol{L}} \boldsymbol{L}^{+/2} \right\| \leq \epsilon/2. \tag{2}$$

We introduce a set of independent random variables, one for each edge e, with a probability p_e associated with the edge which we will fix later. We then let

$$\boldsymbol{Y}_e = \begin{cases} \frac{\boldsymbol{w}(e)}{p_e} \boldsymbol{b}_e \boldsymbol{b}_e^\top & \text{with probability } p_e \\ \boldsymbol{0} & \text{otherwise.} \end{cases}$$

This way, $\tilde{\boldsymbol{L}} = \sum_e \boldsymbol{Y}_e$. Note that $\mathbb{E}\left[\boldsymbol{Y}_e\right] = p_e \frac{\boldsymbol{w}(e)}{p_e} \boldsymbol{b}_e \boldsymbol{b}_e^\top = \boldsymbol{w}(e) \boldsymbol{b}_e \boldsymbol{b}_e^\top$, and so

$$\mathbb{E}\left[\tilde{\boldsymbol{L}}\right] = \sum_e \mathbb{E}\left[\boldsymbol{Y}_e\right] = \boldsymbol{L}.$$

By linearity of Φ,

$$\mathbb{E}\left[\Phi(\tilde{\boldsymbol{L}})\right] = \Phi(\mathbb{E}\left[\tilde{\boldsymbol{L}}\right]) = \boldsymbol{\Pi}_{\boldsymbol{L}}.$$

Let us also define

$$\boldsymbol{X}_e = \Phi(\boldsymbol{Y}_e) - \mathbb{E}\left[\Phi(\boldsymbol{Y}_e)\right] \text{ and } \boldsymbol{X} = \sum_e \boldsymbol{X}_e$$

Note that this ensures $\mathbb{E}\left[\boldsymbol{X}_e\right] = \boldsymbol{0}$. We are now going to fix the edge sampling probabilities, in a way that depends on some overall scaling parameter $\alpha > 0$. We let

$$p_e = \min\left(\alpha \left\|\Phi\left(\boldsymbol{w}(e)\boldsymbol{b}_e\boldsymbol{b}_e^\top\right)\right\|, 1\right)$$

then we see from the definition of $\boldsymbol{Y}_e$ that whenever $p_e < 1$

$$\left\|\Phi(\boldsymbol{Y}_e)\right\| \leq \frac{1}{\alpha}$$

from this, we can conclude, with a bit of work, that for all e

$$\left\|\boldsymbol{X}_e\right\| \leq \frac{1}{\alpha}. \tag{3}$$

We can also show that

$$\left\|\sum_e \mathbb{E}\left[\boldsymbol{X}_e^2\right]\right\| \leq \frac{1}{\alpha}. \tag{4}$$

In the exercises for this lecture, we will ask you to show that Equations (3) and (4) holds.

This means that we can apply Theorem 1.1 to our $\boldsymbol{X} = \sum_e \boldsymbol{X}_e$, with $R = \frac{1}{\alpha}$ and $\sigma^2 = \frac{1}{\alpha}$, to get

$$\Pr\left[\left\|\boldsymbol{\Pi_L} - \boldsymbol{L}^{+/2}\tilde{\boldsymbol{L}}\boldsymbol{L}^{+/2}\right\| \geq \epsilon/2\right] \leq 2n\exp\left(\frac{-0.25\epsilon^2}{(\epsilon+4)/\alpha}\right)$$

Since $0 < \epsilon < 1$, this means that if $\alpha = 40\epsilon^{-2}\log(n/\delta)$, then

$$Pr\left[\left\|\boldsymbol{\Pi_L} - \boldsymbol{L}^{+/2}\tilde{\boldsymbol{L}}\boldsymbol{L}^{+/2}\right\| \geq \epsilon/2\right] \leq \frac{2n\delta^2}{n^2} \leq \delta/2.$$

In the last step, we assumed $n \geq 4$.

Lastly, we'd like to know that the graph $\tilde{G}$ is sparse. The number of edges in $\tilde{G}$ is equal to the number of $\boldsymbol{Y}_e$ that come out nonzero. Thus, the expected value of $\left|\tilde{E}\right|$ is

$$\mathbb{E}\left[\left|\tilde{E}\right|\right] = \sum_e p_e \leq \alpha \sum_e \boldsymbol{w}(e)\left\|\boldsymbol{L}^{+/2}\boldsymbol{b}_e\boldsymbol{b}_e^\top\boldsymbol{L}^{+/2}\right\|$$

We can bound the sum of the norms with a neat trick relating it to the trace of $\boldsymbol{\Pi_L}$. Note that in general for a vector $\boldsymbol{a} \in \mathbb{R}^n$, we have $\left\|\boldsymbol{a}\boldsymbol{a}^\top\right\| = \boldsymbol{a}^\top\boldsymbol{a} = \operatorname{Tr}\left(\boldsymbol{a}\boldsymbol{a}^\top\right)$. And hence

$$\begin{aligned}
\sum_e \boldsymbol{w}(e)\left\|\boldsymbol{L}^{+/2}\boldsymbol{b}_e\boldsymbol{b}_e^\top\boldsymbol{L}^{+/2}\right\| &= \sum_e \boldsymbol{w}(e)\operatorname{Tr}\left(\boldsymbol{L}^{+/2}\boldsymbol{b}_e\boldsymbol{b}_e^\top\boldsymbol{L}^{+/2}\right) \\
&= \operatorname{Tr}\left(\boldsymbol{L}^{+/2}\left(\sum_e \boldsymbol{w}(e)\boldsymbol{b}_e\boldsymbol{b}_e^\top\right)\boldsymbol{L}^{+/2}\right) \\
&= \operatorname{Tr}\left(\boldsymbol{\Pi_L}\right) = n - 1.
\end{aligned}$$

Thus with our choice of α,

$$\mathbb{E}\left[\left|\tilde{E}\right|\right] \leq 40\epsilon^{-2}\log(n/\delta)n.$$

With a scalar Chernoff bound, can show that $\left|\tilde{E}\right| \leq O(\epsilon^{-2}\log(n/\delta)n)$ with probability at least $1-\delta/2$. Thus by a union bound, the this condition and Equation (2) are both satisfied with probability at least $1-\delta$. □

Remark 2.9. Note that

$$\left\|\Phi\left(\boldsymbol{w}(e)\boldsymbol{b}_e\boldsymbol{b}_e^\top\right)\right\| = \boldsymbol{w}(e)\left\|\boldsymbol{L}^{+/2}\boldsymbol{b}_e\boldsymbol{b}_e^\top\boldsymbol{L}^{+/2}\right\| \leq \boldsymbol{w}(e)\left\|\boldsymbol{L}^{+/2}\boldsymbol{b}_e\right\|_2^2.$$

Recall that in Lecture 6, we saw that the effective between vertex v and vertex u is given by $\left\|\boldsymbol{L}^{+/2}(\boldsymbol{e}_u-\boldsymbol{e}_v)\right\|_2^2$, and for an edge e connecting vertex u and v, we have $\boldsymbol{b}_e = \boldsymbol{e}_u - \boldsymbol{e}_v$. That means the norm of the "baby Laplacian" $\boldsymbol{w}(e)\boldsymbol{b}_e\boldsymbol{b}_e^\top$ of a single edge with weight $\boldsymbol{w}(e)$ is exactly $\boldsymbol{w}(e)$ times the effective resistance between the two endpoints of the edge.

We haven't shown how to compute the sampling probabilities efficiently, so right now, it isn't clear whether we can efficiently find $\tilde{G}$. It turns out that if we have access to a fast algorithm for solving Laplacian linear equations, then we can find sufficiently good approximations to the effective resistances quickly, and use these to compute $\tilde{G}$. An algorithm for this is described in [SS11].

References

[SS11] Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

[T+15] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.