Cornell University, Fall 2012 **CS 6820: Algorithms**
Lecture notes on spectral methods in algorithm design

Studying the eigenvalues and eigenvectors of matrices has powerful consequences for at least three areas of algorithm design: graph partitioning, analysis of high-dimensional data, and analysis of Markov chains. Collectively, these techniques are known as *spectral methods* in algorithm design. These lecture notes present the fundamentals of spectral methods.

1 Review: symmetric matrices, their eigenvalues and eigenvectors

This section reviews some basic facts about real symmetric matrices. If $A = (a_{ij})$ is an $n - \times - n$ square symmetric matrix, then $\mathbb{R}^n$ has a basis consisting of eigenvectors of A, these vectors are mutually orthogonal, and all of the eigenvalues are real numbers. Furthermore, the eigenvectors and eigenvalues can be characterized as solutions of natural maximization or minimization problems involving *Rayleigh quotients*.

Definition 1.1. If x is a nonzero vector in $\mathbb{R}^n$ and A is an $n - \times - n$ matrix, then the Rayleigh quotient of x with respect to A is the ratio

$$RQ_A(x) = \frac{x^\mathsf{T} A x}{x^\mathsf{T} x}.$$

Definition 1.2. If A is an $n - \times - n$ matrix, then a linear subspace $V \subseteq \mathbb{R}^n$ is called an *invariant subspace* of A if it satisfies $Ax \in V$ for all $x \in V$.

Lemma 1.3. *If A is a real symmetric matrix and V is an invariant subspace of A, then there is some $x \in V$ such that $RQ_A(x) = \inf\{RQ_A(y) \mid y \in V\}$. Any $x \in V$ that minimizes $RQ_A(x)$ is an eigenvector of A, and the value $RQ_A(x)$ is the corresponding eigenvalue.*

Proof. If x is a vector and r is a nonzero scalar, then $RQ_A(x) = RQ_A(rx)$, hence every value attained in V by the function RQ_A is attained on the unit sphere $S(V) = \{x \in V \mid x^\mathsf{T} x = 1\}$. The function RQ_A is a continuous function on $S(V)$, and $S(V)$ is compact (closed and bounded) so by basic real analysis we know that RQ_A attains its minimum value at some unit vector $x \in S(V)$. Using the quotient rule we can compute the gradient

$$\nabla RQ_A(x) = \frac{2Ax - 2(x^\mathsf{T} A x)x}{(x^\mathsf{T} x)^2}. \tag{1}$$

At the vector $x \in S(V)$ where RQ_A attains its minimum value in V, this gradient vector must be orthogonal to V; otherwise, the value of RQ_A would decrease as we move away from x in the direction of any $y \in V$ that has negative dot product with $\nabla RQ_A(x)$. On the other hand, our assumption that V is an invariant subspace of A implies that the right side of (1) belongs to V. The only way that $\nabla RQ_A(x)$ could be orthogonal to V while also belonging to V is if it is the zero vector, hence $Ax = \lambda x$ where $\lambda = x^\mathsf{T} A x = RQ_A(x)$. □

Lemma 1.4. *If A is a real symmetric matrix and V is an invariant subspace of A, then $V^\perp = \{x \mid x^\mathsf{T} y = 0 \; \forall y \in V\}$ is also an invariant subspace of A.*

Proof. If V is an invariant subspace of A and $x \in V^\perp$, then for all $y \in V$ we have

$$(Ax)^\mathsf{T} y = x^\mathsf{T} A^\mathsf{T} y = x^\mathsf{T} A y = 0,$$

hence $Ax \in V^\perp$. □

Combining these two lemmas, we obtain a recipe for extracting all of the eigenvectors of A, with their eigenvalues arranged in increasing order.

Theorem 1.5. *Let A be an $n - \times - n$ real symmetric matrix and let us inductively define sequences*

$$\begin{aligned} x_1, \ldots, x_n &\in \mathbb{R}^n \\ \lambda_1, \ldots, \lambda_n &\in \mathbb{R} \\ \{0\} = V_0 \subseteq V_1 \subseteq \cdots \subseteq V_n &= \mathbb{R}^n \\ \mathbb{R}^n = W_0 \supseteq W_1 \supseteq \cdots \supseteq W_n &= \{0\} \end{aligned}$$

by specifying that

$$\begin{aligned} x_i &= \operatorname{argmin} \{RQ_A(x) \mid x \in W_{i-1}\} \\ \lambda_i &= RQ_A(x_i) \\ V_i &= \operatorname{span}(x_1, \ldots, x_i) \\ W_i &= V_i^\perp. \end{aligned}$$

Then $x_1, \ldots, x_n$ are mutually orthogonal eigenvectors of A, and $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the corresponding eigenvalues.

Proof. The proof is by induction on i. The induction hypothesis is that $\{x_1, \ldots, x_i\}$ is a set of mutually orthogonal eigenvectors of A constituting a basis of V_i, and $\lambda_1 \leq \cdots \leq \lambda_i$ are the corresponding eigenvectors. Given this induction hypothesis, and the preceding lemmas, the proof almost writes itself. Each time we select a new x_i, it is guaranteed to be orthogonal to the preceding ones because $x_i \in W_{i-1} = V_{i-1}^\perp$. The linear subspace V_{i-1} is A-invariant because it is spanned by eigenvectors of A; by Lemma 1.4 its orthogonal complement W_{i-1} is also A-invariant and this implies, by Lemma 1.3 that x_i is an eigenvector of A and λ_i is its corresponding eigenvalue. Finally, $\lambda_i \geq \lambda_{i-1}$ because $\lambda_{i-1} = \min\{RQ_A(x) \mid x \in W_{i-2}\}$, while $\lambda_i = RQ_A(x_i) \in \{RQ_A(x) \mid x \in W_{i-2}\}$. □

An easy corollary of Theorem 1.5 is the *Courant-Fischer Theorem.*

Theorem 1.6 (Courant-Fischer)**.** *The eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ of an $n - \times - n$ real symmetric matrix satisfy:*

$$\forall k \; \lambda_k = \min_{\dim(V)=k} \left(\max_{x \in V} RQ_A(x) \right) = \max_{\dim(W)=n-k+1} \left(\min_{x \in W} RQ_A(x) \right).$$

Proof. The vector space W_{k-1} constructed in the proof of Theorem 1.5 has dimension $n-k+1$, and by construction it satisfies $\min_{x \in W_{k-1}} RQ_A(x) = \lambda_k$. Therefore

$$\max_{\dim(W)=n-k+1} \left(\min_{x \in W} RQ_A(x) \right) \geq \lambda_k.$$

If $W \subseteq \mathbb{R}^n$ is any linear subspace of dimension $n-k+1$ then $W \cap V_k$ contains a nonzero vector x, because $\dim(W) + \dim(V_k) > n$. Since $V_k = \operatorname{span}(x_1, \ldots, x_k)$ we can write $x = a_1 x_1 + \cdots + a_k x_k$. Rescaling $x_1, \ldots, x_k$ if necessary, we can assume that they are all unit vectors. Then, using the fact that $x_1, \ldots, x_k$ are mutually orthogonal eigenvectors of A, we obtain

$$RQ_A(x) = \frac{\lambda_1 a_1 + \cdots + \lambda_k a_k}{a_1 + \cdots + a_k} \leq \lambda_k.$$

Therefore $\max_{\dim(W)=n-k+1} (\min_{x \in W} RQ_A(x)) \leq \lambda_k$. Combining this with the inequality derived in the preceding paragraph, we obtain $\max_{\dim(W)=n-k+1} \min_{x \in W} RQ_A(x) = \lambda_k$. Replacing A with $-A$, and k with $n-k+1$, we obtain $\min_{\dim(V)=k} (\max_{x \in V} RQ_A(x)) = \lambda_k$. □

2 The Graph Laplacian

Two symmetric matrices play a vital role in the theory of graph partitioning. These are the Laplacian and normalized Laplacian matrix of a graph G.

Definition 2.1. If G is an undirected graph with non-negative edge weights $w(u, v) \geq 0$, the *weighted degree* of a vertex u, denoted by $d(u)$, is the sum of the weights of all edges incident to u. The Laplacian matrix of G is the matrix L_G with entries

$$(L_G)_{uv} = \begin{cases} d(u) & \text{if } u = v \\ -w(u, v) & \text{if } u \neq v \text{ and } (u, v) \in E \\ 0 & \text{if } u \neq v \text{ and } (u, v) \notin E. \end{cases}$$

If D_G is the diagonal matrix whose (u, u)-entry is $d(u)$, and if G has no vertex of weighted degree 0, then the normalized Laplacian matrix of G is

$$\overline{L}_G = D_G^{-1/2} L_G D_G^{-1/2}.$$

The eigenvalues of L_G and $\overline{L}_G$ will be denoted in these notes by $\lambda_1(G) \leq \cdots \leq \lambda_n(G)$ and $\nu_1(G) \leq \cdots \leq \nu_n(G)$. When the graph G is clear from context, we will simply write these as $\lambda_1, \ldots, \lambda_n$ or $\nu_1, \ldots, \nu_n$.

The "meaning" of the Laplacian matrix is best explained by the following observation.

Observation 2.2. The Laplacian matrix L_G is the unique symmetric matrix satisfying the following relation for all vectors $x \in \mathbb{R}^V$.

$$x^{\mathsf{T}} L_G x = \sum_{(u,v) \in E} w(u, v)(x_u - x_v)^2. \tag{2}$$

The following lemma follows easily from Observation 2.2.

Lemma 2.3. *The Laplacian matrix of a graph G is a positive semidefinite matrix. Its minimum eigenvalue is 0. The multiplicity of this eigenvalue equals the number of connected components of G.*

Proof. The right side of (2) is always non-negative, hence L_G is positive semidefinite. The right side is zero if and only if x is constant on each connected component of G (i.e., it satisfies $x_u = x_v$ whenever u, v belong to the same component), hence the multiplicity of the eigenvalue 0 equals the number of connected components of G. □

The normalized Laplacian matrix has a more obscure graph-theoretic meaning than the Laplacian, but its eigenvalues and eigenvectors are actually more tightly connected to the structure of G. Accordingly, we will focus on normalized Laplacian eigenvalues and eigenvectors in these notes. The cost of doing so is that the matrix $\overline{L}_G$ is a bit more cumbersome to work with. For example, when G is connected the 0-eigenspace of L_G is spanned by the all-ones vector $\mathbf{1}$ whereas the 0-eigenspace of $\overline{L}_G$ is spanned by the vector $\mathbf{d}^{1/2} = D_G^{1/2}\mathbf{1}$.

3 Sparsity and expansion

We will relate the eigenvalue $\nu_2(G)$ to two graph parameters called the *expansion* (also known as *conductance*) and the *sparsity* of G. Both of them measure the value of the "sparsest" cut, with respect to subtly differing notions of sparsity. For any set of vertices S, define

$$d(S) = \sum_{u \in S} d(u)$$

and define the edge boundary

$$\partial S = \{e = (u, v) \mid \text{exactly one of } u, v \text{ belongs to } S\}.$$

The *sparsity* of G is

$$\sigma(G) = \min_{(S,\overline{S})} \left\{ d(V) \cdot \frac{w(\partial S)}{d(S)d(\overline{S})} \right\}$$

and the *expansion* of G is

$$\phi(G) = \min_{(S,\overline{S})} \left\{ \frac{w(\partial S)}{\min\{d(S), d(\overline{S})\}} \right\},$$

where the minimum in both cases is over all vertex sets $S \neq \emptyset, V$. Note that for any such S,

$$\frac{d(V)}{d(S)d(\overline{S})} = \frac{d(V)}{\min\{d(S), d(\overline{S})\} \cdot \max\{d(S), d(\overline{S})\}} = \frac{1}{\min\{d(S), d(\overline{S})\}} \cdot \frac{d(V)}{\max\{d(S), d(\overline{S})\}}.$$

The second factor on the right side is between 1 and 2, and it easily follows that

$$\phi(G) \leq \sigma(G) \leq 2\phi(G).$$

Thus, each of the parameters $\sigma(G), \phi(G)$ is a 2-approximation to the other one. Unfortunately, it is not known how to compute a $O(1)$-approximation to either of these parameters in polynomial time. In fact, assuming the Unique Games Conjecture, it is NP-hard to compute an $O(1)$-approximation to either of them.

4 Cheeger's Inequality: Lower Bound on Conductance

There is a sense, however, in which $\nu_2(G)$ constitutes an approximation to $\sigma(G)$. To see why, let us begin with the following characterization of $\nu_2(G)$ that comes directly from Courant-Fischer.

$$\nu_2(G) = \min\left\{ \frac{x^\mathsf{T}\overline{L}_G x}{x^\mathsf{T}x} \middle| x \neq 0,\, x^\mathsf{T} D_G^{1/2}\mathbf{1} = 0 \right\} = \min\left\{ \frac{y^\mathsf{T} L_G y}{y^\mathsf{T} D_G y} \middle| y \neq 0,\, y^\mathsf{T} D_G \mathbf{1} = 0 \right\}.$$

The latter equality is obtained by setting $x = D_G^{1/2} y$.

The following lemma allows us to rewrite the Rayleigh quotient $\frac{y^\mathsf{T} L_G y}{y^\mathsf{T} D_G y}$ in a useful form, when $y^\mathsf{T} D_G \mathbf{1} = 0$.

Lemma 4.1. *For any vector y we have*

$$y^\mathsf{T} D_G y \geq \frac{1}{2d(V)} \sum_{u \neq v} d(u)d(v)(y(u) - y(v))^2,$$

with equality if and only if $y^\mathsf{T} D_G \mathbf{1} = 0$.

Proof.

$$\begin{aligned}
\frac{1}{2}\sum_{u\neq v} d(u)d(v)(y(u)-y(v))^2 &= \frac{1}{2}\sum_{u\neq v} d(u)d(v)[y(u)^2 + y(v)^2] - \sum_{u\neq v} d(u)d(v)y(u)y(v) \\
&= \sum_{u\neq v} d(u)d(v)y(u)^2 - \sum_{u\neq v} d(u)d(v)y(u)y(v) \\
&= \sum_{u,v} d(u)d(v)y(u)^2 - \sum_{u,v} d(u)d(v)y(u)y(v) \\
&= d(V)\sum_u d(u)y(u)^2 - \left(\sum_u d(u)y(u)\right)^2 \\
&= d(V) y^\mathsf{T} D_G y - \left(y^\mathsf{T} D_G \mathbf{1}\right)^2.
\end{aligned}$$

□

A corollary of the lemma is the formula

$$\nu_2(G) = \inf \left\{ d(V) \frac{\sum_{(u,v)\in E(G)} w(u,v)(y(u)-y(v))^2}{\sum_{u<v} d(u)d(v)(y(u)-y(v))^2} \,\middle|\, \text{denominator is nonzero} \right\}, \quad (3)$$

where the summation over $u < v$ in the denominator is meant to indicate that each unordered pair $\{u, v\}$ of distinct vertices contributes exactly one term to the sum. The corollary is obtained by noticing that the numerator and denominator on the right side are invariant under adding a scalar multiple of $\mathbf{1}$ to y, and hence one of the vectors attaining the infimum is orthogonal to $D_G\mathbf{1}$.

Let us evaluate the quotient on the right side of (3) when y is the characteristic vector of a cut $(S, \overline{S})$, defined by

$$y(u) = \begin{cases} 1 & \text{if } u \in S \\ 0 & \text{if } u \in \overline{S}. \end{cases}$$

In that case,

$$\sum_{(u,v)\in E(G)} w(u,v)(y(u)-y(v))^2 = \sum_{(u,v)\in \partial S} w(u,v) = w(\partial S)$$

while

$$\sum_{u<v} d(u)d(v)(y(u)-y(v))^2 = \sum_{u\in S}\sum_{v\in\overline{S}} d(u)d(v) = d(S)d(\overline{S}).$$

Hence,

$$\nu_2(G) \le d(V)\frac{w(\partial S)}{d(S)d(\overline{S})},$$

and taking the minimum over all $(S, \overline{S})$ we obtain

$$\nu_2(G) \le \sigma(G).$$

5 Cheeger's Inequality: Upper Bound on Conductance

The inequality $\nu_2(G) \le \sigma(G)$ is the easy half of Cheeger's Inequality; the more difficult half asserts that there is also an upper bound on $\sigma(G)$ of the form

$$\sigma(G) \le \sqrt{8\nu_2(G)}.$$

Owing to the inequality $\sigma(G) \le 2\phi(G)$, it suffices to prove that

$$\phi(G) \le \sqrt{2\nu_2(G)}$$

and that is, in fact, the next thing we will prove.

For any vector y that is not a scalar multiple of $\mathbf{1}$, define

$$Q(y) = d(V)\frac{\sum_{(u,v)\in E(G)} w(u,v)(y(u)-y(v))^2}{\sum_{u<v} d(u)d(v)(y(u)-y(v))^2}.$$

Given any such y, we will find a cut $(S,\overline{S})$ such that $\frac{w(\partial S)}{\min\{d(S),d(\overline{S})\}} \leq \sqrt{2Q(y)}$; the upper bound $\phi(G) \leq \sqrt{2\nu_2(G)}$ follows immediately by choosing y to be a vector minimizing $Q(y)$. In fact, if we number the vertices of G as $v_1, v_2, \ldots, v_n$ such that $y_1 \leq y_2 \leq \cdots \leq y_n$, we will show that it suffices to take S to be one of the sets $\{y_1, \ldots, y_k\}$ for $1 \leq k < n$.

Note that $Q(y)$ is unchanged when we add a scalar multiple of $\mathbf{1}$ to y. Accordingly, we can assume without loss of generality that

$$\sum_{y_i<0} d(v_i) \leq \sum_{y_i\geq 0} d(v_i)$$
$$\sum_{y_i\leq 0} d(v_i) \geq \sum_{y_i>0} d(v_i)$$

For d-regular graphs, this essentially means that we're setting the median of the components of y to be zero. For irregular graphs, it essentially says that we're balancing the total degree of the vertices with positive $y(u)$ and those with negative $y(u)$.

Now here comes the most unmotivated part of the proof. Define a vector z by

$$z_i = \begin{cases} -y_i^2 & \text{if } y_i < 0 \\ y_i^2 & \text{if } y_i \geq 0. \end{cases}$$

Note also that $Q(y)$ is unchanged when we multiply y by a nonzero scalar. Accordingly, we can assume that $z_n - z_1 = 1$. Now choose a threshold value t uniformly at random from the interval $[z_1, z_n]$ and let

$$S = \{v_i \mid z_i < t\}.$$

We will prove that

$$\frac{\mathbb{E}[w(\partial S)]}{\mathbb{E}[\min\{d(S), d(\overline{S})\}]} \leq \sqrt{2Q(y)}$$

from which it follows that

$$\mathbb{E}[w(\partial S)] \leq \sqrt{2Q(y)} \cdot \mathbb{E}[\min\{d(S), d(\overline{S})\}]$$

and consequently that there is at least one S in the support of our distribution such that

$$w(\partial S) \leq \sqrt{2Q(y)} \cdot \min\{d(S), d(\overline{S})\}.$$

It is surprisingly easy to evaluate $\mathbb{E}[\min\{d(S), d(\overline{S})\}]$. Each vertex v_i contributes $d(v_i)$ to the expression inside the expectation operator when it belongs to the smaller side of the

cut, which happens if and only if t lands between 0 and z_i, an event with probability $|z_i|$. Consequently,

$$\mathbb{E}[\min\{d(S), d(\overline{S})\}] = \sum_u d(u)|z(u)| = \sum_u d(u)y(u)^2 = y^\mathsf{T} D_G y.$$

Meanwhile, to bound the numerator $\mathbb{E}[w(\partial S)]$, observe that an edge (u, v) contributes $w(u, v)$ to the numerator if and only if it is cut, an event having probability $|z(u) - z(v)|$. A bit of case analysis reveals that

$$\forall u, v \;\; |z(u) - z(v)| \leq |y(u) - y(v)| \cdot (|y(u)| + |y(v)|),$$

since the left and right sides are equal when $y(u), y(v)$ have the same sign, and otherwise the left side equals $y(u)^2 + y(v)^2$ while the right side equals $(|y(u)| + |y(v)|)^2$. Combining this estimate of the numerator with Cauchy-Schwartz, we find that

$$\begin{aligned}
\mathbb{E}[w(\partial S)] &\leq \sum_{(u,v)\in E(G)} w(u,v)|y(u) - y(v)|(|y(u)| + |y(v)|) \\
&\leq \left(\sum_{(u,v)\in E(G)} w(u,v)(y(u) - y(v))^2 \right)^{1/2} \left(\sum_{(u,v)\in E(G)} w(u,v)(|y(u)| + |y(v)|)^2 \right)^{1/2} \\
&\leq \left(\frac{Q(y)}{d(V)} \sum_{u<v} d(u)d(v)(y(u) - y(v))^2 \right)^{1/2} \left(\sum_{(u,v)\in E(G)} w(u,v)(2y(u)^2 + 2y(v)^2) \right)^{1/2} \\
&\leq \left(Q(y) y^\mathsf{T} D_G y\right)^{1/2} \left(2 \sum_u d(u)y(u)^2 \right)^{1/2} \\
&= (2Q(y))^{1/2} y^\mathsf{T} D_G y.
\end{aligned}$$

6 Laplacian eigenvalues and spectral partitioning

We've seen a connection between sparse cuts and eigenvectors of the *normalized* Laplacian matrix. However, in some contexts it is easier to work with eigenvalues and eigenvectors of the unnormalized Laplacian, L_G. One can use eigenvectors of L_G for spectral partitioning, provided one is willing to tolerate weaker bounds for graphs with unbalanced degree sequences. For example, if y is an eigenvector of L_G satisfying $L_G y = \lambda_2 y$ then we can express $Q(y)$ as follows:

$$Q(y) = d(V) \frac{y^\mathsf{T} L_G y}{\sum_{u<v} d(u)d(v)(y(u) - y(v))^2} = \frac{\lambda_2 \|y\|^2 d(V)}{\sum_{u<v} d(u)d(v)(y(u) - y(v))^2}.$$

To estimate the denominator, let $d_{\min}$ and d_{avg} denote the minimum and the average degree of G, respectively. We have

$$\begin{aligned}\sum_{u<v} d(u)d(v)(y(u)-y(v))^2 &= \frac{1}{2}\sum_{u\neq v} d(u)d(v)(y(u)-y(v))^2 \\ &\geq \frac{1}{2}d_{\min}^2 \sum_{u\neq v}(y(u)-y(v))^2 \\ &= nd_{\min}^2 \sum_u y(u)^2 = \frac{d(V)}{d_{\text{avg}}} d_{\min}^2 \|y\|^2.\end{aligned}$$

Hence

$$Q(y) \leq \frac{d_{\text{avg}}}{d(V)} \frac{\lambda_2 \|y\|^2 d(V)}{d_{\min}^2 \|y\|^2} = \left(\frac{d_{\text{avg}}}{d_{\min}^2}\right) \lambda_2.$$

7 Spectral sparsification of graphs

For a dense graph G with n vertices and $m \gg n$ edges, it is often desirable to compute a sparse approximation H, i.e. an edge-weighted graph with the same vertex set but with $O(n)$ or $O(n \log n)$ edges, such that

$$(1-\varepsilon)L_G \preceq L_H \preceq (1+\varepsilon)L_G. \tag{4}$$

Such a graph is called a *spectral sparsification* of G. It is useful because it preserves some of the essential features of G. For example, we have seen that for any vertex set S, if x denotes the vector

$$x_u = \begin{cases} 1 & \text{if } u \in S \\ 0 & \text{if } u \notin S \end{cases}$$

then the capacity of the cut $(S, \overline{S})$, with edge set ∂S, is given by

$$c(\partial S) = x^{\mathsf{T}} L_G x.$$

In light of equation (4) we know that a spectral sparsifier H satisfies

$$(1-\varepsilon)x^{\mathsf{T}} L_G x \leq x^{\mathsf{T}} L_H x \leq (1+\varepsilon)x^{\mathsf{T}} L_G x$$

hence a spectral sparsifier preserves the capacity of every cut in G, up to a factor of $1 \pm \varepsilon$.

Random sampling furnishes a simple method for computing a spectral sparsifier of G. We will be designing and analyzing an algorithm that samples edges $e_1, \ldots, e_k$ independently, each drawn from a probability distribution that will be denoted by $\{\pi(e) \mid e \in E\}$. Designing an appropriate sampling distribution will be the subtlest part of the algorithm, and we will defer discussion of how to choose π for now. The number of sampled edges, k, will turn out to be $O(n \log n)$, but for now we'll also leave k as a parameter of the algorithm whose precise value will be specified later.

For any edge $e = (u, v)$ let δ_e denote the vector whose components are defined by

$$(\delta_e)_w = \begin{cases} -1 & \text{if } w = u \\ 1 & \text{if } w = v \\ 0 & \text{otherwise.} \end{cases}$$

The vector δ_e is only well-defined up to sign. In other words, the undirected edge $e = (u, v)$ is equally well represented as $e = (v, u)$, but these two representations lead to the vectors δ_e and $-\delta_e$, respectively. The sign ambiguity will not matter, because we won't be dealing directly with the vector δ_e but instead with the rank-one matrix $\delta_e \delta_e^{\mathsf{T}}$. The equation $(-\delta_e)(-\delta_e^{\mathsf{T}}) = \delta_e \delta_e^{\mathsf{T}}$ assures that we get the same matrix no matter which choice we make for δ_e.

Recall that the Laplacian of a graph G with edge capacities $c(e)$ is given by the weighted sum

$$L_G = \sum_{e \in E} c(e) \delta_e \delta_e^{\mathsf{T}}. \tag{5}$$

Similarly, for our random graph H, if we choose a "rescaled capacity" for each edge e, and set the capacity of e in H to $\hat{c}(e)$ times the number of times e occurs in the multi-set $\{e_1, \ldots, e_k\}$ of randomly sampled edges, then the Laplacian of H will be given by

$$L_H = \sum_{i=1}^{k} \hat{c}(e_i) \delta_{e_i} \delta_{e_i}^{\mathsf{T}} \tag{6}$$

and its expected value will be

$$\mathbb{E}[L_H] = k \cdot \sum_{e \in E} \pi(e) \hat{c}(e) \delta_e \delta_e^{\mathsf{T}}.$$

To equate $\mathbb{E}[L_H]$ with L_G the simplest thing to do is to equate the coefficient of $\delta_e \delta_e^{\mathsf{T}}$ for each edge e, which necessitates setting

$$\begin{aligned} k \hat{c}(e) \pi(e) &= c(e) \\ \hat{c}(e) &= \tfrac{c(e)}{k \pi(e)}. \end{aligned}$$

Thus, the capacities $\hat{c}(e)$ of the sampled edges will be uniquely determined by the number of sampled edges, k, and the sampling distribution, π.

To analyze the quality of the spectral approximation achieved by the sampling algorithm, we need to estimate the extent to which the random matrix L_H may differ from its expectation, $\mathbb{E}[L_H]$. Since L_H is a sum of independent, identically distributed random matrices—namely, the summands on the right side of (6)—it is natural to use the Ahlswede-Winter Inequality. In our application of the inequality, the average of the k summands has expected value $\frac{1}{k}\mathbb{E}[L_H] = \frac{1}{k} L_G$. Thus, to apply the inequality, we need to find a constant $R \geq 1$ such that for each edge e,

$$\begin{aligned} \hat{c}(e) \delta_e \delta^{\mathsf{T}} &\preceq R \cdot \left(\tfrac{1}{k} L_G\right) \\ c(e) \delta_e \delta_e^{\mathsf{T}} &\preceq R \pi(e) \cdot L_G. \end{aligned} \tag{7}$$

The Ahlswede-Winter Inequality will then ensure that with probability at least $1-2n\,\exp(-\frac{\varepsilon^2 k}{4R})$, we have $(1-\varepsilon)L_G \preceq L_H \preceq (1+\varepsilon)L_G$, as desired. If we want this event to happen with probability at least $\frac{1}{2}$, we set $k = 4R\varepsilon^{-2}\ln(4n)$. Thus, the number of edges we need to sample when constructing H is linearly related to the constant R appearing on the right side of (7), and designing a good graph sparsification algorithms boils down to constructing a distribution $\{\pi(e)\}$ that allows R to be as small as possible.

As a naïve first attempt, we could take π to be the uniform distribution, $\pi(e) = \frac{1}{m}$ for all e. Then, noting that the formula (5) justifies the relation $c(e)\delta_e\delta_e^\mathsf{T} \preceq L_G$ for all e, we see that we just need to make R large enough that $R\pi(e) \geq 1$ for all e. Since we are using $\pi(e) = \frac{1}{m}$ this means setting $R = m$. Our naïve idea of setting π to be the uniform distribution has not worked out well: instead of sparsifying G, we have *increasing* the number of edges from m to $k = 4m\varepsilon^{-2}\ln(4n)$.

One might hope that the uniform sampling technique performs better than the above analysis would suggest. After all, our analysis made use of the relation $c(e)\delta_e\delta_e^\mathsf{T} \preceq L_G$, which typically has a large amount of "slack" because L_G is a sum of m positive semidefinite matrices, only one of which is $c(e)\delta_e\delta_e^\mathsf{T}$. However, on closer inspection, the whose idea of uniform edge sampling is doomed to require sampling $\Omega(m)$ edges in the worst case. To see why, consider the case that G is made up of two cliques K_0, K_1, each of size $n/2$, joined by a single edge, e. Let x denote the vector defined by setting $x_u = 1$ if $u \in K_0$, $x_u = 0$ if $u \in K_1$. If we fail to sample edge e when constructing the sparsifier, H, then $x^\mathsf{T}L_H x = 0$ whereas $x^\mathsf{T}L_G x > 0$, which rules out the possibility that H is a spectral sparsifier of G. Thus, if we sample $o(m)$ edges from the uniform distribution, with probability $1 - o(1)$ we will fail to obtain a spectral sparsifier. Our only hope is to use a non-uniform distribution over edges that assigns higher probability to edges, such as the edge e in the foregoing example, that are "spectrally irreplaceable", meaning that they must be included in any spectral sparsifier of G.

Since our goal is to minimize R, a more principled way of designing the distribution π consists of solving the following semidefinite program whose variables are R and $\{\pi(e) \mid e \in E\}$.

$$\begin{array}{lll} \text{minimize} & R & \\ \text{subject to} & c(e)\delta_e\delta_e^\mathsf{T} \preceq R\pi(e)\cdot L_G & \forall e \in E \\ & \sum_{e\in E}\pi(e) = 1 & \\ & \pi(e) \geq 0 & \forall e \in E \end{array} \tag{8}$$

The first constraint can be rewritten as

$$\forall e = (u,v) \in E \quad R \geq \frac{c(e)}{\pi(e)} \cdot \max\left\{ \frac{(\delta_e^\mathsf{T}x)^2}{x^\mathsf{T}L_G x} \,\middle|\, x \neq 0 \right\} \tag{9}$$

and it will be helpful to solve the maximization problem on the right side. Since δ_e is orthogonal to the nullspace of L_G, the quotient $\frac{c(e)(\delta_e^\mathsf{T}x)^2}{\pi(e)\,x^\mathsf{T}L_G x}$ is unchanged if we add any vector in the nullspace of L_G to x. For this reason, among the set of vectors x that attain the maximum

on the right side of (9) there is one that is orthogonal to the nullspace of L_G and we may assume henceforth that x is such a vector. In particular, this means $L_G^+ L_G x = L_G L_G^+ x = x$. Let $y = L_G^{1/2} x$ and . Then

$$\frac{(\delta_e^\mathsf{T} x)^2}{x^\mathsf{T} L_G x} = \frac{(\delta_e^\mathsf{T} (L_G^+)^{1/2} y)^2}{y^\mathsf{T} y}.$$

For any vector w, the maximum of $\frac{(w^\mathsf{T} y)^2}{y^\mathsf{T} y}$ over nonzero vectors y is attained when y is a unit vector in the direction of w, in which case $\frac{(w^\mathsf{T} y)^2}{y^\mathsf{T} y} = w^\mathsf{T} w$. Substituting $w = (L_G^+)^{1/2} \delta_e$ we find that

$$\max \left\{ \frac{(\delta_e^\mathsf{T} x)^2}{x^\mathsf{T} L_G x} \middle| x \neq 0 \right\} = \left((L_G^+)^{1/2} \delta_e \right)^\mathsf{T} (L_G^+)^{1/2} \delta_e = \delta_e^\mathsf{T} L_G^+ \delta_e.$$

Substituting this into (9) and multiplying both sides by $\pi(e)$ we find that

$$R\pi(e) \geq c(e) \delta_e^\mathsf{T} L_G^+ \delta_e.$$

Summing over e,

$$R = R \left(\sum_e \pi(e) \right) \geq \sum_e c(e) \delta_e^\mathsf{T} L_G^+ \delta_e \tag{10}$$

$$= \operatorname{tr} \left(\sum_e c(e) \delta_e \delta^\mathsf{T} L_G^+ \right) \tag{11}$$

$$= \operatorname{tr} \left(L_G L_G^+ \right) = n - c, \tag{12}$$

where c denotes the number of connected components of G, and the last inequality follows from the fact that $L_G L_G^+$ is the projection on $\mathbb{R}^n$ onto the nullspace of L_G.

Our objective of minimizing R will be served if we make the inequality in line (10) tight, which means setting $R = n - c$ and $\pi(e) = \frac{1}{n-c} c(e) \delta_e^\mathsf{T} L_G^+ \delta_e$ for each edge. This choice of R and $\{\pi(e)\}$ is the optimal solution of the semidefinite program (8) and leads to a spectral sparsifier H with $k < 4n\varepsilon^{-2} \ln(4n)$ edges.

Incidentally, the quantity $c(e) \delta_e^\mathsf{T} L_G^+ \delta_e$ is called the *effective resistance of edge* e. It can be interpreted as the resistance between the endpoints of edge e, if one were to build an electrical network in the shape of the graph G, with each edge e' represented by a resistor of resistance $c(e')$. This connection between electrical networks and graph sparsification is just one of many beautiful connections between electrical networks, spectral graph theory, graph algorithms, and random walks. For more on this topic, see Doyle and Snell's short book, "Random Walks and Electric Networks".

8 Spectral Algorithms for Average-Case Problems

One of the common applications of spectral algorithms is to distinguish "signal" from "noise" in a dataset consisting of random vectors. In this section we'll illustrate this application by analyzing two problems.

Learning a mixture of two spherical Gaussians. In this problem one is given m independent, identically distributed random samples $x_1, x_2, \ldots, x_m$, where each x_i can be expressed as $a_i y + z_i$ where y is a fixed vector in $\mathbb{R}^n$, a_i is a uniformly random element of $\{\pm 1\}$, and z_i is drawn from the spherical Gaussian distribution $\mathcal{N}(0, \mathbb{1})$. In other words, the sampling distribution for x_i is a mixture of two Gaussians centered at y and $-y$, respectively. The objective of the algorithm is to output vectors close to the centers of the two distributions.

If $\|y\|$ is large enough, e.g. $\|y\| \gg n^{1/4}$, the data points cluster into two well-separated point clouds surrounding y and $-y$, and one can easily separate the clusters and average the samples in each cluster to obtain vectors that very closely approximate y and $-y$. When $\|y\|$ is smaller, e.g. $\|y\| = O(1)$, the points of the form $y + z_i$ and those of the form $-y + z_i$ occupy highly overlapping regions of space, and more sophisticated algorithms are required in order to find the locations of y and $-y$.

The "planted clique" problem. In this problem one is given a random graph G with n-element vertex set V, whose edge set is sampled by the following process: for some parameter $p < n$, a random p-element subset $P \subset V$ is chosen. Then for each unordered pair of distinct vertices u, v, we include edge u, v in the edge set of G with probability 1 if $\{u, v\} \subseteq P$, and with probability $\frac{1}{2}$ otherwise. In other words, we "plant" a clique on vertex set P in the graph, and every edge that is not in the clique is sampled independently with probability $\frac{1}{2}$.

If p is large enough, e.g. $p \gg \sqrt{n \log n}$, then an overwhelmingly large fraction of the p highest-degree vertices belong to P, and the remaining vertices of P can easily be located by searching for vertices having many more than $p/2$ neighbors among the p highest-degree vertices. When p is smaller, e.g. $p = \Theta(\sqrt{n})$, more sophisticated algorithms are required in order to find the planted clique.

The similarities between the two problems run deeper than the common theme of, "easy to solve when signal is sufficiently strong relative to noise, becoming harder when the signal is weaker." To be more specific, in both problems we can postprocess the input data to obtain a random symmetric positive semidefinite matrix whose expected value contains enough information to solve the problem. Designing and analyzing algorithms for both problems then becomes a matter of understanding when the "signal" in this random matrix stands out more strongly than the "noise". Since this involves bounding the difference between a random matrix and its expected value, it is not surprised that the Ahlswede-Winter Inequality will play a starring role in the analysis.

In the case of the Gaussian mixture problem, the matrix we need to study is the *empirical second moment matrix* of the dataset, namely the matrix

$$X = \frac{1}{m} \sum_{i=1}^{m} x_i x_i^{\mathsf{T}}.$$

It is easy to see that the expected value of this matrix is

$$\begin{aligned}\mathbb{E}X &= \frac{1}{m}\sum_{i=1}^{m} x_i x_i^{\mathsf{T}} \\ &= \mathbb{E}[x_i x_i^{\mathsf{T}}] \\ &= yy^{\mathsf{T}} + \mathbb{E}[a_i]\mathbb{E}[yz_i^{\mathsf{T}} + z_i y^{\mathsf{T}}] + \mathbb{E}[z_i z_i^{\mathsf{T}}] \\ &= yy^{\mathsf{T}} + \mathbb{1}.\end{aligned}$$

The eigenvalues and eigenvectors of $\mathbb{E}X$ are easy to understand. The rank-one matrix yy^{T} has eigenvector y with eigenvalue $\|y\|^2$, and every vector orthogonal to y is an eigenvector with eigenvalue zero. When we add the identity matrix, it changes the eigenvalues (by adding 1 to each) without changing the eigenvectors. Thus, y is an eigenvector of $\mathbb{E}X$ with eigenvalue $1+\|y\|^2$ and the orthogonal complement of y is an $(n-1)$-dimensional eigenspace with eigenvalue 1. Thus, if we were given the matrix $\mathbb{E}X$ without any noise, it would be easy to identify y and $-y$: the maximum eigenvalue of $\mathbb{E}X$ minus one tells us the length of y and $-y$, and the corresponding eigenvector tells us their direction.

When we examine the Laplacian L_G of the random graph in the planted-clique problem, we encounter a very similar story. Letting $A = (a_{uv})$ denote the adjacency matrix of the graph, the entries of A are random variables with expected values

$$\mathbb{E}a_{uv} = \begin{cases} 1 & \text{if } \{u,v\} \subseteq P \\ \frac{1}{2} & \text{if } \{u,v\} \not\subseteq P \end{cases} \tag{13}$$

The Laplacian matrix is

$$L_G = \sum_{u \neq v} a_{uv}\delta_{uv}\delta_{uv}^{\mathsf{T}},$$

hence

$$\begin{aligned}\mathbb{E}L_G &= \sum_{u\neq v} \mathbb{E}a_{uv}\delta_{uv}\delta_{uv}^{\mathsf{T}} \\ &= \frac{1}{2}\sum_{u\neq v\in V} \delta_{uv}\delta_{uv}^{\mathsf{T}} \;+\; \sum_{u\neq v\in P} \delta_{uv}\delta_{uv}^{\mathsf{T}} \\ &= \tfrac{1}{2}L_{K(V)} + \tfrac{1}{2}L_{K(P)}\end{aligned}$$

where $K(V)$ and $K(P)$ denotes cliques on the sets V and P, respectively. It is not hard to work out the eigenvalues and eigenvectors of $\frac{1}{2}L_{K(V)}$ and $\frac{1}{2}L_{K(P)}$. Let $\mathbf{1}_p$ denote the indicator vector of P (i.e., the vector with 1 or 0 in its u^{th} entry depending whether u belongs to P or not) and similarly let $\mathbb{1}_p$ denote the diagonal matrix whose (u,u) entry is 1 if $u \in P$ and 0 if not. We have

$$\begin{aligned}L_{K(V)} &= n\mathbb{1} - \mathbf{1}\mathbf{1}^{\mathsf{T}} \\ L_{K(P)} &= p\mathbb{1}_p - \mathbf{1}_p\mathbf{1}_p^{\mathsf{T}}\end{aligned}$$

The matrix $\mathbb{E}L_G = \frac{1}{2}L_{K(V)} + \frac{1}{2}L_{K(P)}$ is an average of graph Laplacians, so one of its eigenvectors is $\mathbf{1}$, with eigenvalue 0. There is an eigenspace of dimension $n-p$ consisting of vectors x such that $\mathbf{1}^\mathsf{T}x = 0$ and $x_u = x_v$ for all $u, v \in P$. Any such x satisfies $L_{K(V)}x = nx$ and $L_{K(P)}x = 0$, so $(\mathbb{E}L_G)x = \frac{1}{2}nx$. Finally, there is an eigenspace of dimension $p-1$ consisting of vectors x such that $\mathbf{1}_p^\mathsf{T}x = 0$ and $x_u = 0$ for all $u \notin P$. Any such x satisfies $L_{K(V)}x = nx$ and $L_{K(P)}x = px$, so $(\mathbb{E}L_G)x = \frac{1}{2}(n+p)x$. In summary, the eigenvalues of $\mathbb{E}L_G$ are $0, \frac{1}{2}n, \frac{1}{2}(n+p)$ with multiplicities $1, n-p, p-1$, respectively, and given any non-zero vector in the eigenspace

From the formulae for $L_{K(V)}$ and $L_{K(P)}$, one can see that if x is a vector supported on P that is orthogonal to $\mathbf{1}$ (hence also orthogonal to $\mathbf{1}_p$) then $\frac{1}{2}(L_{K(V)} + L_{K(P)})x = \frac{n+p}{2}x$. Similarly, if x is a vector supported on $V \setminus P$ that is orthogonal to $\mathbf{1}$ then $\frac{1}{2}(L_{K(V)}+L_{K(P)})x = \frac{n}{2}x$. These two eigenspaces have dimensions $p-1$ and $n-p-1$, respectively, and together they span the $(n-2)$-dimensional space of vectors orthogonal to both $\mathbf{1}$ and $\mathbf{1}_p$. At this point we have discovered eigenspaces of dimensions $1, p-1, n-p-1$, so there's one remaining one-dimensional eigenspace and it is orthogonal to the ones we've already enumerated. The eigenspaces of dimensions $p-1$ and $n-p-1$ span the orthogonal complement of $\langle \mathbf{1}, \mathbf{1}_p \rangle$, so the remaining eigenspace belongs to the span of $\mathbf{1}$ and $\mathbf{1}_p$. In the two-dimensional space spanned by $\mathbf{1}$ and $\mathbf{1}_p$, the vectors orthogonal to $\mathbf{1}$ are the scalar multiples of $p\mathbf{1} - n\mathbf{1}_p$. The eigenvalue associated to this eigenspace is $\frac{n+p}{2}$.

A Additional tools for working with symmetric matrices

This appendix contains some additional tools that are useful in the design and analysis of algorithms involving symmetric matrices.

A.1 Standard matrix functions

There is a standard way of extending any function that maps $\mathbb{R}$ to $\mathbb{R}$ into a function mapping $\mathrm{Sym}_n(\mathbb{R})$ to $\mathrm{Sym}_n(\mathbb{R})$. In this section we define these "standard matrices functions" and present some basic examples and properties.

Definition A.1. If $f : \mathbb{R} \to \mathbb{R}$ is any function, the matrix extension of f is the unique function from $\mathrm{Sym}_n(\mathbb{R})$ to $\mathrm{Sym}_n(\mathbb{R})$ satisfying $f(\mathrm{diag}(\lambda_1, \ldots, \lambda_n)) = \mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_n))$ and $f(QDQ^{-1}) = Qf(D)Q^{-1}$ for every orthogonal matrix Q and diagonal matrix D.

The only subtlety in the definition of the matrix extension of f is that any given $A \in \mathrm{Sym}_n(\mathbb{R})$ can be written as QDQ^{-1} in more than one way, and one needs to verify that the definition of $f(A)$ does not depend on the choice of representation $A = QDQ^{-1}$. We leave this verification to the reader.

Some immediate consequences of the definition are the following.

1. If $A \in \mathrm{Sym}_n(\mathbb{R})$ has eigenvalues $\lambda_1, \ldots, \lambda_n$ and corresponding eigenvectors $x_1, \ldots, x_n$, then $f(A)$ has eigenvalues $f(\lambda_1), \ldots, f(\lambda_n)$ and corresponding eigenvectors $x_1, \ldots, x_n$.

2. If A is symmetric and Q is orthogonal, then $f(QAQ^{-1}) = Qf(A)Q^{-1}$.

3. If f and g are two functions from $\mathbb{R}$ to $\mathbb{R}$ and $f \circ g$, $f + g$, $f \cdot g$ are the functions defined by

$$(f \circ g)(\lambda) = f(g(\lambda)), \quad (f + g)(\lambda) = f(\lambda) + g(\lambda), \quad (f \cdot g)(\lambda) = f(\lambda)\, g(\lambda)$$

   then for all $a \in \mathrm{Sym}_n(\mathbb{R})$,

$$(f \circ g)(A) = f(g(A)), \quad (f + g)(A) = f(A) + g(A), \quad (f \cdot g)(A) = f(A)\, g(A).$$

The following are some useful properties and examples.

1. For any two functions f, g and any matrix $A \in \mathrm{Sym}_n(\mathbb{R})$, the matrices $f(A)$ and $g(A)$ commute with one another. This is due to the identity $f \cdot g = g \cdot f$.

2. If f is a polynomial function $f(\lambda) = \sum_{i=0}^{m} c_i \lambda^i$, then $f(A) = \sum_{i=0}^{n} c_i A^i$, where A^0 is interpreted as the identity matrix.

3. If f is represented by a power series $f(\lambda) = \sum_{i=0}^{\infty} c_i \lambda^i$ that converges on the open interval $(-R, R)$, then $f(A) = \sum_{i=0}^{\infty} c_i A^i$ for every matrix A whose eigenvalues are all contained in the interval $(-R, R)$.

4. An important special case of the preceding example is the matrix exponential function, defined by

$$e^A = \sum_{i=0}^{\infty} \frac{1}{i!} A^i.$$

5. If f is the function

$$f(\lambda) = \begin{cases} \lambda^{-1} & \text{if } \lambda \neq 0 \\ 0 & \text{if } \lambda = 0 \end{cases}$$

   then $f(A)$ is denoted by A^+ and is called the Moore-Pensore pseudoinverse (or, simply, pseudoinverse) of A. When A is invertible, A^+ is the inverse of A. More generally, $A^+ A = A\, A^+$ is the matrix that represents the orthogonal projection of $\mathbb{R}^n$ onto the column space of A.

A.2 The Golden-Thompson Inequality

Theorem A.2 (Golden-Thompson Inequality)**.** *For any matrices $A, B \in \mathrm{Sym}_n(\mathbb{R})$,*

$$\mathrm{tr}\left(e^A e^B\right) \geq \mathrm{tr}\left(e^{A+B}\right).$$

A.3 The PSD ordering on symmetric matrices

Let $\mathrm{Sym}_n(\mathbb{R})$ denote the vector space of n-by-n symmetric matrices over $\mathbb{R}$. If $A - B$ is positive semidefinite we write $A \succeq B$ or $B \preceq A$. This relation is a partial order: it is reflexive (the zero matrix is PSD), antisymmetric (if a matrix and its negation are PSD, it is the zero matrix), and transitive (the sum of two PSD matrices is PSD).

A.4 The Ahlswede-Winter Inequality

The Ahlswede-Winter Inequality is a counterpart of the Chernoff bound, for sums of independent random symmetric PSD matrices rather than sums of independent random scalars.

Theorem A.3 (Ahlswede-Winter Inequality)**.** *Suppose $X_1, X_2, \ldots, X_k$ are mutually independent random, symmetric, positive semidefinite $n \times n$ matrices, and let $U = \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^k X_i\right]$. If $R \geq 1$ is a scalar such that for all i, $X_i \preceq R \cdot U$ with probability 1, then for all $\varepsilon \in (0,1)$,*

$$\Pr\left[(1-\varepsilon)U \preceq \tfrac{1}{k}\sum_{i=1}^k X_i \preceq (1+\varepsilon)U\right] \;\geq\; 1 - 2n \cdot \exp\left(-\tfrac{\varepsilon^2 k}{4R}\right). \tag{14}$$

The proof is similar to the proof of the Chernoff bound. Letting X denote the random sum $\sum_{i=1}^k X_i$, the standard proof of the Chernoff bound uses the exponential generating function $\Phi(t) = \mathbb{E}[e^{tX}]$. Here, the expression e^{tX} is matrix-valued. To make Φ scalar-valued we instead use $\Phi(t) = \mathbb{E}[\mathrm{tr}(e^{tX})]$. This accounts for the extra factor of n in the failure probability on the right side of (14). (The trace of the identity matrix is n, not 1.) The main difficulty that arises in the proof of Ahlswede-Winter, relative to the proof of Chernoff, is that the exponentials of non-commuting matrices do not commute, so the identity $e^{tX} = \prod_{i=1}^n e^{tX_i}$ does not hold. However, the Golden-Thompson Inequality justifies the inequality

$$\mathrm{tr}(e^{tX}) \leq \mathrm{tr}\left(\prod_{i=1}^n e^{tX_i}\right)$$

which is good enough to complete the proof.

Proof. The proof begins by reducing to the case where U is the identity matrix. Since U is symmetric positive semidefinite, it can be written as QDQ^{-1}, where Q is orthogonal and D is a diagonal matrix with non-negative entries arranging in non-increasing order. Replacing each X_i with QX_iQ^{-1}, we can assume henceforth that $U = D$. If the nullspace of D has dimension $d > 0$, then the entries in the final d rows and columns of D are all equal to zero. The relation $X_i \preceq R \cdot D$ implies that any vector in the nullspace of D must also belong to the nullspace of X_i for each i. Thus, for each i, the entries in the final d rows and columns of X_i are all equal to zero. To prove a lower bound on the probability of the event $(1-\varepsilon)D \preceq \frac{1}{k}\sum_{i=1}^k X_i \preceq (1+\varepsilon)D$, it suffices to confine our attention to the square submatrices occupying the first $n - d$ rows and columns of each matrix involved.

Having thus reduced to the case that U is a non-singular diagonal matrix D, we may replace each X_i with $D^{-1/2}X_iD^{-1/2}$ to finally reduce to the case when $\mathbb{E}[\frac{1}{k}\sum_{i=1}^{k} X_i]$ is the identity matrix, $\mathbb{1}$.

Now let $X = \frac{1}{k}\sum_{i=1}^{k} X_i$ and consider the function

$$\Phi(t) = \mathbb{E}\left[\mathrm{tr}\left(e^{tX}\right)\right].$$

□