ORIE 6334 Spectral Graph Theory October 27, 2016

Lecture 19

Lecturer: David P. Williamson *Scribe: Faisal Alkaabneh*

1 Chernoff bounds

Today we will look at a matrix analog of the standard scalar Chernoff bounds. This matrix analog will be used in the next lecture when we talk about graph sparsification. While we're more interested in the application of the theorem than its proof, it's still useful to see the similarities and the differences of moving from the proof of the result for scalars to the same result for matrices.

Scalar Chernoff bounds get used all over the place in theoretical computer science, so much so that one now sees the phrase "By a standard application of Chernoff bounds..." without even seeing the theorem references or the random variables defined.

Theorem 1 ((Scalar) Chernoff bound) *Let $X_1, X_2, ..., X_k$ independent random variables with $X_i \in [0, R]$. Let $\mu_{\min} \leq E\left[\sum_{i=1}^k X_i\right] \leq \mu_{\max}$. Then for all $\delta \geq 0$,*

$$\Pr\left[\sum_{i=1}^k X_i \geq (1+\delta)\mu_{\max}\right] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R}$$
$$\Pr\left[\sum_{i=1}^k X_i \leq (1-\delta)\mu_{\min}\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu_{\min}/R}.$$

By a simple application of calculus, one can further show that if $\delta \leq 1$,

$$\Pr\left[\sum_{i=1}^k X_i \geq (1+\delta)\mu_{\max}\right] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R} \leq e^{-\delta\mu_{\max}/3R}$$
$$\Pr\left[\sum_{i=1}^k X_i \leq (1-\delta)\mu_{\min}\right] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu_{\min}/R} \leq e^{-\delta^2\mu_{\max}/2R}$$

The corresponding matrix version looks remarkably similar, except for the dimension d of the matrix appearing in the bound.

[0]This lecture is extremely indebted to a lecture of Nick Harvey at the Sixth Cargèse Workshop on Combinatorial Optimization, `http://www.cs.ubc.ca/~nickhar/Cargese2.pdf`, as will be completely obvious if you go look at his notes.

Theorem 2 (Tropp 2011) *Let $X_1, X_2, ..., X_k$ be random symmetric $d \times d$ matrices. s.t. $0 \preceq X_i \preceq R \cdot I$ for some constant R. Suppose $\mu_{\min} I \preceq \sum_{i=1}^{k} E[X_i] \preceq \mu_{\max} I$. Then for all $\delta \geq 0$,*

$$\Pr\left[\lambda_{\max}\left(\sum_{i=1}^{k} X_i\right) \geq (1+\delta)\mu_{\max}\right] \leq d \cdot \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R}$$
$$\Pr\left[\lambda_{\min}\left(\sum_{i=1}^{k} X_i\right) \leq (1-\delta)\mu_{\min}\right] \leq d \cdot \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu_{\min}/R}.$$

Then if $\delta \leq 1$,

$$\Pr\left[\lambda_{\max}\left(\sum_{i=1}^{k} X_i\right) \geq (1+\delta)\mu_{\max}\right] \leq d \cdot \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R} \leq d \cdot e^{-\delta\mu_{\max}/3R}$$
$$\Pr\left[\lambda_{\min}\left(\sum_{i=1}^{k} X_i\right) \leq (1-\delta)\mu_{\min}\right] \leq d \cdot \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu_{\min}/R} \leq d \cdot e^{-\delta^2\mu_{\max}/2R}$$

2 Proof of scalar version

We now walk through the proof of the scalar version in order to set up the similarities of the proof of the matrix version. We will use the following.

Theorem 3 (Markov's inequality) *For X a random variable s.t. $X \geq 0$,*

$$\Pr[X \geq a] \leq \frac{E[X]}{a}.$$

Proof: Since X is nonnegative, we have that

$$E[X] \geq a \cdot \Pr[X \geq a].$$

□

We will show the following two claims.

Claim 4 *For any $\theta > 0$,*

$$\Pr\left[\sum_{i=1}^{k} X_i \geq t\right] \leq e^{-\theta t} \prod_{i=1}^{k} E\left[e^{\theta X_i}\right].$$

Claim 5 *For X random variable $X \in [0,1]$,*

$$E\left[e^{\theta X}\right] \leq 1 + (e^{\theta} - 1)E[X].$$

Now let us prove the claims, and show how they imply the scalar Chernoff bound.

Proof of Claim 4:

$$\begin{aligned}
\Pr\left[\sum_{i=1}^{k} X_i \geq t\right] &= \Pr\left[\sum_{i=1}^{k} \theta X_i \geq \theta t\right] \\
&= \Pr\left[\exp\left(\sum_{i=1}^{k} \theta X_i\right) \geq \exp(\theta t)\right] \\
&\leq e^{-\theta t} \cdot E\left[\exp\left(\sum_{i=1}^{k} \theta X_i\right)\right] \qquad \text{by Markov's inequality} \\
&= e^{-\theta t} \cdot E\left[\prod_{i=1}^{k} \exp(\theta X_i)\right] \\
&= e^{-\theta t} \prod_{i=1}^{k} E\left[\exp(\theta X_i)\right],
\end{aligned}$$

where the last equality follows by the independence of the X_i. □

Proof of Claim 5: Because $e^{\theta x}$ is a convex function, we know that $e^{\theta x} \leq 1 + (e^{\theta} - 1)x$ for $x \in [0, 1]$. Hence,

$$E\left[e^{\theta x}\right] \leq 1 + (e^{\theta} - 1)E\left[x\right].$$

□

Proof of Theorem 1: We'll only prove the upper bound, with $R = 1$. We can show that

$$\begin{aligned}
\prod_{i=1}^{k} E\left[\sum_{i=1}^{k} e^{\theta X_i}\right] &\leq \prod_{i=1}^{k} \Big(1 + (e^{\theta} - 1)E\left[X_i\right]\Big) \qquad \text{by Claim 5} \\
&= \exp\left(\sum_{i=1}^{k} \log(1 + (e^{\theta} - 1)E\left[X_i\right])\right) \\
&\leq \exp\left(\sum_{i=1}^{k} (e^{\theta} - 1)E\left[X_i\right]\right) \qquad \text{using } \log(1 + x) \leq x \\
&\leq \exp[(e^{\theta} - 1)\mu_{\max}].
\end{aligned}$$

Applying Claim 4 to the above with with $t = (1+\delta)\mu_{\max}, \theta = \ln(1+\delta)$, we get

$$\Pr\left[\sum_{i=1}^{k} X_i \geq (1+\delta)\mu_{\max}\right] \leq \exp[-\ln(1+\delta)\cdot(1+\delta)\mu_{\max}]\cdot\exp(\delta\mu_{\max})$$
$$= \Big(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\Big)^{\mu_{\max}}.$$

□

3 Concepts and facts needed for matrix version

We'd like to extend the same logic as used in the scalar version to matrices, but it isn't clear that some of the concepts we've used will carry through, like the convexity of matrices, or what the log of a matrix even is. So before diving into the proof, we need to think about what concepts extend and which ones don't.

We first define the concepts of a *spectral mapping*. For function f, symmetric matrix A, with eigenvalues λ_i and orthonormal eigenvectors $x_1, x_2, ..., x_n$, recall that

$$A = \sum_{i=1}^{k} \lambda_i x_i x_i^T = XDX^T$$

for $D = \text{diag}(\lambda_i)$, where X has x_i as its ith column. Then we define $f(A) \equiv Xf(D)X^T$ where $f(D)_{ii} = f(D_{ii})$; for example, for $f(x) = x^k$, $f(A) = f(A^k) = XD^kX^T$, as we've already seen. We can extend concepts of monotonicity and concavity to matrices as follows.

Definition 1 *Function f is* operator monotone *if $A \succeq B$ implies that $f(A) \succeq f(B)$. f is* operator concave *if*

$$f((1-\alpha)A + \alpha B) \succeq (1-\alpha)f(A) + \alpha f(B)$$

for $\alpha \in [0,1]$ and for all A, B.

Sadly, f monotone does not imply that f is operator monotone, and f concave does not imply that f is operator concave. ☹

To get around these problems we will use a careful combination of things that are known to hold. Some facts we will need:

Fact 1 *If $f(x) \leq g(x)$ for all $x \in [l,u]$, then for symmetric A with $\lambda_{\min}(A) \geq l, \lambda_{\min}(A) \leq u$ then $f(A) \preceq g(A)$.*

Fact 2 *If X, Y are random matrices and $X \preceq Y$, then $E[X] \preceq E[Y]$.*

Fact 3 (Weyl monotonicity) *For $A, B \in \mathbb{R}^{d\times d}$ symmetric, $A \preceq B$, $\lambda_i(A) \leq \lambda_i(B)$ for all i.*

Corollary 6 *If f is monotone, then $A \preceq B$ implies $\mathrm{tr}(f(A)) \leq \mathrm{tr}(f(B))$.*

Fact 4 log *is operator concave.*

Recall that $A \succ 0$ implies that A is positive definite (that is, for symmetric A, A has all positive eigenvalues).

Definition 2 *If $A, B \succ 0$, then $A \odot B \equiv \exp(\log(A) + \log(B))$.*

The following fact is the key ingredient to the proof.

Fact 5 *If $A_1, A_2, ..., A_k$ independent random matrices such that $A_i \succ 0$ for all i, then*

$$E\left[\mathrm{tr}(A_1 \odot A_2 \odot ... \odot A_k)\right] \leq \mathrm{tr}\left(E(A_1) \odot E(A_2) \odot ... \odot E(A_k)\right).$$

4 Proof of matrix version

We can now give the proof of the matrix version of Chernoff bounds. We state the following two claims so that the parallels with Claims 4 and 5 are clear.

Claim 7 *For any $\theta > 0$,*

$$\Pr\left[\lambda_{\max}(\sum_{i=1}^{k} X_i) \geq t\right] \leq e^{-\theta t} \cdot \mathrm{tr}\left[\bigodot_{i=1}^{k} E\left[e^{\theta X_i}\right]\right].$$

Claim 8 *Let X be random symmetric matrix such that $0 \preceq X \preceq I$. Then*

$$E\left[e^{\theta X}\right] \preceq I + (e^{\theta} - 1)E\left[X\right].$$

We now give the proofs of these claims.

Proof of Claim 7: The proof is somewhat analogous to the Proof of Claim 4.

$$\begin{aligned}
\Pr\left[\lambda_{\max}\left(\sum_{i=1}^{k} X_i\right) \geq t\right] &= Pr\left[\lambda_{\max}\left(\sum_{i=1}^{k} \theta X_i\right) \geq \theta t\right] \\
&= \Pr\left[\exp\left(\lambda_{\max}\left(\sum_{i=1}^{k} \theta X_i\right)\right) \geq \exp(\theta t)\right] \\
&\leq e^{-\theta t} \cdot E\left[\exp\left(\lambda_{\max}\left(\sum_{i=1}^{k} \theta X_i\right)\right)\right] \qquad \text{by Markov's inequality.}
\end{aligned}$$

We notice that

$$\exp\left(\lambda_{\max}\left(\sum_{i=1}^{k} \theta X_i\right)\right) = \lambda_{\max}\left(\exp\left(\sum_{i=1}^{k} \theta X_i\right)\right) \leq \operatorname{tr}\left(\exp\left(\sum_{i=1}^{k} \theta X_i\right)\right),$$

by the properties of spectral mapping, and because the trace dominates the maximum eigenvalue given that the matrices are positive semidefinite. Then

$$\begin{aligned}
\Pr\left[\lambda_{\max}\left(\sum_{i=1}^{k} X_i\right) \geq t\right] &\leq e^{-\theta t} \cdot E\left[\operatorname{tr}\left(\exp\left(\sum_{i=1}^{k} \theta X_i\right)\right)\right] \\
&= e^{-\theta t} \cdot E\left[\operatorname{tr}\left(\exp\left(\sum_{i=1}^{k} \log e^{\theta X_i}\right)\right)\right] \\
&= e^{-\theta t} \cdot E\left[\operatorname{tr}(e^{-\theta X_1} \odot e^{-\theta X_2} \odot ... \odot e^{-\theta X_k})\right] \qquad \text{by definition of } \odot \\
&\leq e^{-\theta t} \cdot \operatorname{tr}\Big(E\left[e^{-\theta X_1}\right] \odot E\left[e^{-\theta X_2}\right] \odot ... \odot E\left[e^{-\theta X_k}\right]\Big) \qquad \text{by Fact 5.}
\end{aligned}$$

□

Proof of Claim 8: The proof is analogous to the proof of Claim 5. For $x \in [0,1]$, $e^{\theta x} \leq 1 + (e^{\theta} - 1)x$ by convexity. Since $0 \preceq X \preceq I$, $\lambda_{\min}(x) \geq 0$ and $\lambda_{\max}(x) \leq 1$. Then by Fact 1,

$$e^{\theta X} \leq I + (e^{\theta} - 1)X.$$

Then by Fact 2,

$$E\left[e^{\theta X}\right] \leq I + (e^{\theta} - 1)E\left[X\right].$$

□

We can now complete the proof of the Matrix Chernoff bound.

Proof of Theorem 2: We only prove the upper bound, and we assume $R = 1$. Then by the operator concavity of the log function,

$$\sum_{i=1}^{k} \log E\left[e^{\theta X_i}\right] = k \sum_{i=1}^{k} \frac{1}{k} \log E\left[e^{\theta X_i}\right] \preceq k \log\left(\sum_{i=1}^{k} \frac{1}{k} E\left[e^{\theta X_i}\right]\right). \qquad (1)$$

Then we have that

$$
\begin{aligned}
&\operatorname{tr}\left(E\left[e^{\theta X_1}\right] \odot \ldots \odot E\left[e^{\theta X_k}\right]\right) \\
&\quad = \operatorname{tr}\left(\exp\left(\sum_{i=1}^{k} \log\left(E\left[e^{\theta X_i}\right]\right)\right)\right) && \text{by definition of } \odot \\
&\quad \leq \operatorname{tr}\left(\exp\left(k \log\left(\sum_{i=1}^{k} \frac{1}{k} E\left[e^{\theta X_i}\right]\right)\right)\right) && \text{by 1 and Corollary 6} \\
&\quad \leq d \cdot \lambda_{\max}\left(\exp\left(k \log\left(\sum_{i=1}^{k} \frac{1}{k} E\left[e^{\theta X_i}\right]\right)\right)\right) && \text{since trace is at most } d\lambda_{\max} \\
&\quad = d \cdot \exp\left(k \log \lambda_{\max}\left(\sum_{i=1}^{k} \frac{1}{k} E\left[e^{\theta X_i}\right]\right)\right) && \text{by operator mapping} \\
&\quad \leq d \cdot \exp\left(k \log \lambda_{\max}\left(I + (e^{\theta} - 1) \sum_{i=1}^{k} \frac{1}{k} E\left[e^{\theta X_i}\right]\right)\right) && \text{by Claim 8 and Weyl Monotonicity} \\
&\quad = d \cdot \exp\left(k \log\left(1 + \frac{1}{k}(e^{\theta} - 1) \cdot \lambda_{\max}\left(\sum_{i=1}^{k} E\left[X_i\right]\right)\right)\right) \\
&\quad \leq d \cdot \exp\left(\left(e^{\theta} - 1\right) \mu_{\max}\right) && \text{using } \log(1+x) \leq x.
\end{aligned}
$$

Then by plugging in the above to 7 and using $t = (1+\delta)\mu_{\max}$, $\theta = \ln(1+\delta)$, we get

$$
\begin{aligned}
\Pr\left[\lambda_{\max}\left(\sum_{i=1}^{k} X_i\right) \geq (1+\delta)\mu_{\max}\right] &\leq d \cdot \exp\left(-\ln(1+\delta)(1+\delta)\mu_{\max}\right) \exp(\delta\mu_{\max})) \\
&= d \cdot \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}}.
\end{aligned}
$$

□

In the next lecture, we will apply this result to sampling edges from graphs to get sparse versions of the graph.