CS 388R: Randomized Algorithms, Fall 2023 10/18/2023

Lecture 16: Matrix concentration and graph sparsification

Prof. Eric Price *Scribe: Steven Xu, Bennett Liu*

NOTE: THESE NOTES HAVE NOT BEEN EDITED OR CHECKED FOR CORRECTNESS

1 Overview

In the previous lecture, learned about online bipartite matching. This lecture, we will develop some background required for graph sparsification. In particular, we will try to prove the Rudelson-Vershyni theorem by using an extension of Bernstein's inequality for symmetric matrices.

2 Bernstein Concentration Inequality

The Bernstein Concentration Inequality is a concentration inequality for the sum of bounded independent real random variables—similar to the Chernoff bounds we use more commonly, but taking into account the variance as well as the boundedness of our variables.

Claim 1. Bernstein Concentration Inequality. *Suppose $X_1, \ldots, X_n$ are independent, centered random variables where $|X_i| \leq K$ for all i, and let*

$$X = \sum_{i=1}^{n} X_i, \quad \sigma_i^2 = \mathrm{Var}[X_i], \quad \sigma^2 = \mathrm{Var}[X] = \sum_{i=1}^{n} \sigma_i^2.$$

Then

$$\mathbb{P}\left[X \geq t\right] \leq \exp\left(-\frac{1}{4}\min\left(\frac{t^2}{\sigma^2}, \frac{t}{K}\right)\right).$$

Intuitively, this inequality states that X behaves like a normal distribution around the mean and an exponential distribution further out. This idea will be important to proving the inequality.

2.1 Proof of Bernstein's Inequality

Author's note: the proof was not covered in class.

First, we capture the notion of "normal around the mean and exponential on the tail" in the idea of a subgamma variable.

Definition 2. Subgamma Random Variable. *A centered variable X is* subgamma(σ^2, c) *with variance proxy σ^2 and exponential scale c if*

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{1}{2}\sigma^2\lambda^2\right) \text{ for all } |\lambda| \leq \frac{1}{c}$$

While this definition doesn't seem to correspond to the behavior we're trying to model, it turns out they are roughly equivalent.

Proposition 3. *If a random variable X is* $\mathrm{subgamma}(\sigma^2, c)$ *then*

$$\max(\mathbb{P}[X \geq t], \mathbb{P}[X \leq -t]) \leq \exp\left[-\frac{1}{2}\min\left(\frac{t^2}{\sigma^2}, \frac{t}{c}\right)\right].$$

The converse also holds with a loss in parameters.

Proof. We'll only prove the forward direction since we don't actually need the converse anywhere. For those who've seen a proof of the Chernoff bounds, this will be very similar. Suppose X is $\mathrm{subgamma}(\sigma^2, c)$. By Markov's inequality, we know that for $|\lambda| \leq \frac{1}{c}$,

$$\mathbb{P}[X \geq t] = \mathbb{P}[\exp(\lambda X) \leq \exp(\lambda t)] \leq \mathbb{E}[\exp(\lambda X)]\exp(-\lambda t) \leq \exp\left(\frac{1}{2}\sigma^2\lambda^2 - t\lambda\right).$$

To get the tightest inequality possible, we choose λ to minimize the convex quadratic $f(\lambda) = \frac{1}{2}\sigma^2\lambda^2 - t\lambda$. We know that f achieves its minimum at $\lambda = -(-t)/(2(\sigma^2/2)) = t\sigma^{-2}$. However, that value might be greater than $\frac{1}{c}$, so in that case, we take $\lambda = \frac{1}{c}$. Finally, our minimum value of f is

$$\begin{aligned}
f\left(\min\left(\frac{t}{\sigma^2}, \frac{1}{c}\right)\right) &= \frac{1}{2}\sigma^2\left[\min\left(\frac{t}{\sigma^2}, \frac{1}{c}\right)\right]^2 - t\min\left(\frac{t}{\sigma^2}, \frac{1}{c}\right) \\
&= \frac{1}{2}\min\left(\frac{t^2}{\sigma^2}, \frac{\sigma^2}{c^2}\right) - \min\left(\frac{t^2}{\sigma^2}, \frac{t}{c}\right) \\
&= \frac{1}{2}\min\left(\frac{t^2}{\sigma^2}, \frac{1}{c}\frac{\sigma^2}{c}\right) - \min\left(\frac{t^2}{\sigma^2}, \frac{t}{c}\right) \\
&\leq \frac{1}{2}\min\left(\frac{t^2}{\sigma^2}, \frac{t}{c}\right) - \min\left(\frac{t^2}{\sigma^2}, \frac{t}{c}\right) \\
&= -\frac{1}{2}\min\left(\frac{t^2}{\sigma^2}, \frac{t}{c}\right).
\end{aligned}$$

The inequality holds since if $t^2/\sigma^2 \geq \sigma^2/c^2$, then $t \geq \sigma^2/c$. Finally,

$$\mathbb{P}[X \geq t] \leq \exp\left(\frac{1}{2}\sigma^2\lambda^2 - t\lambda\right) \leq \exp\left(-\frac{1}{2}\min\left(\frac{t^2}{\sigma^2}, \frac{t}{c}\right)\right).$$

The proof so far only bounds the positive tail of X. To bound the negative tail of X, observe that if X is subgamma, then $-X$ is subgamma with the same parameters. □

Now we can show Bernstein's inequality by proving that the sum X is $\mathrm{subgamma}(2\sigma^2, 2K)$. To do this, we'll first show that each X_i is $\mathrm{subgamma}(2\sigma_i^2, 2K)$, and then show that the sum of subgamma variables is subgamma.

Proposition 4. *Let X be a random variable such that $|X| \leq K$, and let $\sigma^2 = \mathrm{Var}[X]$. Then X is* $\mathrm{subgamma}(2\sigma^2, 2K)$.

Proof. Let $|\lambda| \leq 1/[2K]$. Note that $|\lambda X| \leq 1/2$.

To help untangle the soup of upcoming equations, here's the gist of the proof. Read this alongside the equations.

1. We do Taylor series expansion on $\mathbb{E}[\exp(\lambda X)]$ to get a polynomial in the moments of X.
2. With some manipulation, we turn those into moments of $|X/K| \leq 1$, so we're guaranteed that higher moments get smaller. This allows us to replace the second moment onwards with second moments.
3. After pulling some terms out, we bound our remaining series with a geometric series, then bound that with a constant using the fact that $|\lambda| \leq 1/[2K]$.
4. Once the dust settles, we're left with the first two terms of the Taylor series expansion of $\exp(\lambda^2\sigma^2)$, which must be at most $\exp(\lambda^2\sigma^2)$ itself since every term is positive.

$$\begin{aligned}
\mathbb{E}[\exp(\lambda X)] \leq \mathbb{E}[\exp(|\lambda X|)] &= \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{1}{n!}|\lambda X|^n\right] = \sum_{n=0}^{\infty} \frac{|\lambda|^n}{n!}\,\mathbb{E}[|X|^n] && (1)\\
&= 1 + \sum_{n=2}^{\infty} \frac{|\lambda K|^n}{n!}\,\mathbb{E}\left[\left|\frac{X}{K}\right|^n\right] \leq 1 + \sum_{n=2}^{\infty} \frac{|\lambda K|^n}{n!}\,\mathbb{E}\left[\left|\frac{X}{K}\right|^2\right] && (2)\\
&= 1 + \sum_{n=2}^{\infty} \frac{|\lambda K|^n}{n!}\frac{\sigma^2}{K^2}\\
&= 1 + |\lambda K|^2 \frac{\sigma^2}{K^2}\sum_{n=0}^{\infty} \frac{|\lambda K|^n}{(n+2)!} \leq 1 + \frac{1}{2}\lambda^2\sigma^2 \sum_{n=0}^{\infty} |\lambda K|^n && (3)\\
&= 1 + \frac{1}{2}\lambda^2\sigma^2 \frac{1}{1-|\lambda K|} \leq 1 + \lambda^2\sigma^2\\
&\leq \exp(\lambda^2\sigma^2). && (4)
\end{aligned}$$

Thus X is subgamma$(2\sigma^2, 2K)$. □

Now we will show that the sum of two subgamma variables is subgamma.

Proposition 5. *Suppose X_1, X_2 are independent random variables s.t. X_1 is* subgamma(σ_1^2, c_1) *and X_2 is* subgamma(σ_2^2, c_2)*. Then the sum $X_1 + X_2$ must be* subgamma$(\sigma_1^2 + \sigma_2^2, \max(c_1, c_2))$.

Proof. Let $\lambda \leq 1/\max(c_1, c_2)$. Then $\lambda \leq 1/c_1$ and $\lambda \leq 1/c_2$. Using the subgamma property of X_1, X_2, we see that

$$\begin{aligned}
\mathbb{E}[\exp(\lambda(X_1 + X_2))] &= \mathbb{E}[\exp(\lambda X_1)]\,\mathbb{E}[\exp(\lambda X_2)]\\
&\leq \exp\left(\frac{1}{2}\sigma_1^2\lambda^2\right)\exp\left(\frac{1}{2}\sigma_2^2\lambda^2\right)\\
&= \exp\left(\frac{1}{2}(\sigma_1^2 + \sigma_2^2)\lambda^2\right).
\end{aligned}$$

Thus $X_1 + X_2$ is subgamma$(\sigma_1^2 + \sigma_2^2, \max(c_1, c_2))$. □

Finally, we can prove Bernstein's inequality.

Theorem 6. Bernstein Concentration Inequality. *Suppose $X_1, \ldots, X_n$ are independent, centered random variables where $|X_i| \leq K$ for all i, and let*

$$X = \sum_{i=1}^{n} X_i, \quad \sigma_i^2 = \mathrm{Var}[X_i], \quad \sigma^2 = \mathrm{Var}[X] = \sum_{i=1}^{n} \sigma_i^2.$$

Then

$$\mathbb{P}\left[X \geq t\right] \leq \exp\left(-\frac{1}{4}\min\left(\frac{t^2}{\sigma^2}, \frac{t}{K}\right)\right).$$

Proof. We know that X_i is subgamma$(2\sigma_i^2, 2K)$ and, so X must be subgamma$(2\sum_{i=1}^{n}\sigma_i^2, 2K)$. But since $\sigma^2 = \sum_{i=1}^{n}\sigma_i^2$, X is subgamma$(2\sigma^2, 2K)$. Finally,

$$\mathbb{P}[X \geq t] \leq \exp\left[-\frac{1}{2}\min\left(\frac{t^2}{2\sigma^2}, \frac{t}{2K}\right)\right] = \exp\left[-\frac{1}{4}\min\left(\frac{t^2}{\sigma^2}, \frac{t}{K}\right)\right].$$

□

3 Matrix Bernstein

This section will attempt to extend Bernstein's inequality to symmetric matrices. First, we will do a quick review of matrix norms.

3.1 Matrix Norms

Definition 7. Spectral Norm. *The spectral norm of an n by n matrix A is*

$$\|A\| = \max_{i\in[n]} \sigma_n \text{ where } \{\sigma_i\}_{i\in[n]} \text{ are the singular values of } A$$

Definition 8. Operator Norm. *The operator norm of an n by n matrix A is*

$$\|A\|_{\mathrm{op}} = \sup_{v\in\mathbb{R}^n\setminus\{0\}} \frac{\|Av\|}{\|v\|}.$$

In other words, $\|A\|_{\mathrm{op}}$ is the max factor A will increase the length of a vector will increase by.

For a symmetric matrix A, the singular values coincide with the absolute value of the eigenvalues— if $Q\Lambda Q^T$ is an eigen-decomposition of A, then $A^TA = A^2 = (Q\Lambda Q^T)(Q\Lambda Q^T) = Q\Lambda^2 Q^T$, so the singular values are $\sigma = \sqrt{\lambda^2} = |\lambda|$ where λ is an eigenvalue of A.

For any matrix A, the operator norm and the spectral norm are equal. Let $U\Sigma V^T$ be a singular value decomposition of A. Then

$$\begin{aligned}
\|A\|_{\text{op}} &= \sup_{v\in\mathbb{R}^n\setminus\{0\}} \frac{\|Av\|}{\|v\|} = \sup_{v\in\mathbb{R}^n\setminus\{0\}} \frac{\|U\Sigma V^T v\|}{\|v\|} \\
&= \sup_{v\in\mathbb{R}^n\setminus\{0\}} \frac{\|\Sigma V^T v\|}{\|v\|} && (U \text{ preseves the norm}) \\
&= \sup_{v\in\mathbb{R}^n\setminus\{0\}} \frac{\|\Sigma v\|}{\|V v\|} && (\text{substitute } V^T v \text{ with } v) \\
&= \sup_{v\in\mathbb{R}^n\setminus\{0\}} \frac{\|\Sigma v\|}{\|v\|} = \max_i |\Sigma_{ii}| = \|A\|
\end{aligned}$$

3.2 Matrix Bernstein Inequality

Finally, we are ready for Matrix Bernstein.

Claim 9. Bernstein Concentration Inequality for Matrices. *Suppose $X_1, \ldots, X_m$ are independent, symmetric random n by n matrices s.t. for all i,*

$$\mathbb{E}[X_i] = 0, \quad \|X_i\| \leq K,$$

and let

$$X = \sum_{i=1}^{m} X_i, \quad \sigma^2 = \left\|\mathbb{E}[X^2]\right\|$$

Then

$$\mathbb{P}\left[\|X\| \geq t\right] \leq 2n \exp\left(-\frac{1}{4}\min\left(\frac{t^2}{\sigma^2}, \frac{t}{K}\right)\right).$$

We will not prove this, but we will attempt to build a bit of intuition for this claim. **Author's note**: We didn't go over the stuff below in class and it might be completely wrong. Take it with a grain of salt.

3.2.1 Change of Basis

Let QDQ^T be an orthonormal eigen-decomposition of X. This means that $v_i = Q_{i*}$ are the eigenvectors of X and $\lambda_i = D_{ii}$ are their respective eigenvalues. To get everything to be well-defined as random variables, we'll choose the ordering of eigenvectors uniform randomly. Note that this requires the eigenvalues to be unique, but we'll ignore that detail.

Now we will change basis using Q. Let $Y_i = Q^T X_i Q$ and $Y = Q^T X Q = D$. Our first leap of faith will be to pretend that Q is independent from pairs of X_i. Q is a unitary transformation (rotation and reflection) resulting from the sum X, and the set of unitary transformations is compact, so the hope is that it can't leak too much information.

If we take the leap, we have that $Y_1, \ldots, Y_m$ are independent and symmetric, $Y = \sum_i Y_i$, and

$$\mathbb{E}[Y_i] = \mathbb{E}[Q^T X_i Q] = \mathbb{E}[Q^T]\,\mathbb{E}[X_i]\,\mathbb{E}[Q] = 0, \quad \|Y_i\| = \|X_i\| \le K$$
$$\left\|\mathbb{E}[Y^2]\right\| = \Big\| \sum_{i,j} \mathbb{E}[Q^T X_i X_j Q] \Big\| = \Big\| \mathbb{E}[Q^T] \Big[\sum_{i,j} \mathbb{E}[X_i]\,\mathbb{E}[X_j] \Big] \,\mathbb{E}[Q] \Big\| = \sigma^2.$$

We also have that the eigenvectors of Y are $e_1, \ldots, e_n$ and their respective eigenvalues are $\lambda_1, \ldots, \lambda_n$, so we have $\|Y\| = \max_i |\lambda_i| = \|X\|$. [**Notation**: e_i is the vector where the ith entry is 1 and the others are 0.] Since we ordered the eigenvalues randomly, the distribution of all the λ_i should be the same.

$$\mathbb{P}[\|Y\| \ge t] = \mathbb{P}[\max_{i \in [n]} |\lambda_i| \ge t] \le \sum_{i=1}^{n} \mathbb{P}[|\lambda_i| \ge t] = n\,\mathbb{P}[|\lambda_1| \ge t].$$

3.2.2 Bounding $\mathbb{P}[|\lambda_1| \ge t]$

Now we will attempt to bound $\mathbb{P}[|\lambda_1| \ge t]$. We know that $\sum_i Y_i e_1 = \lambda_1 e_1$, so $\sum_i (Y_i)_{11} = \lambda_1$. But the spectral norm is equivalent to the operator norm for symmetric matrices, so

$$|(Y_i)_{11}| \le \|Y_i e_1\| \le \|Y\| \|e_1\| = \|Y\| \le K.$$

Also, since the eigenvectors of Y and Y^2 are the same,

$$\begin{aligned}
\sigma^2 = \left\|\mathbb{E}[Y^2]\right\| &= \max_{i \in [n]} \left\|\mathbb{E}[Y^2] e_i\right\| = \max_{i \in [n]} \left\|\mathbb{E}[\lambda_i^2] e_i\right\| \\
&= \max_{i \in [n]} \mathbb{E}[\lambda_i^2] = \mathbb{E}[\lambda_1^2] = \mathrm{Var}[\lambda_1].
\end{aligned}$$

Next, we will take our second leap of faith by assuming that $(Y_1)_{11}, \ldots, (Y_m)_{11}$ are independent. Roughly speaking, what our change of basis did is eliminate the off-diagonal entries of X so Y is diagonal. Then while we have an obvious dependency between the off-diagonal entries of $Y_1, \ldots, Y_m$, no such thing exists for the diagonal entries. The hope is then that since $X_1, \ldots, X_m$ are independent, the diagonal entries are as well.

We have now satisfied all the requirements to use the regular Bernstein Inequality on $\lambda_1 = \sum_i (Y_i)_{11}$. To summarize, we know that:

$$\{(Y_i)_{11}\}_i \text{ are independent}, \quad \mathbb{E}[(Y_i)_{11}] = 0, \quad |(Y_i)_{11}| \le K, \quad \mathrm{Var}[\lambda_1] \le \sigma^2.$$

Finally, applying regular Bernstein, we get

$$\begin{aligned}
\mathbb{P}[|\lambda_1| \ge t] &\le 2 \exp\left(-\frac{1}{4} \min\left(\frac{t^2}{\sigma^2}, \frac{t}{K} \right) \right) \\
\mathbb{P}[\|X\| \ge t] = \mathbb{P}[\|Y\| \ge t] &\le n\,\mathbb{P}[|\lambda_1| \ge t] \\
&\le 2n \exp\left(-\frac{1}{4} \min\left(\frac{t^2}{\sigma^2}, \frac{t}{K} \right) \right).
\end{aligned}$$

4 Rudelson-Vershynin

Rudelson-Vershynin is a concentration inequality for the covariance matrix of a set of vectors x_i, which is defined as $\frac{1}{m}\sum_{i=1}^m x_i x_i^T$. We will prove this inequality using Matrix Bernstein.

Theorem 10. Rudelson-Vershynin [RV05].

Let $K \geq 1$, $x_1, \dots, x_m \in \mathbb{R}^n$ be independent random vectors s.t. for all i,

$$\max_{i\in[m]} \|x_i\| \leq K, \quad \left\|\mathbb{E}[x_i x_i^T]\right\| \leq 1.$$

Then there exists some C s.t. if $CK\sqrt{\frac{1}{m}\log n} \leq 1$,

$$\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m x_i x_i^T - \frac{1}{m}\,\mathbb{E}\left[\sum_{i=1}^m x_i x_i^T\right]\right\|\right] \lesssim CK\sqrt{\frac{1}{m}\log n}.$$

Proof. Let $Y_i = x_i x_i^T - \mathbb{E}[x_i x_i^T]$. Then $\mathbb{E}[Y_i] = 0$ and

$$\|Y_i\| \leq \|x_i x_i^T\| + \|\mathbb{E}[x_i x_i^T]\| \leq K^2 + 1 \leq 2K^2.$$

Also,

$$\begin{aligned}
\left\|\sum_{i=1}^m \mathbb{E}[Y_i^2]\right\| &\leq \sum_{i=1}^m \left\|\mathbb{E}[Y_i^2]\right\| = \sum_{i=1}^m \left\|\mathbb{E}[(x_i x_i^T - \mathbb{E}[x_i x_i^T])^2]\right\| \\
&= \sum_{i=1}^m \left\|\mathbb{E}[x_i x_i^T x_i x_i^T] - 2\,\mathbb{E}[\mathbb{E}[x_i x_i^T] x_i x_i^T] + \mathbb{E}[x_i x_i^T]^2\right\| \\
&= \sum_{i=1}^m \left\|\|x_i\|^2\,\mathbb{E}[x_i x_i^T] - \mathbb{E}[x_i x_i^T]^2\right\| \\
&\leq \sum_{i=1}^m \left(\|x_i\|^2 \|\mathbb{E}[x_i x_i^T]\| + \mathbb{E}[x_i x_i^T]^2\right) \\
&\leq \sum_{i=1}^m (K^2 + 1) = m(K^2+1) \leq 2mK^2.
\end{aligned}$$

Now let $E = \left\|\frac{1}{m} Y_i\right\|$. Applying Matrix Bernstein to $\sum_{i=1}^m Y_i$, we get that

$$\begin{aligned}
\mathbb{P}\left[E \geq t\right] &= \mathbb{P}\left[\left\|\sum_{i=1}^m Y_i\right\| \geq mt\right] \\
&\leq 2n \exp\left(-\frac{1}{4}\min\left(\frac{(mt)^2}{2mK^2}, \frac{mt}{2K^2}\right)\right) \\
&= 2n \exp\left(-\frac{1}{4}\min\left(\frac{mt^2}{2K^2}, \frac{mt}{2K^2}\right)\right) \\
&= 2n \exp\left(-\frac{m}{8K^2}\min(t^2, t)\right) \\
&= f(t).
\end{aligned}$$

To bound our expectation, we know that

$$\mathbb{E}[E] = \int_0^\infty \mathbb{P}[E \geq t]dt \leq \int_0^\infty \min(1, f(t))dt.$$

We will evaluate this integral in pieces. First, we need to find at what point $f(t) \leq 1$:

$$f(t) \leq 1, \quad \exp\left(-\frac{m}{8K^2}\min(t^2, t)\right) \leq \frac{1}{2n}, \quad \min(t^2, t) \geq \frac{8K^2}{m}\ln 2n.$$

Now let $s^2 = 8K^2/m \cdot \ln 2n$, and suppose $s \leq 1$. Then

$$\begin{aligned}
\int_0^\infty \min(1, f(t))dt &= s + \int_s^\infty f(t)dt \\
&= s + \int_s^1 2n\exp\left(-\frac{m}{8K^2}t^2\right)dt + \int_1^\infty 2n\exp\left(-\frac{m}{8K^2}t\right)dt \\
&= s + A + B
\end{aligned}$$

For the first part[1],

$$\begin{aligned}
A = \int_s^1 2n\exp\left(-\frac{m}{8K^2}t^2\right)dt &= 2n\sqrt{\frac{8K^2}{m}}\int_{\ln 2n}^1 e^{-u^2}du \\
\leq \frac{2ns}{\sqrt{\ln 2n}}\int_{\ln 2n}^\infty e^{-u^2}du &= \frac{2ns}{\sqrt{\ln 2n}}\left[\Theta(e^{-u^2}z^{-1})\right]_{u=\ln 2n} \\
= \Theta\left(s(\ln 2n)^{-3/2}(2n)^{1-\ln 2n}\right) &= O(s)
\end{aligned}$$

For the second part,

$$\begin{aligned}
B &= \int_1^\infty 2n\exp\left(-\frac{m}{8K^2}t\right)dt = 2n\frac{8K^2}{m}\exp\left(-\frac{m}{8K^2}\right) \\
&= 2n\frac{s^2}{\ln 2n}\exp\left(-\frac{\ln 2n}{s^2}\right) = s^2(\ln 2n)^{-1}(2n)^{1-s^{-2}} \\
&= O(s^2) = O(s).
\end{aligned}$$

Thus $\mathbb{E}[E] \leq s + A + B = s + O(s) + O(s) = O(s)$.

However, $s = K\sqrt{\frac{8}{m}\ln 2n} \leq CK\sqrt{\frac{1}{m}\ln n}$ for some C, so we have that

$$\text{if } CK\sqrt{\frac{1}{m}\ln n} \leq 1, \text{ then } \mathbb{E}[E] \lesssim CK\sqrt{\frac{1}{m}\ln n}.$$

□

References

[RV05] Rudelson, Mark, and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)* 54.4 (2007): 21-es.

[1] See https://math.stackexchange.com/questions/3703576/asymptotic-rate-of-decrease-of-error-function for how to bound the Gaussian integral.