CS 388R: Randomized Algorithms, Fall 2021 21 October, 2021

Lecture 16: Matrix Concentration and Graph Sparsification

Prof. Eric Price *Scribe: Aditya Parulekar, Rochan Avlur*

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

1 Overview

In the last lecture we discussed online matching algorithms.

In this lecture we learn about matrix concentration inequalities, which we will use to eventually find results on graph sparsification. In particular, we derive the Bernstein concentration inequality for scalar random variables, extend this result to symmetric matrices, and then prove the Rudelson-Vershynin (RV) theorem.

2 Introduction

Imagine a setting in which we have a graph G that is possibly quite dense, i.e., $G = (|E| \geq n^2$ with $|V| = n)$. Densely connected graphs can make certain algorithms inefficient. Rather, it is beneficial to operate on a sparse graph G' that captures the structure of G. In particular, we want to operate on a sparse graph where the size of most cuts are approximately equal to those in the original dense graph.

For instance, suppose our dense graph consists of two densely connected clusters that share a edge with a single intermediate vertex. Suppose we generate a sparse graph by randomly sampling a fixed number of edges from our original graph. With some non-zero probability, we may not sample the edges connecting the two clusters. This could lead our downstream algorithm to believe that the original graph was disconnected which is a very significant error. We want our sparse graph to capture key properties such as these.

To perform graph sparsification, we are going to need concentration inequalities on matrices. But first, let's look at the following concentration inequality for real numbers

3 Bernstein's Concentration Inequality

Theorem 1 (Bernstein's Concentration Inequalities)**.** *Let $X_1, X_2, \ldots, X_m$ be independent random reals satisfying*

$$\mathbb{E}[X_i] = 0, \qquad \max |X_i| \leq k \ \forall i, \qquad \sum_i \mathbb{E}[X_i^2] \leq \sigma^2$$

Then, we have for some constant c

$$\mathbb{P}\left[\left|\sum_i X_i\right| \geq t\right] \leq 2\exp\left(-c \cdot \min\left(\frac{t^2}{\sigma^2}, \frac{t}{k}\right)\right)$$

This can be analyzed by showing that $\sum_i X_i$ is *subgamma* with certain parameters. To see how to do this, see previous year's Lecture 17 notes. This is equivalent to multiplicative chernoff bound. However, using multiplicative chernoff would give us a significantly worse bound.

Now, we state a variant of this concentration inequality for matrices.

Theorem 2 (Matrix Bernstein's Concentration Inequalities)**.** *Let $X_1, X_2, \ldots, X_m$ be independent random symmetric matrices in $\mathbb{R}^{n\times n}$ satisfying*

$$\mathbb{E}[X_i] = 0 \qquad \max_i \|X_i\|^1 \leq k \qquad \left\|\sum_i \mathbb{E}[X_i^2]\right\| \leq \sigma^2$$

Then, we have for some constant c

$$\mathbb{P}\left[\left\|\sum_i X_i\right\| \geq t\right] \leq 2n\exp\left(-c \cdot \min\left(\frac{t^2}{\sigma^2}, \frac{t}{k}\right)\right)$$

4 Approximating Covariance Matrices

Motivation Given a sample of m vectors $x_1, x_2, \ldots, x_m \in \mathbb{R}^n$, how well does the empirical covariance matrix $\frac{1}{m}\sum_i x_i x_i^\top$ concentrate about the true covariance matrix $\frac{1}{m}\sum_i \mathbb{E}[x_i x_i^\top]$?

Using the matrix version of Bernstein's inequality, we prove the following concentration of empirical covariance matrices

Theorem 3 (Rudelson Vershynin [RV05])**.** *Let $x_1, x_2, \ldots, x_m \in \mathbb{R}^n$ be independent random real vectors satisfying*

$$\max \|x_i\| \leq K, \qquad K \geq 1, \qquad \mathbb{E}[x_i x_i^\top] \leq 1$$

Then,

$$\mathbb{E}\left[\left\|\frac{1}{m}\sum_i x_i x_i^\top - \frac{1}{m}\sum_i \mathbb{E}[x_i x_i^\top]\right\|\right] \lesssim K\sqrt{\frac{\log n}{m}}$$

for $K\sqrt{\frac{\log n}{m}} \leq 1$

Proof. We will use matrix-Bernstein. Let $Y_i := x_i x_i^\top - \mathbb{E}[x_i x_i^\top]$. Clearly, $\mathbb{E}[Y_i] = 0$. We have

$$\begin{aligned}\max \|Y_i\| &\leq \max \|x_i x_i^\top\| + \|\,\mathbb{E}[x_i x_i^\top]\| \\ &\leq K^2 + 1 \leq 2K^2\end{aligned}$$

[1]Here, $\|A\|$ refers to the spectral norm, which is defined as

$$\|A\| = \max_{\|u\|_2=\|v\|_2=1} u^\top A v.$$

For symmetric matrices, this is equivalent to the largest eigenvalue of A. For vectors and scalars, $\|\cdot\|$ defaults to l^2-norm and absolute value respectively.

and

$$\begin{aligned}
\sigma^2 &= \left\| \sum_i \mathbb{E}[Y_i^2] \right\| \\
&= \left\| \sum_i \mathbb{E}[x_i x_i^\top x_i x_i^\top - x_i x_i^\top \mathbb{E}[x_i x_i^\top] - \mathbb{E}[x_i x_i^\top] x_i x_i^\top + \mathbb{E}[x_i x_i^\top]^2] \right\| \\
&= \left\| \sum_i \mathbb{E}[x_i x_i^\top x_i x_i^\top] - \mathbb{E}[x_i x_i^\top]\,\mathbb{E}[x_i x_i^\top] - \mathbb{E}[x_i x_i^\top]\,\mathbb{E}[x_i x_i^\top] + \mathbb{E}[x_i x_i^\top]^2 \right\| \\
&= \left\| \sum_i \mathbb{E}[x_i x_i^\top x_i x_i^\top] + \mathbb{E}[x_i x_i^\top]^2 \right\| \\
&\leq \sum_i \left\| \mathbb{E}[x_i x_i^\top x_i x_i^\top] \right\| + \sum_i \left\| \mathbb{E}[x_i x_i^\top]^2 \right\| \qquad (1)
\end{aligned}$$

using linearity of expectation and the triangle inequality. Now, let

$$u = \arg\min_v v^\top \mathbb{E}[x_i x_i^\top x_i x_i^\top] v$$

Since $\mathbb{E}[x_i x_i^\top x_i x_i^\top]$ is symmetric,

$$\begin{aligned}
\left\| \mathbb{E}\left[x_i x_i^\top x_i x_i^\top \right] \right\| &= u^\top \mathbb{E}\left[x_i x_i^\top x_i x_i^\top \right] u \\
&= \mathbb{E}\left[u^\top x_i x_i^\top x_i x_i^\top u \right] \\
&= \mathbb{E}\left[x_i^\top x_i u^\top x_i x_i^\top u \right] \\
&= \mathbb{E}\left[\|x_i\|_2^2 \|x_i^\top u\|_2^2 \right]
\end{aligned}$$

Now, here, since both terms in the expectation are positive, we have that the product of their expectations is less than the expectation of the products. So,

$$\begin{aligned}
\mathbb{E}\left[\|x_i\|_2^2 \|x_i^\top u\|_2^2 \right] &\leq \mathbb{E}\left[\|x_i\|_2^2 \right] \mathbb{E}\left[\|x_i^\top u\|_2^2 \right] \\
&\leq K^2\, \mathbb{E}\left[\|x_i^\top u\|_2^2 \right] \\
&= K^2\, \mathbb{E}\left[u^\top x_i x_i^\top u \right] \\
&= K^2 u^\top \mathbb{E}\left[x_i x_i^\top \right] u \\
&\leq K^2 \left\| \mathbb{E}\left[x_i x_i^\top \right] \right\|
\end{aligned}$$

Using this and the fact that $\mathbb{E}[x_i x_i^\top] = 1$, we plug into (1) to get

$$\sigma^2 \leq \sum_i K^2 + 1 \leq 2mK^2$$

So, by Bernstein, with $\sigma^2 = 2mK^2, k = 2K^2$

$$\mathbb{P}\left[\left\| \frac{1}{m} \sum_i Y_i \right\| \geq t \right] = \mathbb{P}\left[\left\| \sum_i Y_i \right\| \geq mt \right]$$

$$\leq 2n \exp\left(-c \min\left(\frac{m^2t^2}{2mK^2}, \frac{mt}{2K^2}\right)\right)$$
$$\leq 2n \exp\left(\frac{-cmt^2}{2K^2}\right)$$

for $t \geq 1$. So, setting $t = K\sqrt{\frac{\log \frac{n}{\delta}}{m}}$, we get

$$\mathbb{P}\left[\left\|\frac{1}{m}\sum_i Y_i\right\| \geq K\sqrt{\frac{\log \frac{n}{\delta}}{m}}\right] \leq \delta$$

□

5 Intro to Graph Sparsification

Next time, we will get into the applications of this to graph sparsification. As an example of how this works, say you have an unweighted, undirected graph $G = (E, V)$. Construct a matrix $U \in \mathbb{R}^{|E| \times |V|}$. If the ith edge in U connects vertices (v_1, v_2), then U_{i,v_1} is set to 1 and U_{i,v_2} is set to -1, i.e., ith edge represented by row vector u_i has $u_{v_1} = 1$ and $u_{v_2} = -1$ (order in which value is assigned does not matter). Then, the **Laplacian** of G is defined as $L_G = U^\top U$, and can also be expressed as $L_G = D - A$, where D is a diagonal matrix with the degrees of the vertices on the diagonal, and A is the adjacency matrix of G. The graph Laplacian gives us information about the covariance of distribution of the rows of U.

If we were to simply sample edges from this graph, then Rudelson-Vershynin on the rows of this matrix gives the edge count to have a good approximation of the covariance matrix, which is $U^\top U$. However, this does not work very well for all graphs. Consider the example graph introduced in the beginning. For such barbell shaped graphs, the sample range term K is bad when we try to apply Rudelson-Vershynin. Next time, we will see how to get around this by sampling with a different probability distribution.

References

[RV05] Mark Rudelson, Roman Vershynin. Sampling from large matrices: an approach through geometric functional analysis. *arXiv: math/0503442 [math.FA]*, 2005.