36-709: Advanced Statistical Theory **Spring 2019**

Lecture 13: March 5

Lecturer: Alessandro Rinaldo *Scribe: Tim Barry*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

In these notes we finish our discussion of matrix concentration inequalities. We then see some applications of matrix concentration inequalities to covariance estimation and networks.

Remark on notation: In these notes the norm function $||\cdot||$ refers to the operator norm for matrices and the Euclidean norm for vectors.

13.1 Matrix concentration inequalities

Recall the matrix Bernstein inequality.

Theorem 13.1 *(Matrix Bernstein inequality) Let $X_1, \dots, X_n$ be mean-zero, symmetric, $d \times d$ random matrices such that $||X_i|| \leq C$ almost surely for all $i \in \{1, \dots, n\}$. Then for all $t \geq 0$,*

$$\mathbb{P}\left\{ \left|\left| \sum_{i=1}^n X_i \right|\right| \geq t \right\} \leq 2d \exp\left\{ \frac{-t^2}{2(\sigma^2 + Ct/3)} \right\},$$

where $\sigma^2 = \left|\left| \sum_{i=1}^n \mathbb{E}[X_i^2] \right|\right|$ is the norm of the matrix variance of the sum.

Recall that, for a symmetric $d \times d$ matrix A, $||A|| = \max_{i=1,\dots d} |\lambda_i(A)| = \max\{\lambda_{\max}(A), \lambda_{\max}(-A)\}$, where $\{\lambda_1, \dots, \lambda_d\}$ is the spectrum of A. By union bound,

$$\mathbb{P}\left(\max\{\lambda_{\max}(A), \lambda_{\max}(-A)\} \geq t\right) \leq \mathbb{P}\left(\lambda_{\max}(A) \geq t\right) + \mathbb{P}\left(\lambda_{\max}(-A) \geq t\right). \tag{13.1}$$

Last time we saw that

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{i=1}^n X_i \right) \geq t \right) \leq d \exp\left\{ \frac{-t^2}{2(\sigma^2 + Ct/3)} \right\}.$$

As it turns out the same bound holds for $\lambda_{\max}\left(-\sum_{i=1}^n X_i\right)$. Just repeat the proof using $-\sum_{i=1}^n X_i$ instead of $\sum_{i=1}^n X_i$. Combine these bounds by 13.1 to achieve the result.

We can derive matrix concentration inequalities for other types of matrices as well. First we define sub-Gaussian and sub-exponential random matrices. Recall that $\mathcal{S}^{d \times d}$ is the set of $d \times d$ symmetric matrices, and $\mathcal{S}_+^{d \times d}$ is the set of $d \times d$ symmetric, positive-semidefinite matrices. Define the moment-generating function $\psi_Q : \mathbb{R} \to \mathcal{S}^{d \times d}$ of a random matrix Q by

$$\psi_Q(\lambda) = \mathbb{E}\left[e^{\lambda Q}\right].$$

Definition 13.2 *(Sub-Gaussian matrices) A centered symmetric random matrix $Q \in \mathcal{S}^{d\times d}$ is sub-Gaussian with parameter $V \in S_{+}^{d\times d}$ if, for all $\lambda \in \mathbb{R}$,*

$$\psi_Q(\lambda) \preceq e^{\frac{\lambda^2 V}{2}}.$$

Definition 13.3 *(Sub-exponential matrices) A centered, symmetric random matrix $Q \in \mathcal{S}^{d\times d}$ is sub-exponential with parameters $V \in S_{+}^{d\times d}$ and $\alpha > 0$ if, for all $|\lambda| < \frac{1}{\alpha}$,*

$$\psi_Q(\lambda) \preceq e^{\frac{\lambda^2 V}{2}}.$$

Notice that, similar to the scalar case, a sub-Gaussian matrix with parameter V is sub-exponential with parameters V and 0. We are now ready to state the matrix analogue of Hoeffding's inequality.

Theorem 13.4 *(Matrix Hoeffding's inequality) Let $X_1, \ldots, X_n$ be centered, independent, symmetric, $d \times d$ random matrices that are sub-Gaussian with parameters $V_1, \ldots, V_n$. Then for all $t \geq 0$,*

$$\mathbb{P}\left\{\left|\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right|\right| \geq t\right\} \leq 2d \exp\left\{-\frac{nt^2}{2\sigma^2}\right\},$$

where $\sigma^2 = \left|\left|\frac{1}{n}\sum_{i=1}^{n} V_i\right|\right|$.

There also exists a matrix analogue of Bernstein's inequality that holds for random matrices that satisfy the sub-exponential condition. This concentration inequality is a more general version of Theorem 13.1. Indeed, random matrices with bounded operator norm are sub-exponential. We do not state the more general Bernstein's matrix inequality in these lecture notes; see instead [WW2019].

We can extend the matrix concentration inequality 13.1 to matrices that are non-symmetric or (more generally) non-square. The idea is to use a common linear algebra trick involving block matrices. Let A be a $d_1 \times d_2$ matrix. Let the $(d_1 + d_2) \times (d_1 + d_2)$ matrix Q be defined by

$$Q = \begin{bmatrix} 0_{d_1 \times d_2} & A \\ A^T & 0_{d_2 \times d_1} \end{bmatrix}.$$

We can show that $||Q|| = ||A||$. Additionally, if $Q_1, \ldots, Q_n$ are independent matrices of the above form, we can straightforwardly bound

$$\left|\left|\frac{1}{n}\sum_{i=1}^{n} \text{var}(Q_i)\right|\right|.$$

Using these facts, we can derive a Bernstein-type bound for the sum of independent, non-symmetric matrices with bounded operator norm. See [WW2019] exercise 6.10 for details.

Under certain conditions, we can replace d in 13.1 by $d_{\text{int}}\left(\sum_{i=1}^{n} \mathbb{E}[X_i^2]\right)$, where

$$d_{\text{int}}(A) = \frac{\text{tr}(A)}{||A||}$$

for positive semi-definite A. This produces a sharper bound. See for [T14] details.

13.2 Applications of matrix Bernstein inequality

Our first application of the matrix Bernstein inequality is covariance matrix estimation. We previously saw that we can bound the covariance matrix of a sub-Gaussian random vector using an exponential tail bound. Here, we derive an exponential tail bound on the covariance matrix of a bounded random vector.

Theorem 13.5 *Let $X_1, \ldots, X_n$ be independent, centered random vectors in $\mathbb{R}^d$. Suppose that, for all $i \in \{1, \ldots, n\}$, Var$(X_i) = \Sigma$ and $||X_i||_2 \leq \sqrt{C}$ almost everywhere for some $C > 0$. Then for all $t \geq 0$,*

$$\mathbb{P}\left\{\left|\left|\hat{\Sigma}_n - \Sigma\right|\right| \geq t\right\} \leq 2d \exp\left\{\frac{-nt^2}{2C\left(||\Sigma|| + 2t/3\right)}\right\},$$

where $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ is the sample covariance.

Proof: Define $Q_i = X_i X_i^T - \Sigma$. We use matrix Bernstein inequality to bound $\sum_{i=1}^n (1/n) Q_i = \sum_{i=1}^n (1/n) X_i X_i^T - \sum_{i=1}^n (1/n)\Sigma = \hat{\Sigma}_n - \Sigma$. To apply matrix Bernstein inequality we must verify several conditions on the $(1/n)Q_i$s.

i. **mean zero**: $\mathbb{E}\left[(1/n)Q_i\right] = (1/n)\left[\mathbb{E}\left(X_i X_i^T\right) - \Sigma\right] = 0$.

ii. **symmetric**: $X_i X_i^T$ and Σ are both symmetric, and so the difference $(1/n)Q_i = (1/n)(X_i X_i^T - \Sigma)$ is symmetric.

iii. **d × d**: The dimension is obvious.

iv. **independent**: The independence of the X_is implies the independence of the Q_is.

v. **bound on $||(1/n)Q_i|| = (1/n)||Q_i||$** : Observe by triangle inequality that

$$||Q_i|| = ||X_i X_i^T - \Sigma|| \leq ||X_i X_i^T|| + || - \Sigma|| = ||X_i X_i^T|| + ||\Sigma||. \tag{13.2}$$

We bound $||X_i X_i^T||$ using Cauchy-Schwarz:

$$||X_i X_i^T|| = \sup_{y \in \mathbb{R}^d : ||y|| = 1} y^T X_i X_i^T y \leq \sup_{y \in \mathbb{R}^d : ||y|| = 1} ||y|| \cdot ||X_i|| \cdot ||y|| \cdot ||X_i|| = ||X_i||^2 \leq C.$$

Moreover, by Jensen's inequality,

$$||\Sigma|| = ||\mathbb{E}\left(X_i X_i^T\right)|| \leq \mathbb{E}||X_i X_i^T|| \leq \mathbb{E}C = C.$$

Combining our bounds for $||\Sigma||$ and $||X_i X_i^T||$ with 13.2, we find $||(1/n)Q|| \leq 2C/n$.

vi. **bound on $\sigma^2 = ||\sum_{i=1}^n \mathbb{E}([(1/n)Q_i])^2|| = (1/n)||\mathbb{E}(Q_i^2)||$**. Observe that

$$\begin{aligned}\mathbb{E}(Q_i^2) &= \mathbb{E}[(X_i X_i^T - \Sigma)^2] = \mathbb{E}[(X_i X_i^T)^2] - 2\mathbb{E}[X_i X_i^T]\Sigma + \Sigma^2 = \mathbb{E}[(X_i X_i^T)^2] - \Sigma^2 \\ &\preceq \mathbb{E}[(X_i X_i^T)^2] \text{ (because } \Sigma^2 \text{ positive semi-definite) } = \mathbb{E}[X_i X_i^T X_i X_i^T] = \mathbb{E}[X_i ||X_i||^2 X_i^T] \\ &= ||X_i||^2 \mathbb{E}[X_i X_i^T] \preceq C\Sigma.\end{aligned}$$

We therefore see that $\sigma^2 \leq C||\Sigma||/n$.

Now we are ready to apply matrix Bernstein inequality:

$$P\left\{\left|\left|\hat{\Sigma}_n - \Sigma\right|\right| \geq t\right\} = \mathbb{P}\left\{\left|\left|\sum_{i=1}^n (1/n)Q_i\right|\right| \geq t\right\} \leq 2d \exp\left\{\frac{-t^2}{2(C||\Sigma||/n + 2Ct/(3n))}\right\} = 2d \exp\left\{\frac{-nt^2}{2C(||\Sigma|| + 2t/3)}\right\}.$$

■

Our next application of the matrix Bernstein inequality is to network models. Let G be a random, undirected graph on n nodes. Let $A \in \mathbb{R}^{n \times n}$ be the adjacency matrix of G, i.e.,

$$A_{ij} = \begin{cases} 1 \text{ if node } i \text{ is connected to node } j \\ 0 \text{ else} \end{cases}.$$

Because G is undirected, A is symmetric. Assume that, for all $i \in \{1, \ldots, n\}$, $A_{ii} = 0$ (i.e., no loops). Additionally, assume that $A_{ij} \sim \text{bern}(p_{ij})$ and that the A_{ij}s are independent.

We assume G takes a special form to simplify the estimation problem. Let $k \in \mathbb{N}$. Let $B \in \mathbb{R}^{k \times k}$ be a matrix of probabilities. Let $C : \{1, \ldots n\} \to \{1, \ldots, k\}$ be a surjective function. Assume there exists a partition of $\{1, \ldots, n\}$ into k communities such that

$$p_{ij} = \begin{cases} B_{C(i),C(j)} \text{ if } i \neq j \\ 0 \text{ if } i = j. \end{cases} .$$

This is called the stochastic block model. The stochastic block model is simpler than the general graphical model because the stochastic block model is parameterized only by probabilities between and within blocks.

The simplest kind of stochastic block model is the so-called pointed partition model. For this model we assume

$$p_{ij} = \begin{cases} p \text{ if } i, j \text{ in same community (and } i \neq j) \\ q \text{ if } i, j \text{ in different communities} \\ 0 \text{ if i = j} \end{cases} .$$

We can write $B = (p - q)I_k + q1_k1_k^T$, where I_k is the $k \times k$ identity matrix and 1_k is the (column) vector of ones in $\mathbb{R}^k$.

When we estimate a pointed partition model, we assume p, q and the communities themselves are unknown. Our goal typically is to recover the communities. To do this we can use spectral clustering, which requires us to know the number of communities k. The idea is as follows:

1. Compute the k leading eigenvectors $u_1, \ldots, u_k \in \mathbb{R}^n$ of A.
2. Form the $n \times k$ matrix $E = [u_1, \ldots, u_k]$. The rows of E form n points in $\mathbb{R}^k$. Apply the $k-$means clustering algorithm to the rows of E. The clusters provide an estimate of the communities in the graph G.

To prove that spectral clustering works (i.e., recovers the communities), we must show that $||A - \mathbb{E}(A)||$ is well-controlled. To do this we can use the matrix Bernstein inequality. We save this for next time.

References

[WW2019] M. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, 2019.

[T14] J. Tropp, "An introduction to matrix concentration inequalities," *Foundations and Trends in Machine Learning*, 8.1-2 (2015): 1-230.