**36-709: Advanced Statistical Theory I** **Spring 2020**

# Lecture 10: February 20

*Lecturer: Siva Balakrishnan* *Scribe: Alec McClean*

In the previous lecture we introduced two general concentration bounds for random matrices (RMs), Matrix-Hoeffding and Matrix-Bernstein, the first of which we proved last lecture using Lieb's inequality and the Chernoff method. In these notes we recap the definitions of Matrix-Hoeffding and Matrix-Bernstein, show an application of the Matrix-Bernstein concentration bound, discuss sparse covariance estimation, and basis pursuit and its application to compressed sensing.

## 10.1 Review

### 10.1.1 Matrix-Hoeffding

**Theorem 10.1** *Let* $\mathbf{Q_1}, ..., \mathbf{Q_n} \in \mathbb{S}^{d \times d}$ *be zero-mean random matrices. If* $\mathbf{Q_i}$ *are* $\mathbf{V_i}$ *sub-Gaussian for all* $i$*, such that*

$$\mathbb{E} \exp(t\mathbf{Q_i}) \leq \exp\left(\frac{t^2 \mathbf{V_i}^2}{2}\right)$$

*then*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{Q_i}\right\|_{op} \geq t\right) \leq 2d \exp\left(\frac{-nt^2}{2\sigma^2}\right)$$

*where*

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

### 10.1.2 Matrix-Bernstein

**Theorem 10.2** *If* $\mathbf{Q_i}$ *are bounded such that* $\|\mathbf{Q_i}\|_{op} \leq b$ *then*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{Q_i}\right\|_{op}\geq t\right)\leq 2d\exp\left(\frac{-nt^2}{2(bt+\sigma^2)}\right)$$

*where*

$$\sigma^2=\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{V}(\mathbf{Q_i})\right\|_{op}$$

## 10.2 Applying Matrix-Bernstein to Covariance Estimation

**Corollary 10.3** *Let $x_1,...,x_n\in\mathbb{R}^d$ be IID zero-mean random vectors with covariance $\mathbf{\Sigma}$ such that $\|x_i\|_2\leq\sqrt{b}$. Then for all $t>0$, the sample covariance matrix $\widehat{\mathbf{\Sigma}}=\frac{1}{n}\sum_{i=1}^{n}x_ix_i^T$ satisfies*

$$\mathbb{P}\left(\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_{op}\geq t\right)\leq 2d\exp\left(\frac{-nt^2}{2b(\|\mathbf{\Sigma}\|_{op}+t)}\right)$$

**Proof:** Define $\mathbf{Q_i}:=x_ix_i^T-\mathbf{\Sigma}$ and note that $\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{Q_i}\|_{\text{op}}=\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|_{\text{op}}$. These matrices have controlled operator norm:

$$\begin{aligned}
\|\mathbf{Q_i}\|_{\text{op}}&=\|x_ix_i^T-\mathbf{\Sigma}\|_{\text{op}}\\
&\leq\|x_ix_i^T\|_{\text{op}}+\|\mathbf{\Sigma}\|_{\text{op}} && \text{by the triangle inequality}\\
&\leq\|x_i\|_2^2+\|\mathbf{\Sigma}\|_{\text{op}} && \text{converting operator to Frobenius norm}\\
&\leq b+\|\mathbf{\Sigma}\|_{\text{op}} && \text{by definition } x_i
\end{aligned}$$

Since $\mathbf{\Sigma}=\mathbb{E}[x_ix_i^T]$, we have $\|\mathbf{\Sigma}\|_{\text{op}}=\max_{v\in\mathbb{S}^{d-1}}\mathbb{E}[\langle v,x_i\rangle^2]\leq b$. Therefore

$$\|\mathbf{Q_i}\|_{\text{op}}\leq b+\|\mathbf{\Sigma}\|_{\text{op}}\leq 2b.$$

Turning to the variance of $\mathbf{Q_i}$, we have that $\mathbb{V}(\mathbf{Q_i})=\mathbb{E}[(x_ix_i^T)^2]-\mathbf{\Sigma}^2\leq\mathbb{E}[(x_ix_i^T)^2]$. Therefore we can bound $\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{V}(\mathbf{Q_i})\|_{\text{op}}$ like so:

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{V}(\mathbf{Q_i}) \right\|_{\text{op}} &\leq \left\| \mathbb{E}[(x_i x_i^T)^2] \right\|_{\text{op}} && \text{by triangle inequality and IID} \\
&\leq \left\| \mathbb{E}[\|x_i\|_2^2 x_i x_i^T] \right\|_{\text{op}} \\
&\leq b \left\| \mathbb{E}[x_i x_i^T] \right\|_{\text{op}} && \text{by definition } x_i \\
&= b \left\| \mathbf{\Sigma} \right\|_{\text{op}}
\end{aligned}$$

So, applying the **Matrix-Bernstein** bound we conclude that:

$$\begin{aligned}
\mathbb{P}\left( \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\text{op}} \geq t \right) &= \mathbb{P}\left( \left\| \frac{1}{n} \sum_{i=1}^{n} Q_i \right\|_{\text{op}} \geq t \right) \\
&\leq 2d \exp\left( \frac{-nt^2}{2b\left( \left\| \mathbf{\Sigma} \right\|_{\text{op}} + 2t \right)} \right)
\end{aligned}$$

■

This result gives us similar, but slightly worse, rates of convergence as we might obtain by other means (such as leveraging the sub-Gaussianity of $x_i$). Consider the example where $x_i$ are chosen uniformly from the sphere $\mathbb{S}^{d-1}(\sqrt{d})$, so that $\|x_i\|_2 = \sqrt{d}$ for all $i = 1, ..., n$. By construction, we have $\mathbb{E}[x_i x_i^T] = \mathbf{\Sigma} = \mathbf{I_d}$, and hence $\|\mathbf{\Sigma}\|_{\text{op}} = 1$. So, by Matrix-Bernstein we have

$$\mathbb{P}\left( \|\widehat{\mathbf{\Sigma}} - \mathbf{I_d}\|_{\text{op}} \geq t \right) \leq 2d \exp\left( \frac{-nt^2}{2d(1+2t)} \right) \quad \text{for all } t \geq 0.$$

This bound implies that

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{I_d}\|_{\text{op}} \preceq \sqrt{\frac{d \log d}{n}} + \frac{d \log d}{n}$$

with high probability, so the sample covariance is consistent as long as $\frac{d \log d}{n} \to 0$. This result is almost optimal. The $\log d$ factor could be removed into this particular case by using the fact that $x_i$ is a sub-Gaussian random vector.

## 10.3 Sparse Covariance Estimation

Above, we considered estimating a general unstructured covariance matrix with the sample covariance. When a covariance matrix has additional structure, such as sparsity, then faster rates of estimation are possible using different estimators from the sample covariance matrix.

Let $x_1, ..., x_n \in \mathbb{R}^d$ be $\sigma$-sub-Gaussian zero-mean random vectors, and assume that the covariance matrix $\Sigma$ is sparse with adjacency matrix $A$, such that $n > d$. Our approach will be to use hard thresholding on the entries of $\widehat{\Sigma}$ to get a tighter bound on $\|\widehat{\Sigma} - \Sigma\|_{\text{op}}$.

### 10.3.1 A Naive Approach: the Union Bound

To gain intuition for our threshold, we first show a naive bound on $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty$ by considering an element-wise union bound on $|\widehat{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}|$. We can consider the diagonal and off-diagonal elements separately. First, note that, since all $x_i$ are $\sigma$-sub-Gaussian, then

$$\widehat{\mathbf{\Sigma}}_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}$$

is $(\sigma^4, \sigma^2)$-sub-exponential for all $j$ and $k$. So, applying the union bound and Bernstein bound for random variables to both the diagonal and off-diagonal elements, we conclude that, as long as $n > \log d$, then with probability at least $1 - \delta$

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \leq C\sigma^2\sqrt{\frac{\log d}{n}}$$

Using the same method, we can show the a very similar bound for the operator norm with only an extra $d$ term included. With probability at least $1 - \delta$, we have

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\text{op}} \leq Cd\sigma^2\sqrt{\frac{\log d}{n}}$$

### 10.3.2 Hard Thresholding

The above analysis suggests that a suitable hard threshold would be $t := C\sigma^2\sqrt{\frac{\log d}{n}}$ with hard thresholding operator:

$$\tilde{\mathbf{\Sigma}}_{ij} = \widehat{\mathbf{\Sigma}}_{ij}\mathbb{I}\left(\left|\widehat{\mathbf{\Sigma}}_{ij}\right| \geq 2t\right)$$

**Theorem 10.4** *With $x_1, ..., x_n$, the threshold $t$, and the hard thresholding operator $\tilde{\mathbf{\Sigma}}$ defined above, then with probability $1 - \delta$*

$$|\tilde{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}| \leq 4t\mathbf{A}_{ij} \quad \forall\ i, j$$

*and*

$$\left\|\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_{op} \leq 4t\left\|\mathbf{A}\right\|_{op}$$

**Proof:** Suppose that $\mathbf{\Sigma}_{ij} = 0$, then $\mathbf{A}_{ij} = 0$ and $\tilde{\mathbf{\Sigma}}_{ij} = 0$. Then, suppose the opposite, that $\mathbf{\Sigma}_{ij} \neq 0$. Then, by applying the triangle inequality and the definition of $\tilde{\mathbf{\Sigma}}_{ij}$ we have that

$$\left|\tilde{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}\right| \leq \left|\tilde{\mathbf{\Sigma}}_{ij} - \widehat{\mathbf{\Sigma}}_{ij}\right| + \left|\widehat{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}\right| \leq 2t + \left|\widehat{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}\right|.$$

Further, we established that $\left|\widehat{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}\right| \leq 2t$ with probability $1 - \delta$ when bounding $\left\|\widehat{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}\right\|_{\infty}$. So, we conclude that, with probability at least $1 - \delta$,

$$\left|\tilde{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}\right| \leq 4t\mathbf{A}_{ij} \quad \forall\ i, j.$$

So, we have shown that the matrix $\mathbf{B} := |\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}|$ and the adjacency matrix $\mathbf{A}$ satisfy the element-wise inequality $\mathbf{B}_{ij} \leq 4t\mathbf{A}_{ij}$. Since both $\mathbf{A}$ and $\mathbf{B}$ have only non-negative entries, we are guaranteed that $\|\mathbf{B}\|_{\text{op}} \leq 4t\|\mathbf{A}\|_{\text{op}}$ and so

$$\|\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\text{op}} \leq 4t\|\mathbf{A}\|_{\text{op}}.$$

■

## 10.4 Basis Pursuit and Compressed Sensing

Consider a high-dimensional linear model, where we observe $y \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n \times d}$ such that

$$y = \mathbf{X}\theta^*$$

where $\theta^* \in \mathbb{R}^d$ and we want to recover $\theta^*$. In the situation where $n < d$, this is an underdetermined system, so there is a whole subspace of solutions. But, what if we assume (or are told) that there is a sparse solution to the system? Then, we know that there exists some vector $\theta^* \in \mathbb{R}^d$ with at most $s \ll d$ non-zero entries such that $y = \mathbf{X}\theta^*$.

A natural candidate for solving this (non-convex) optimization problem is the $\ell_0$-"norm":

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_0 \qquad \text{such that } \mathbf{X}\theta = y.$$

However, this optimization problem is computationally intractable. So, instead, we replace $\ell_0$-minimization with $\ell_1$-minimization, which leads us to

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \qquad \text{such that } \mathbf{X}\theta = y.$$

This is now a solvable convex program! Generally, this optimization problem is referred to as ***basis pursuit***.

### 10.4.1 Exact recovery and the restricted nullspace

Now, the natural question following from our norm-switch above is: ***when is solving basis pursuit (i.e. optimization in $\ell_1$) the same as solving the optimization in the $\ell_0$-norm?***

Let us suppose there exists $\theta^* \in \mathbb{R}$ such that $y = \mathbf{X}\theta^*$. Further, $\theta^*$ has support $S \subset \{1, 2, ..., d\}$, i.e. $\theta^*_j = 0$ for $j \in S^c$. The success of basis pursuit will depend on how the nullspace of $\mathbf{X}$ is related to this support and the geometry of the $\ell_1$-ball. Let us define

$$\mathcal{C}(S) = \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}$$

corresponding to the cone of vectors whose $\ell_1$-norm on the support dominates their $\ell_1$-1 norm off the support. We can link the nullspace of $\mathbf{X}$ to $\mathcal{C}(S)$ with the following definition:

**Definition 10.5** *The matrix $\mathbf{X}$ satisfies the **restricted nullspace property** with respect to $S$ if $\mathcal{C}(S) \cap null(\mathbf{X}) = \{0\}$.*

Finally, we can related the restricted nullspace property to the success of the basis pursuit program with the following theorem

**Theorem 10.6** *The following two properties are equivalent:*

1. *The matrix $\mathbf{X}$ satisfies the restricted nullspace property with respect to $S$.*
2. *For any vector $\theta^* \in \mathbb{R}^d$ with support $S$, basis pursuit applied with $y = \mathbf{X}\theta^*$ has a unique solution $\widehat{\theta} = \theta^*$.*

**Proof:** In these notes, we will show that 1 $\implies$ 2. Assume there exists some $\tilde{\theta} \neq \theta^*$ that also solves basis pursuit, i.e. $\mathbf{X}\tilde{\theta} = y$ and $\|\tilde{\theta}\|_1 \leq \|\theta\|_1$ for all $\theta$. Then, we also have that $\|\tilde{\theta}\|_1 = \|\theta^*\|_1$. Now, we define the error vector $\Delta := \tilde{\theta} - \theta^*$. By construction, we have that $\mathbf{X}\theta^* = \mathbf{X}\tilde{\theta} = \mathbf{X}(\theta^* + \Delta)$. So, $\mathbf{X}\Delta = 0$ and $\Delta \in null(\mathbf{X})$.

Next, we note that $\theta^*_{S^c} = 0$ and show the following:

$$\begin{aligned}
\|\theta^*_S\|_1 &= \|\theta^*\|_1 \\
&\geq \|\tilde{\theta}\|_1 \\
&= \|\theta^* + \Delta\|_1 \\
&= \|\theta^* + \Delta_S + \Delta_{S^c}\|_1 \\
&= \|\theta^*\|_1 + \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \\
&\geq \|\theta^*_S\|_1 - \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1
\end{aligned}$$

Rearranging this final inequality, we see that $\|\Delta_S\|_1 \geq \|\Delta_{S^c}\|_1$, which implies that $\Delta \in \mathcal{C}(S)$. So, we have shown that $\Delta \in null(\mathbf{X})$ and $\Delta \in \mathcal{C}(S)$. By our assumption of property 1, the only $\Delta$ that can satisfy this conclusion is $\Delta = 0$, so $\tilde{\theta} = \theta^*$, and we have a unique solution to basis pursuit.

■

### 10.4.2 Application: Compressed Sensing

Finally, we turn to compressed sensing, which is based on the $\ell_1$-relaxation we described above alongside the random projection method. The goal is to both compress and reconstruct a signal $\beta^*$. Compressed sensing is motivated by the inherent wastefulness of the classical approach to exploiting sparsity for signal compression. The standard approach is to compute, for a given signal $\beta^* \in \mathbb{R}^d$, the full vector $\theta^* = \mathbf{\Psi}^T\beta^* \in \mathbb{R}^d$ using a wavelet transform, and then to *discard* all but the top $s$ coefficients. Compressed sensing aims to avoid precomputing the full vector $\theta^*$ before just discarding $d - s$ coefficients.

In compressed sensing, we take $n \ll d$ random projections of the original signal $\beta^* \in \mathbb{R}^d$, each of the form $y_i = \langle x_i, \beta_i^* \rangle := \sum_{j=1}^d x_{ij}\beta_j^*$, where $x_i \in \mathbb{R}^d$ is a random vector (e.g. $x_{ij} \sim N(0,1)$). Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be our measurement matrix, where the $i^{th}$ row of $\mathbf{X}$ is $x_i^T$ and $y \in \mathbb{R}^n$ is our concatenated set of random projections. Then, our problem of reconstruction becomes finding a solution $\beta \in \mathbb{R}^d$ to the underdetermined linear system $\mathbf{X}\beta = \mathbf{X}\beta^*$ such that $\mathbf{\Psi}^T\beta$ is as sparse as possible. In other words, if we define $\widetilde{\mathbf{X}} := \mathbf{X}\mathbf{\Psi}$, then in the transform domain we are trying to solve

$$\min_{\theta\in\mathbb{R}^d} \|\theta\|_1 \quad \text{such that } y = \widetilde{\mathbf{X}}\theta.$$

Now, this optimization should look very familiar, as it is the basis pursuit linear program for which we just proved Theorem 10.6.