10-755: Advanced Statistical Theory I **Fall 2017**

Lecture 9: October 2

Lecturer: Alessandro Rinaldo *Scribes: Nic Dalmasso*

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

9.1 Matrix Bernstein Inequality - Application on Random Graphs

We will now consider an application to random graphs, mainly referencing Chung and Radcliffe [CR11]. Let A be a $n \times n$ matrix, symmetric with binary entries everywhere apart from the main diagonal - i.e. $A_{ij} = \{0, 1\} \forall i \neq j$. We can further specify by saying that for each of the elements of the matrix A:

$$A_{ij} \sim \text{Bernoulli}(p_{ij}) \quad \forall i \neq j \tag{9.1}$$

All independent of each others. We can notice that A can be seen as the adjacency matrix of a graph on $\{1, ..., n\}$ such that $i \sim j$ if $A_{ij} = 1$. We then have $\binom{n}{2}$ independent Bernoulli random variables, creating what's called an inhomogeneous random graph. If we have that $p_{ij} = p \; \forall i \neq j$ then we obtain what's called the "Erdös-Rényi" model.

The end goal of this setting is to able to apply matrix Bernstein inequality to graphs in order to do **community detection**, which is the idea of being able to retrieve communities simply from the adjacency matrix A of a graph and the number of communities k.

9.1.1 Stochastic Block Model

A very widely used method for modeling communities in networks is the stochastic block model.
Let $C : \{1, ..., n\} \to \{1, ..., k\}$ with $k \leq n$ and C being surjective.
Since C is surjective hence $\{C^{-1}(j),\, j = 1, ..., k\}$ is a partition of $\{1, ..., n\}$ by definition.
In this paradigm, k is the number of communities and I use C to determine to which community a given node belongs to. We define B as $k \times k$ matrix, symmetric and with values in $[0, 1]$ in the following way:

$$p_{ij} = B_{C(i),C(j)} \quad \forall i < j \tag{9.2}$$

Hence we model the probabilities of belonging to the same community rather than the specific pair of nodes.

Example: Let's define the matrix B in the following way:

$$B = (p - q)I_k + q1_k 1_k^T, \quad p \in (0, 1)\; q \in (0, 1)\; p > q \tag{9.3}$$

Where I_k is the $k \times k$ identity matrix and 1_k is a k-dimensional vectors of all 1. By definition of the B in equation 9.3, B has value p on the diagonal and q everywhere else. Since $p > q$ in this case hence this models the fact that it is easier to have friends or relationships within the same community rather than across communities.

9.1.2 Spectral Clustering

Considering the adjacency matrix A and the setup defined above one algorithm to perform community detection is Spectral Clustering. The algorithm works in the following way:

- Computer the k leading eigenvectors of A v_k and stack them together in a $n \times k$ matrix $V = [v_1, ..., v_k]$;
- When considering the matrix V we can think about that as n points in $\mathbb{R}^K$, so we basically look at the rows of such matrix;
- We cluster these points in k groups using k-means.

Let's now define the *degree* of a node i, which is the number of nodes to which node i is connected, so:

$$\deg(i) = \sum_j A_{ij} \tag{9.4}$$

According to the expected values of such degree we have two different regimes for graphs in general:

- *Dense Regime*: Expected degree is $\sim np$, with p bounded away from 0 (less realistic, easier to deal with);
- *Sparse Regime*: Expected degree is $O(n)$ (more realistic, more challenging) - the regime we are interested in.

Now let $\alpha_n = \max_{i<j} p_{ij}$ the largest possible probability.
This implies that $(n-1)\alpha_n$ is an upper bound for the largest expected degree.
It also implies that, in a sparse regime, we would need $\alpha_n \to 0$ with n.

The goal in this setting is to upper bound the following quantity:

$$\|A - \mathbb{E}[A]\|_{op} \tag{9.5}$$

In order to approach this problem, we write the quantity above as:

$$A - \mathbb{E}[A] = \sum_{i<j} A^{(i,j)} \tag{9.6}$$

Where $A^{(i,j)}$ are $\binom{n}{2}$ matrices of the following form:

$$A^{(i,j)} = (\xi_{ij} - p_{ij})(E^{(i,j)} + E^{(j,i)}) \tag{9.7}$$

Where:

- $\xi_{ij} \sim \text{Bernoulli}(p_{ij})$ independent;
- $E^{(j,i)}$ is a $n \times n$ matrix with all zeros except in the (j,i) position, in which there is a 1.

We have that $A^{(i,j)}$ are hence independent, centered and $\left\|A^{(i,j)}\right\|_{op} \leq 1$.
From the definition in equation 9.7 we have that:

$$\mathbb{E}[(A^{(i,j)})^2] = p_{ij}(1-p_{ij})[E^{(i,j)} + E^{(j,i)}] \implies \left\|\sum_{i<j} \mathbb{E}[(A^{(i,j)})^2]\right\|_{op} \leq n\alpha_n \tag{9.8}$$

So we can apply matrix Bernstein inequality and we get the following upper bound:

$$\mathbb{P}\left(\|A - \mathbb{E}[A]\|_{op} > t\right) \leq 2n \exp\left\{-\frac{t^2}{2(n\alpha_n + t/3)}\right\} \tag{9.9}$$

Given the fact that $\alpha_n \geq c_1 \frac{\log(n)}{n}$, with $c_1 \in \mathbb{R}$ constant, then for any $r > 0 \in \mathbb{R}$ there exists a constant $c_2 = c_2(c_1, r)$ such that, by setting $t = \sqrt{c_2 n \alpha_n \log(n)}$, we get:

$$\mathbb{P}\left(\|A - \mathbb{E}[A]\|_{op} > t\right) \leq \frac{1}{n^r} \tag{9.10}$$

Another examples such as this one can be found in J.Tropp's monograph [JT15].

9.2 Linear Regression

Consider the case in which we observe n pairs of $(Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^d$. In linear regression we model the relationship between Y and X in the following way:

$$Y = X\beta^* + \epsilon \quad \implies \implies \mathbb{E}[Y] = X\beta^* \tag{9.11}$$

Where:

- $X \in \mathbb{R}^{n \times d}$ are assumed to be deterministic;
- $\epsilon = [\epsilon_1, ..., \epsilon_n]^T \in \mathbb{R}^n$ and we assume $\epsilon_i \in \text{SG}(\sigma^2)$ independent.

Equation 9.11 implies that the expected value is a linear function of β^*, but that does not imply that the model is linear.

Example: Let's consider the case in which we have the following:

$$Y_i = \sum_{j=1}^{d} \beta_j^* f_j(t_i) + \epsilon_i \tag{9.12}$$

Where $f_j : [0,1] \to \mathbb{R}$ and $\{t_1, ..., t_n\} \subset [0,1]$.
We can turn this into a linear regression by setting $X_{ij} = f_j(t_i)$, hence the model becomes linear in β^* but it is not linear given general f_j.

In reality this model is very flawed, as the following strong assumptions were are taking might not hold:

1. The true model might not be linear to start with;
2. There are no reasons to assume the X to be deterministic;
3. Errors might not have the same parameter σ^2 in practice.

A great reference to explore more of such problem is Buja et al. [AB14].
In this case a more correct framework would be the following:

1. $(Y, X) \in \mathbb{R} \times \mathbb{R}^d \sim P$ and we observe n pairs $(Y_1, X_1), ..., (Y_n, X_n)$;
2. We consider the regression function $\mu(x)$ without making any parametric assumption, and we then have:

$$\mathbb{E}[Y|X = x] = \mu(x) \tag{9.13}$$

In general we might use linear regression for two different goals:

1. **Prediction**
   Suppose we observe a new response vector $\tilde{X}$ and we would like to predict $\tilde{Y}$ based on our estimate of β^*, which we call $\widehat{\beta}$. We then want to minimize:

$$\frac{1}{n}\mathbb{E}_{Y,\tilde{Y}}\left[\left\|\tilde{Y} - \tilde{X}\widehat{\beta}\right\|^2\right] = \frac{1}{n}\mathbb{E}_Y\left[\left\|\tilde{X}(\beta^* - \widehat{\beta})\right\|^2\right] + \frac{1}{n}\mathbb{E}_\epsilon\left[\|\epsilon\|^2\right] \tag{9.14}$$

   With the equality given by the fact that we assume the model to be linear, so $\tilde{Y} = X\beta^* + \epsilon$. On the RHS, the second terms does not go to zero, as it's the noise σ^2, but the first does. Hence we want to minimize the mean squared error:

$$\min_{\widehat{\beta}\in\mathbb{R}^d} MSE(\widehat{\beta}) = \min_{\widehat{\beta}\in\mathbb{R}^d} \frac{1}{n}\mathbb{E}\left[\left\|\tilde{X}(\beta^* - \widehat{\beta})\right\|^2\right], \quad \text{where } X\beta^* = \mathbb{E}[Y] \tag{9.15}$$

2. **Parameter Estimation**
   In this case we are interested in the estimation of β^* and so we want to minimize:

$$\min_{\widehat{\beta}\in\mathbb{R}^d} \mathbb{E}\left[\left\|\beta^* - \widehat{\beta}\right\|^2\right] \tag{9.16}$$

9.2.1 Least Squares in High Dimension

Classical linear regression solves the least square problem with a solution of the form:

$$\widehat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{9.17}$$

It is the estimator optimal for β^* under strong assumption, in the sense that it is the one with minimal variance. However, it requires $X^T X$ to be invertible, so $n \geq d$ and matrix needs to be full rank as well. What is $d > n$? We could still find a solution to the following problem:

$$\min_{\widehat{\beta} \in \mathbb{R}^d} \|Y - X\beta\|^2 \tag{9.18}$$

Because the function that maps $\beta \to \|Y - X\beta\|^2$ is convex, so it is enough for us to check the first order optimality condition:

$$\nabla_\beta \|Y - X\beta\|^2 = 0 \Longleftrightarrow X^T X\beta = X^T Y \tag{9.19}$$

Any β satisfying the RHS of equation above will minimize $\beta \to \|Y - X\beta\|^2$. This solution is not unique because, given $\Delta \in \text{NullSpace}(X)$ and β satisfying first order optimality condition, then also $\beta + \Delta$ is a solution. In this case each solution can be written as follows:

$$\widehat{\beta} = \left(X^T X\right)^\dagger X^T Y + \left(I - \left(X^T X\right)^\dagger \left(X^T Y\right)\right) z \tag{9.20}$$

Where $\left(X^T X\right)^\dagger$ is the Moore-Penrose **Pseudo Inverse** of $\left(X^T X\right)$ and z any arbitrary vector $\in \mathbb{R}^d$ (the second part of the RHS is the way to express a vector in the Kernel(X)). A solution with minimal norm is the following:

$$\widehat{\beta} = \left(X^T X\right)^\dagger X^T Y \tag{9.21}$$

Definition 9.1 (Pseudo Inverse) *Let A a $n \times m$ matrix, then $A^\dagger$ is a $m \times n$ matrix called Moore-Penrose pseudo inverse if it satisfies the following properties:*

1. $AA^\dagger A = A$
2. $A^\dagger AA^\dagger = A$
3. $\left(AA^\dagger\right)^\dagger = AA^\dagger$
4. $\left(A^\dagger A\right)^\dagger = A^\dagger A$

In the case in which $m = n$ and A is invertible, then $A^\dagger = A^{-1}$.

In order to construct the pseudo inverse we can consider the singular value decomposition (SVD) of the matrix A:

$$A = UDV^T, \qquad D = \text{diag}(\sigma_1, ..., \sigma_k, 0, ..., 0) \tag{9.22}$$

Where $\sigma_1, ..., \sigma_k$ are the singular values of A and $k = \text{rank}(A) < \min\{n, m\}$. We can then construct the pseudo inverse in the following way:

$$A^\dagger = UD^{-1}V^T, \qquad D^{-1} = \text{diag}(\sigma_1^{-1}, ..., \sigma_k^{-1}, 0, ..., 0) \tag{9.23}$$

Here is a citation, just for fun.

References

[CR11] F. Chung and M. Radcliffe, "On the Spectra of General Random Graphs" *The Electronic Journal of Combinatorics*, 2001, Volume 18, Issue 1.

[JT15] J. Tropp, "An Introduction to Matrix Concentration Inequalities", *Found. Trends Mach. Learning*, 2015, Vol. 8, num. 1-2, pp. 1-230

[AB14] A. Buja and al., "The Conspiracy of Random Predictors and Model Violations against Classical Inference in Regression", *Statistical Science*, 2014