8.1 Continue on Matrix Bernstein Inequality

**Theorem 8.1** (Matrix Bernstein Inequality) \( X_1, \ldots, X_n \) are independent, zero mean, \( d \times d \) and symmetry matrices, such that \( \|X_i\|_{op} \leq C \), then we have

\[
P(\|\sum_{i=1}^{n} X_i\|_{op} \geq t) \leq 2d \exp\left\{ -\frac{t^2}{2(\sigma^2 + \frac{tC}{3})} \right\}
\]

where \( \sigma^2 = \|\sum_{i=1}^{n} EX_i^2\|_{op} = \|\sum_{i=1}^{n} Var(X_i)\|_{op} \)

8.1.1 Review the end of last lecture

**Proof:** The proof of Matrix Bernstein Inequality is following:

**Step 1:** Bound on moment generating function by Chernoff bound.

\[
P(\|\sum_{i=1}^{n} X_i\|_{op} \geq t) \leq P(\lambda_{\text{max}}(\sum_{i=1}^{n} X_i) \geq t) \leq e^{-\lambda t} E[\lambda_{\text{max}}(e^{\lambda \sum_{i=1}^{n} X_i})] \leq e^{-\lambda \sum_{i=1}^{n} X_i} E[tr(e^{\lambda \sum_{i=1}^{n} X_i})]
\]

**Step 2:** Apply monotonicity and Lieb’s inequality.

Note: The tool we have used in this step and may be in later steps:

1. Operator monotonicity of \( \log \): If \( 0 \leq A \leq B \), then \( \log(A) \leq \log(B) \)
2. Monotonicity of \( tr(e^A) \): If \( A \leq B \), then \( tr(e^A) \leq tr(e^B) \)
3. Lieb’s inequality

By these previous tools, we have

\[
P(\lambda_{\text{max}}(\sum_{i=1}^{n} X_i) \geq t) \leq \inf_{\lambda} \left\{ e^{\lambda t} tr\left(\exp\left\{ \sum_{i=1}^{n} \log E[e^{\lambda X_i}] \right\} \right) \right\}
\]

8-1
8.1.2 Continue on the proof

Step 3: Bound $E[e^{AX_i}]$.

**Lemma 8.2** Let a function $g : (0, \infty) \to [0, \infty)$, and $A_1, \ldots, A_n$ be PSD matrix such that $E[e^{AX_i}] \leq \exp\{g(\lambda)A_i\}$, $\lambda > 0$, $\forall i$. Then we have

$$P(\lambda_{\text{max}}(\sum_{i=1}^{n} X_i) \geq t) \leq d \cdot \inf_{\lambda} \exp\{-\lambda t + g(\lambda) \cdot \lambda_{\text{max}}(\sum_{i=1}^{n} A_i)\}$$

**Proof:** By log operator monotonicity, since $E[e^{AX_i}] \leq \exp\{g(\lambda)A_i\}$, we have $\log(E[e^{AX_i}]) \leq g(\lambda)A_i$.

By Monotonicity of $\text{tr}(e)$, we have $\text{tr}(\exp\{\sum_{i=1}^{n} \log(E[e^{AX_i}])\}) \leq \text{tr}(\exp\{\sum_{i=1}^{n} g(\lambda)A_i\})$

Notice that $\text{tr}(\Sigma) \leq d \cdot \lambda_{\text{max}}(\Sigma)$. Therefore, after extract all the constants terms, we have

$$P(\lambda_{\text{max}}(\sum_{i=1}^{n} X_i) \leq \inf_{\lambda} \{e^{\lambda} \cdot \text{tr}(\exp\{\sum_{i=1}^{n} \log(E[e^{AX_i}])\})\}$$

$$\leq \inf_{\lambda} \{e^{\lambda} \cdot \text{tr}(\exp\{\sum_{i=1}^{n} g(\lambda)A_i\})\}$$

$$\leq \inf_{\lambda} \{e^{\lambda} \cdot d \cdot \lambda_{\text{max}}(\exp\{\sum_{i=1}^{n} g(\lambda)A_i\})\}$$

$$\leq d \cdot \inf_{\lambda} \exp\{-\lambda t + g(\lambda) \cdot \lambda_{\text{max}}(\sum_{i=1}^{n} A_i)\}$$

Then we have $P(\lambda_{\text{max}}(\sum_{i=1}^{n} X_i) \leq d \cdot \inf_{\lambda} \exp\{-\lambda t + g(\lambda) \cdot \|\sum_{i=1}^{n} A_i\|_{\text{op}}\}$ if we have $E[e^{AX_i}] \leq \exp\{g(\lambda)A_i\}$.

Notice that if we assume that $\|X_i\|_{\text{op}} \leq 1$, thus let $C = 1$.

**Lemma 8.3** $E[e^{AX_i}] \leq \exp\{(e^{\lambda} - \lambda - 1)E[X_i^2]\}$

**Proof:** Define the function

$$f_{\lambda}(x) = \begin{cases} \frac{e^{\lambda x} - \lambda x - 1}{x^2} & \text{if } x \neq 0 \\ \frac{\lambda^2}{2} & \text{if } x = 0 \end{cases}$$

By taking the first derivative, we know that $f_{\lambda}(x)$ is increasing thus if $x \leq 1$, then $f_{\lambda}(x) \leq f_{\lambda}(1)$, thus $f_{\lambda}(X_i) \leq f(1) \cdot I$.

By spectral theorem and the assumption of $\|X_i\|_{\text{op}}$, we have

$$e^{AX_i} = I + \lambda X_i + X_i^T f_{\lambda}(X_i) X_i \leq I + \lambda X_i + f_{\lambda}(1)X_i^2$$

Take expectation for both side and use the fact that $1 + x \leq e^x$ and $X_i$ is zero mean, we will have

$$E[e^{AX_i}] \leq E[I + \lambda X_i + f_{\lambda}(1)X_i^2] = I + f_{\lambda}(1)E[X_i^2] \leq \exp\{f_{\lambda}(1)E[X_i^2]\} = \exp\{(e^{\lambda} - \lambda - 1)E[X_i^2]\}$$
Step 4: Apply two lemmas in the step 3 and warp the proof.

Let $A_i = E[X_i^2]$, and $g(\lambda) = e^\lambda - \lambda - 1$, then based on Lemma 8.3, we have

$$E[e^{\lambda X_i}] \leq \exp((e^\lambda - \lambda - 1)E[X_i^2]) = \exp(g(\lambda)A_i)$$

Notice that $A_i = E[X_i^2]$ is PSD matrix thus we satisfied the condition of Lemma 8.2.

Apply Lemma 8.2 and notice that let $A = \begin{bmatrix} A_{i1} & \cdots & A_{in} \end{bmatrix}$

Recall in step 3, we have the assumption $\lambda_{\max}(\sum_{i=1}^n A_i) \leq 1$, but in the question we only have the condition $\|X_i\|_{op} \leq C$, therefore, we replace $X_i$ as $\frac{\sigma}{C}$ in the inequality and have

$$P(\lambda_{\max}(\sum_{i=1}^n \frac{X_i}{C}) \geq t) \leq d * \inf_{\lambda} \exp\{-\lambda t + g(\lambda) \lambda_{\max}(\sum_{i=1}^n A_i)\} \leq d * \inf_{\lambda} \exp\{-\lambda t + g(\lambda) \sigma^2 \}$$

Minimize the RHS to achieve the narrowest bound by taking the derivative and set ot zero, we have $-t + (e^\lambda - 1) \sigma^2 = 0$, then solve the equation and $\inf_{\lambda} = \log\left(\frac{t}{C} + 1\right)$.

By plugging in the $\inf_{\lambda}$, and let $t^* = tC$, then notice that $\inf_{\lambda} = \log\left(\frac{tC}{C} + 1\right)$ now, and we will have

$$P(\lambda_{\max}(\sum_{i=1}^n X_i) \geq t^*) = P(\lambda_{\max}(\sum_{i=1}^n X_i) \geq tC) \leq d * \inf_{\lambda} \exp\{-\lambda \frac{t^*}{C} + g(\lambda) \frac{\sigma^2}{C^2}\}$$

$$= d * \exp\{-\frac{t^*}{C} \log\left(\frac{t^* C}{\sigma^2} + 1\right) + \left(\frac{t^* C}{\sigma^2} + 1 - \log\left(\frac{t^* C}{\sigma^2} + 1\right) - 1\right) \frac{\sigma^2}{C^2}\}$$

$$= d * \exp\left\{\frac{t^*}{C} - \left(\frac{t^* C}{\sigma^2} + \frac{\sigma^2}{C^2} \log\left(\frac{t^* C}{\sigma^2} + 1\right)\right)\right\}$$

$$= d * \exp\left\{\frac{\sigma^2}{C^2} - \left(\frac{t^* C}{\sigma^2} + 1\right) \log\left(\frac{t^* C}{\sigma^2} + 1\right)\right\}$$

If we let $h(u) = (1 + u) \log(1 + u) - u$ and replace the $t^*$ with $t$, then we have

$$P(\lambda_{\max}(\sum_{i=1}^n X_i) \geq t) \leq d * \exp\{-\frac{\sigma^2}{C^2} h\left(\frac{tC}{\sigma^2}\right)\} \quad \text{where} \quad u = \frac{tC}{\sigma^2}$$

Notice that $h(u) \geq \frac{u^2}{2(1 + u)}$, then

$$P(\lambda_{\max}(\sum_{i=1}^n X_i) \geq t) \leq d * \exp\{-\frac{\sigma^2}{C^2} \frac{u^2}{2(1 + \frac{u}{C})}\} = d * \exp\{-\frac{\sigma^2}{C^2} \frac{\left(\frac{tC}{\sigma^2}\right)^2}{2(1 + \frac{tC}{\sigma^2})}\} = d * \exp\{-\frac{t^2}{2(\sigma^2 + \frac{tC}{\sigma^2})}\}$$

where $\sigma^2 = \|\sum_{i=1}^n E[X_i^2]\|$.

Based on the value of $t$, we have

$$P(\lambda_{\max}(\sum_{i=1}^n X_i) \geq t) \leq \begin{cases} d * \exp\{-\frac{3t^2}{8\sigma^2}\} & \text{if} \quad t \leq \frac{\sigma^2}{C} \\ d * \exp\{-\frac{3t^2}{16\sigma^2}\} & \text{if} \quad t > \frac{\sigma^2}{C} \end{cases}$$

Lecture 8: September 27

8-3
8.2 Extension and Remarks on Matrix Bernstein Inequality

1. There exists a Bounded in Expectation version, if all the assumption in Theorem 8.1 are satisfied, then

\[ E[|| \sum_{i=1}^{n} X_i ||_{op}] \leq C \cdot [\sigma \sqrt{\log(d)} + C \cdot \log(d)] \]

where \( \sigma = \sqrt{|| \sum_{i=1}^{n} EX_i^2 ||} \)

2. There exists a Bounded difference inequality version.

**Theorem 8.4** Let \( x = (x_1, \ldots, x_n) \) be independent random variables and there exists a function \( H \) such that \( H : \mathbb{R}^n \to \mathbb{R}^{d \times d} \). If there exists a sequence of matrix \( A_i \) such that \( (H(x_1, \ldots, x_i, \ldots, x_n) - H(x_1, \ldots, x_i', \ldots, x_n)) \leq A_i^2 \) for all \( i = 1, \ldots, n \) and let \( \sigma^2 = || \sum_{i=1}^{n} A_i^2 || \), then we have

\[ P(\lambda_{\max}(H(x) - EH(x)) \geq t) \leq d \cdot e^{-\frac{t^2}{2\sigma^2}} \]

3. Weakening the assumption is possible (proof in the book). The weakened Bernstein condition is

\[ E[X_i^p] \leq \frac{p!}{2} C^{p-2} E[X_i^2], \quad \text{for } p = 3, 4, \ldots \]

4. We define \( X \) are \( d \times d \), symmetry, zero mean is sub-Gaussian(\( \Sigma \)) matrix if \( E[e^{\lambda X}] \leq \exp\{\frac{\lambda^2}{2} \Sigma\} \), for some PD matrix \( \Sigma \) and \( \forall \lambda \in \mathbb{R} \), or is sub-Exponential(\( V, \alpha \)) if \( E[e^{\lambda X}] \leq \exp\{\frac{\lambda^2}{2} V\} \), for some PD matrix \( V \) and \( \forall |\lambda| \leq \frac{1}{4} \).

If we insert this bound to the Matrix Bernstein Inequality, we will have a Hoeffding/Bernstein inequality which is

\[ P(|| \sum_{i=1}^{n} X_i || \geq t) \leq 2d \exp\{-\frac{t^2}{2\sigma^2}\} \]

if \( X_i \in SG(\Sigma_i) \) are independent and \( \sigma^2 = || \sum_{i=1}^{n} \Sigma_i || \)

5. The theorem can be extended to no-symmetric or rectangular matrix with Jordan-Wielandt theorem (Steward & Sum, 1990).

If \( B \) is a \( d_1 \times d_2 \) matrix or a \( d \times d \) but not symmetric matrix, let \( A \) be the pilation of \( B \) such that \( A = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} \).

Then \( A \) will be a \( (d_1+d_2) \times (d_1+d_2) \) and symmetric. Since \( A^2 = \begin{bmatrix} BB^T & 0 \\ 0 & B^T B \end{bmatrix} \), then \( A \)'s non-zero eigenvalues are \( \pm \) singular value of \( B \) and \( ||A|| = ||B|| \). Then the matrix inequality can be re-written as,

\[ P(\lambda_{\max}(\sum_{i=1}^{n} B_i) \geq t) \leq (d_1 + d_2) \cdot \exp\{-\frac{t^2}{2(\sigma^2 + \frac{tC}{\sqrt{3}})}\} \]

where \( \sigma^2 = \max\{|| \sum_{i=1}^{n} E[B_i B_i^T] ||, || \sum_{i=1}^{n} E[B_i^T B_i] ||\} \)
8.3 Application of Matrix Bernstein Inequality

8.3.1 Covariance Estimation

**Theorem 8.5** Let $X_1, \ldots, X_n$ are independent, zero mean vectors in $\mathbb{R}^d$ such that $||X_i||^2 \leq C_d, \forall i$. Let $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T$, then,

$$P(||\hat{\Sigma} - \Sigma||_op \geq t) \leq 2d \exp \left\{ - \frac{nt^2}{2C_d(||\Sigma|| + \frac{t}{3})} \right\}$$

**Proof:** Let $Q_i = X_i X_i^T - \Sigma$, then $Q_i$ is symmetric and zero mean. Consider,

$$||Q_i||_op \leq ||X_i X_i^T||_op + ||\Sigma||_op = ||X_i||^2 + ||\Sigma||_op \leq C_d + ||\Sigma||_op$$

$$||\Sigma||_op = \max_{z \in \mathbb{S}^{d-1}} z^T E [X_i X_i^T]^2 = \max_{z \in \mathbb{S}^{d-1}} E [(z^T X_i)^2] \leq ||z||^2 ||X_i||^2 \leq 1 * C_d = C_d$$

Thus, $||Q_i||_op \leq 2C_d$. Since $Q_i$ is zero mean, then,

$$E Q_i^2 = \text{Var}[Q_i] = E [(X_i X_i^T)^2] - \Sigma^2 \leq E [(X_i X_i^T)^2] = E [||X_i||^2 X_i X_i^T] = C_d * E [X_i X_i^T] = C_d \Sigma$$

Thus, $||E Q_i^2||_op \leq C_d ||\Sigma||_op$. Therefore, let $\sigma^2 = ||\sum_{i=1}^{n} E Q_i^2||_op = nC_d ||\Sigma||_op$ and $C = 2C_d$. By applying the Matrix Bernstein Inequality and the extension 8.3.5, we have

$$P(||\hat{\Sigma} - \Sigma||_op \geq t) = P(||\sum_{i=1}^{n} Q_i||_op \geq nt) \leq 2d \exp \left\{ - \frac{n \sigma^2 t^2}{2n + 3} \right\} = 2d \exp \left\{ - \frac{n t^2}{2C_d(||\Sigma|| + \frac{t}{3})} \right\}$$

If we assume that $||X_i|| \leq K \sqrt{E[||X_i||^2]} = K \sqrt{\text{Tr}(\Sigma)} \leq K \sqrt{d ||\Sigma||_op}$, then $C_d = K^2 d ||\Sigma||_op$. In this case, with high probability, we have,

$$\frac{||\hat{\Sigma} - \Sigma||_op}{||\Sigma||_op} \leq C \max \left\{ \sqrt{\frac{d * \log(d)}{n}}, \frac{d * \log(d)}{n} \right\}$$

8.3.2 Random Graph

Let $A$ be a $n \times n$ symmetric matrix with 0 element on the diagonal and $A_{ij} \in \{0, 1\}$, for $i \neq j$. We can consider $A_{ij} \sim \text{Bernoulli}(p_{ij})$ independently. $A$ is usually called the adjacency matrix of a graph on node $\{1, 2, \ldots, n\}$ such that $i$ and $j$ are connected if $A_{ij} = 1$, and we have $\binom{n}{2}$ independent Bernoulli.

References

