Lecture 17

Last time

- Covariance Estimation
- Clustering a Gaussian Mixture.

Today

- Matrix Calculus
- Matrix Chernoff
- Matrix Hoeffding

Matrix Calculus

Our goal today is to show the following extension of Hoeffding's inequality for matrices

Ind. mean-zero

$$\mathbb{P}\left(\left\| \sum_{i=1}^{n} A_i \right\|_{op} \geq t \right) \leq ???$$

We'll need some notation.

Def (Functions of matrices): Given a function $f: \mathbb{R} \to \mathbb{R}$ and a $X \in S^d$ with spectral decomposition

$$X = \sum_{i=1}^{d} \lambda_i u_i u_i^T \quad \text{define} \quad f(X) := \sum_{i=1}^{d} f(\lambda_i) u_i u_i^T.$$

$\dashv$

Examples:

- If $f(t) = t^{-1} \Rightarrow f(X) = X^{-1}$
- If $f(t) = \sum_{p=1}^{\infty} \alpha_p t^p \Rightarrow f(X) = \sum_{p=1}^{\infty} \alpha X^p$

We can Taylor expand!

$\dashv$

Just as with scalars, symmetric matrices also have an order.

Def (Löwner ordering): Given $X, Y \in S^d$ we say that $X \preceq Y$ if

$$Y - X \in S^d_+.$$

$\dashv$

Lemma (*) Suppose that $X \preceq Y$, then:

1) For all $A \in \mathbb{R}^{d \times k}$,

$$A^T X A \preceq A^T Y A.$$

2) All eigenvalues satisfy

ith largest

$$\lambda_i(X) \leq \lambda_i(Y)$$

3) Let $f : \mathbb{R} \to \mathbb{R}$ be nonincreasing, then

$$\operatorname{tr}(f(X)) \leq \operatorname{tr}(f(Y)).$$

$\dashv$

Given the last item it is natural to wonder whether

$$(\checkmark) \quad X \preceq Y \quad \Rightarrow \quad f(X) \preceq f(Y)$$

for nonincreasing f. In general this is not the case. Indeed,

$$t \mapsto e^t$$

is a counterexample (check!)

Def (Matrix monotone) A function is matrix monotone if (✓) holds $\forall X, Y$. $\dashv$

Lemma (:) The functions $t \mapsto t^{-1}$, $t \mapsto t^{1/2}$, and $t \mapsto \log t$ are matrix monotone. $\dashv$

Matrix Chernoff

Recall the strategy we use for scalar random variables:

① Sub-Gaussian MGF + Markov's $\Rightarrow$ Tail bounds for 1 r.v.

② Sum of iid sub-Gaussian + Peeling argument $\Rightarrow$ Sum is also sub-Gaussian.

In turn, the peeling argument is delicate. Define the MGF of a random matrix Q as $\Psi_Q : \mathbb{R} \to S^d$ given by

$$\Psi_Q(\gamma) = \mathbb{E}[e^{\gamma Q}] = \sum_{k=0}^{\infty} \frac{\gamma^k}{k!} \mathbb{E}[Q^k].$$

Just as in the scalar case this controls concentration.

Lemma (Matrix Chernoff Method): Let $\mathcal{A}$ a random sym. matrix with MGF defined in some interval $(-a, a)$. Then for all $t > 0$, we have

$$\mathbb{P}(\lambda_1(\mathcal{A}) \geq t) \leq \operatorname{tr}(\Psi_{\mathcal{A}}(\gamma)) e^{-\gamma t} \quad \forall \gamma \in [0, a).$$

As a consequence

$$\mathbb{P}(\underbrace{\|\mathcal{A}\|_{op}}_{\max\{\lambda_1(\mathcal{A}),\, -\lambda_n(\mathcal{A})\}} \geq t) \leq 2 \operatorname{tr}(\Psi_{\mathcal{A}}(\lambda)) e^{-\gamma t} \quad \forall \gamma \in [0, a)$$

Proof: Taking scalar exponentials

$$\begin{aligned}
\mathbb{P}(\lambda_1(\mathcal{A}) \geq t) &= \mathbb{P}(e^{\lambda_1(\gamma \mathcal{A})} \geq e^{\gamma t}) \\
&= \mathbb{P}(\lambda_1(e^{\gamma \mathcal{A}}) \geq e^{\gamma t}) \\
&\overset{\text{Markov's}}{\leq} \mathbb{E}\, \lambda_1(e^{\gamma \mathcal{A}})\, e^{-\gamma t} \\
&\overset{\operatorname{tr}(A) = \sum \lambda_i(A)}{\leq} \mathbb{E} \operatorname{tr}(e^{\gamma \mathcal{A}})\, e^{-\gamma t} \\
&= \operatorname{tr} \Psi_{\mathcal{A}}(\gamma)\, e^{-\gamma t}.
\end{aligned}$$

$\square$

This seems to suggest that we want $\Psi_{\mathcal{A}}(\gamma)$ to be nicely controlled by a "Gaussian tail" as before.

Def: A random symmetric matrix is V-sub-Gaussian if

$$\Psi_{\mathbb{Q}}(\lambda) \leq e^{\lambda V/2} \quad \forall \lambda \in \mathbb{R}. \top$$

Example: Suppose $\mathbb{Q} = \varepsilon B$ with $\varepsilon \sim \mathrm{Unif}\{\pm 1\}$ and $B \in S^d$ fixed. Then $\mathbb{E}\,\mathbb{Q}^{2k+1} = 0$ and $\mathbb{E}\,\mathbb{Q}^{2k} = B^{2k}$. So,

$$\mathbb{E}\, e^{\lambda \mathbb{Q}} = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} B^{2k} \underset{(\text{why?})}{\preceq} \sum_{k=0}^{\infty} \frac{1}{k!}\left(\frac{\lambda^2 B^2}{2}\right) = e^{\lambda^2 B^2/2}. \top$$

Unlike before we don't have a sum rule because

$$e^{A+B} \neq e^A e^B$$

in general (it only holds for commuting matrices). To "fix" the peeling argument we will use a deep result from analysis.

Theorem (Lieb inequality): Let $H \in S^d$, and define $f: S^d_+ \to \mathbb{R}$ given by

$$f(X) = \operatorname{tr} \exp(H + \log X).$$

Then, f is concave on S_+^d. $\dashv$

We will not prove this result.

Lemma (1.3) Let $a_1, \ldots, a_n \in S^d$ be independent with $\psi_{a_i}(\cdot)$ defined over an interval $J \subseteq \mathbb{R}$. Let $S_n = \sum_{i=1}^n a_i$. Then,

$$\operatorname{tr}(\psi_{S_n}(\gamma)) \le \operatorname{tr}\left(\exp\left(\sum_{i=1}^n \log \psi_{a_i}(\gamma)\right)\right)$$

$\forall \gamma \in J$.

Consequently, Chernoff method

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum^n a_i\right\|_{op} \ge t\right) \le 2\operatorname{tr}\left(e^{\sum^n \log \psi_{a_i}(\gamma)}\right)e^{-\gamma t}.$$

Proof: Expanding

$$\operatorname{tr}(\psi_{S_n}(\gamma)) = \operatorname{tr} \mathbb{E} e^{\gamma S_n}$$

$$= \operatorname{tr} \mathbb{E} e^{\gamma S_{n-1} + \log \exp(\gamma a_n)}$$

$$= \mathbb{E}_{S_{n-1}} \mathbb{E}_{a_n} \operatorname{tr} e^{\gamma S_{n-1} + \log \exp(\gamma a_n)}$$

Lieb + Jensen's

$$\le \mathbb{E}_{S_{n-1}} \operatorname{tr} e^{\gamma S_{n-1} + \log \psi_{a_n}(\gamma)}$$

$$\leq \cdots$$

Recursing the same argument $\rightarrow$ $\leq \operatorname{tr} e^{\sum_{i=1}^n \log \psi_{a_i}(\gamma)}$.

$\square$

Theorem (Hoffding): Suppose $a_1, \ldots, a_n$ are zero-mean, V_i-sub-Gaussian random matrices in S^d. Then,

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n a_i\right\|_{op} \geq t\right) \leq 2\,\operatorname{rank}\left(\sum V_i\right) e^{\frac{-nt^2}{2\sigma^2}}$$
$$\leq 2d\, e^{-\frac{nt^2}{2\sigma^2}} \qquad \forall t>0,$$

where $\sigma^2 = \left\|\frac{1}{n}\sum V_i\right\|_{op}$.

Proof: Let $V = \sum_{i=1}^n V_i$. From Lemma ((ii)) it suffices to bound $\operatorname{tr}(\exp(\sum \log \psi_{a_i}(\lambda)))$. By sub-Gaussianity and the monotonicity of the matrix log (Lemma (i))

$$\sum_{i=1}^n \log \psi_{a_i}(\gamma) \preceq \frac{\gamma^2}{2}\sum_{i=1}^n V_i.$$

Moreover since $t \mapsto e^t$ is increasing

Lemma (*) gives

$$\operatorname{tr}\left(\exp\left(\sum_{i=1}^{n} \log \varphi_{a_i}(\gamma)\right)\right) \leq 2\operatorname{tr}\left(e^{\frac{\gamma^2}{2}V}\right).$$

Thus, by Lemma (((·))),

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i}^{n} a_i\right\|_{op} \geq t\right) \leq 2\operatorname{tr}\left(e^{\frac{\gamma^2 V}{2}}\right)e^{-\gamma n t}.$$

Note that

$$\operatorname{tr}(e^{A}) \leq \operatorname{rank}(A)\, e^{\|A\|_{op}},$$

moreover $\left\|\frac{\gamma^2 V}{2}\right\|_{op} = \frac{\gamma^2}{2} n \sigma^2$. So

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum a_i\right\|_{op} \geq t\right) \leq 2\operatorname{rank}(V)\, e^{\frac{\gamma^2}{2}n\sigma^2 - \gamma n t}.$$

The best bound is given by taking $\gamma = t/\sigma^2$, which yields the claim. $\square$

Remark: The additional d factor in the bound is in general unavoidable.