10-704: Information Processing and Learning **Spring 2012**

Lecture 1: September 20

Lecturer: Alessandro Rinaldo *Scribes: Xiaoyi Gu*

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

1.1 Discretization

Definition 1.1 *(sub-gaussian vector) $X \in \mathbb{R}^d$ is $SG(\sigma^2)$ when $X^T v \in SG(\sigma^2)$ for all $v \in S^{d-1}$.*

Example: 1. If coordinates of X are independent $SG(\sigma^2)$.
2. $X \sim N_d(0, \Sigma) \Rightarrow X \in SG(\|\Sigma\|_{op})$. (exercise)

Theorem 1.2 *Let $X \in \mathbb{R}^d$ be $SG(\sigma^2)$, then $\mathbb{E}[\|X\|] \leq 4\sigma\sqrt{d}$ and $\|X\| \leq 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}$ for all $\delta \in (0,1)$ with probability $\geq 1-\delta$.*

Proof: First notice that $\|X\| = (\sum_{i=1}^d X_i^2)^{1/2} = \max_{\theta \in B_d} \theta^T X$ where B_d is the unit ball in $\mathbb{R}^d$. Let $N_{1/2}$ be a $\frac{1}{2}$ covering of B_d in Euclidean norm. Then $|N_{1/2}| \leq (1 + \frac{2}{1/2})^d \leq 5^d$. Next, $\forall \theta \in B_d$, $\exists\ z = z(\theta) \in N_{1/2}$ such that $\|\theta - z\| \leq 1/2$, or equivalently, $\exists\ w$ such that $\theta = z + w$ and $\|w\| \leq 1/2$. So

$$\max_{\theta \in B_d} \theta^T X \leq \max_{z \in N_{1/2}} z^T X + \max_{w \in \frac{1}{2} B_d} w^T X.$$

Notice that $\max_{w \in \frac{1}{2}B_d} w^T X = \frac{1}{2}\max_{\theta \in B_d} \theta^T X$, we get

$$\max_{\theta \in B_d} \theta^T X \leq 2 \max_{z \in N_{1/2}} z^T X.$$

(In fact, it holds that $\|X\| \leq \frac{1}{1-\varepsilon} \max_{z \in N_{1/2}} z^T X$ for $\varepsilon < 1$).
Then by maximal inequality for sub-gaussians,

$$\mathbb{E}[\|X\|] = \mathbb{E}[\max_{\theta \in B_d} \theta^T X] \leq 2\sigma\sqrt{2\log|N_{1/2}|} \leq 4\sigma\sqrt{d}$$

since $\log|N_{1/2}| \leq d\log 5$.
Next, for all $t \geq 0$,

$$\begin{aligned}
\mathbb{P}(\|X\| \geq t) &= \mathbb{P}(\max_{\theta \in B_d} \theta^T X \geq t) \\
&\leq \mathbb{P}(\max_{z \in N_{1/2}} z^T X \geq t/2) \\
&\leq \sum_{z \in N_{1/2}} \mathbb{P}(z^T X \geq t/2) \\
&\leq |N_{1/2}| e^{-t^2/8\sigma^2} \\
&\leq 5^d e^{-t^2/8\sigma^2}.
\end{aligned}$$

Set RHS $\leq \delta \in (0,1)$ and solve for δ, we get $t = \sqrt{8 \log 5}\sqrt{d}\sigma + 2\sigma\sqrt{2\log(1/\delta)}$. ∎

Remark: The same argument will lead to bounds on $\|A\|_{op}$ using the fact

$$\|A\|_{op} = \max_{x \in S^{d-1}} \|Ax\| \leq \frac{1}{1-\varepsilon} \max_{x \in N_\varepsilon} \|Ax\|$$

for $\varepsilon \in (0,1)$.

1.2 Covariance Matrix Estimation in $\|\cdot\|_{op}$ Norm

Let Σ be a $d \times d$ PSD matrix, $X_1, ..., X_n \sim N(0, \Sigma)$ i.i.d. satisfying the sub-gaussian property. The covariance matrix estimator $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$. Then

$$\max_{i,j} |\hat{\Sigma}_{ij} - \Sigma_{ij}| \leq C\sqrt{\frac{t + \log d}{n}}$$

with probability $\geq 1 - e^{-t}$. This result is consistent even if $d = e^n$.

Before moving on, let's review our matrix algebra.

1.2.1 Review of Matrix Algebra

Singular Value Decomposition(SVD)

Let A be an $m \times n$ matrix, SVD asserts that A can be decomposed into $A = UDV^T$, where D is an $r \times r$ diagonal matrix, or $D = \text{diag}(\sigma_1, ..., \sigma_r, 0, ..., 0)$ for $\sigma_1 \geq ... \geq \sigma_r > 0$ singular values and $r = \text{rank}(A)$. U is an $m \times r$ matrix of orthonormal columns, which are the left singular vectors of A, and V is an $r \times n$ matrix of orthonormal columns that are the right singular vectors of A.

Operator Norm

- Note that σ_1 is the largest singular value of A. The operator norm of A is defined as its largest singular value and the following equalities hold:

$$\|A\|_{op} = \sigma_1 = \max_{x \in \mathbb{R}^n, \|x\| \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{x \in S^{n-1}} \|Ax\| = \max_{y \in S^{n-1}} \max_{x \in S^{n-1}} y^T Ax.$$

  This defines a norm over $m \times n$ matrices.

- If A is symmetric, then

$$\|A\|_{op} = \max_{x \in S^{n-1}} |x^T Ax|.$$

- If A is PSD ($x^T Ax \geq 0 \ \forall x \in \mathbb{R}^n$), then

$$\|A\|_{op} = \max_{x \in S^{n-1}} x^T Ax = \lambda_{max}(A)$$

  where $\lambda_{max}(A)$ is the largest eigenvalue of A.

- The Frobenius norm of A is defined as $\|A\|_F = (\sum_i \sum_j A_{ij}^2)^{1/2}$.
- Fact about $\|\cdot\|_{op}$: $\|Ax\| \leq \|A\|_{op}\|x\|$ for all x.
- Weyl Inequality: If A and B are $m \times n$ matrices with singular values $\sigma_1(A) \geq ... \geq \sigma_{\min\{m,n\}}(A)$ and $\sigma_1(B) \geq ... \geq \sigma_{\min\{m,n\}}(B)$, then $\max_k |\sigma_k(A) - \sigma_k(B)| \leq \|A - B\|_{op}$.

Now we continue with covariance matrix estimation.

Theorem 1.3 *Let $X_1, ..., X_n \in \mathbb{R}^d$ be i.i.d vectors of mean 0 and covariance Σ such that $X_i \in SG(\sigma^2)$ for all i. Then, for $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$, we have*

$$\mathbb{P}(\frac{\|\Sigma - \hat{\Sigma}\|_{op}}{n} \leq C \max\{\sqrt{\frac{d + \log(2/\delta)}{n}}, \frac{d + \log(2/\delta)}{n}\}) \geq 1 - \delta$$

for $\delta \in (0, 1)$.

Note: 1. If $X_i \sim N(0, \Sigma)$, $\sigma^2 = \lambda_{max}(\Sigma) = \|\Sigma\|_{op}$.
2. If $d \gg n$, the result is not consistent.

Proof: Use the discretization argument and the following fact from HW1,

$$X \in SG(\sigma^2) \Rightarrow X^2 - \mathbb{E}[X^2] \in SE(\alpha^2, \nu)$$

where $\alpha = \nu = 16\sigma^2$ and the fact from class that $\mathbb{E}[|X|^k] \leq (2\sigma^2)^{k/2} k\Gamma(k/2)$.
To set up discretization argument, need

Lemma 1.4 *Let A be symmetric and N_ε an ε-covering of S^{d-1}, $\varepsilon \in (0, 1)$. Then*

$$\|A\|_{op} = \underset{x \in S^{d-1}}{max} |x^T A x| \leq \frac{1}{1 - 2\varepsilon} \underset{Z \in N_\varepsilon}{max} |Z^T A Z|$$

Proof: We have to consider 2 cases:
case 1: $\|A\|_{op} = \underset{x \in S^{d-1}}{max}\, x^T A x$
case 2: $\|A\|_{op} = \underset{x \in S^{d-1}}{max}\, -x^T A x$.
Regardless, let x^* be the point in S^{d-1} achieves the optimum and let $z = z(x^*) \in N_\varepsilon$ s.t. $\|z - x^*\| \leq \varepsilon$. Then

$$\begin{aligned}
|(x^*)^T A x^* - z^T A z| &= |z^T A z - (x^*)^T A x^*| \\
&= |(x^*)^T A (x^* - z) + z^T A (x^* - z)| \\
&\leq |(x^*)^T A (x^* - z)| + |z^T A (x^* - z)| \\
&\leq \|x^*\| \|A(x^* - z)\| + \|z\| \|A(x^* - z)\| \text{ by Holder} \\
&\leq \|x^*\| \|A\|_{op} \|x^* - z\| + \|z\| \|A\|_{op} \|x^* - z\| \\
&\leq 2\varepsilon \|A\|_{op}
\end{aligned}$$

So, for case 1,

$$\|A\|_{op} = (x^*)^T A x^* \leq 2\varepsilon \|A\|_{op} + z^T A z.$$

For case 2,

$$\|A\|_{op} = -(x^*)^T A x^* \leq 2\varepsilon \|A\|_{op} - z^T A z.$$

Take maximum over $z \in N_\varepsilon$ on RHS to get the result. ■
Set $Q = \hat{\Sigma} - \Sigma$, symmetric,
let $\{v_1, \dots, v_N\}$ be a 1/4-covering of $B_d \implies N \leq q^d$. So $\|Q\|_{op} \leq 2 \underset{i=1,\dots,N}{max} |v_i^T Q v_i|$ by lemma. Hence, $\forall t > 0$,

$$\begin{aligned}\mathbb{P}(\|Q\|_{op} \geq t) &\leq \mathbb{P}\left(\underset{i=1,\dots,N}{max} |v_i^T Q v_i| \geq \frac{t}{2}\right) \\ &\leq \sum_{i=1}^{N} \mathbb{P}\left(|v_i^T Q v_i| \geq \frac{t}{2}\right).\end{aligned}$$

To be continued ...

■

References

[PM07] P. Massart, "Concentration inequalities and model selection," *Berlin: Springer*, 2007, Vol. 6.

[ML05] M. Ledoux, "The concentration of measure phenomenon," *American Mathematical Soc.*, 2005, No. 89.