Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 16

Yun Yang

- Random matrices and covariance estimation

Bounds for general matrices: Background

- Matrix-valued function: Any function $f : \mathbb{R}^d \to \mathbb{R}^d$ can be extended to a map from $\mathcal{S}^{d\times d}$ to itself through

$$f(Q) = U\,\text{diag}(\gamma(Q))U^T,$$

  where $Q = U\,\text{diag}(\gamma(Q))U^T$ is the SVD of $Q \in \mathcal{S}^{d\times d}$.
- Unitary invariant property: for any unitary matrix V,

$$f(VQV^T) = Vf(Q)V^T.$$

- Spectral mapping property: the eigenvalues of the $f(Q)$ are simply the eigenvalues of Q transformed by f.
- Examples: matrix exponential $e^Q = \sum_{k=0}^{\infty} \frac{Q^k}{k!}$, defined for all $Q \in \mathcal{S}^{d\times d}$; matrix logarithm $\log Q$, defined for all $Q \succ 0$.

Tail conditions for matrices

- Moments: jth moment of a symmetric random matrix Q is defined by $\mathbb{E}[Q^j]$.
- Variance: $\mathrm{Var}(Q) = \mathbb{E}[Q^2] - (\mathbb{E}[Q])^2 \succeq 0$ (Exercise).
- If Q has polynomial moments of all orders, then its cumulative generating function $\Pi_Q : \mathbb{R} \to \mathcal{S}^{d\times d}$ is given by

$$\Pi_Q(\lambda) = \log \mathbb{E}[e^{\lambda Q}].$$

Definition

A zero-mean symmetric random matrix $Q \in \mathcal{S}^{d\times d}$ is sub-Gaussian with matrix parameter $V \in \mathcal{S}_+^{d\times d}$ if

$$\Pi_Q(\lambda) \preceq \frac{\lambda^2 V}{2} \quad \text{for all } \lambda \in \mathbb{R}.$$

Example

- $Q = \varepsilon B$, where B is a fixed symmetric matrix, and ε is a Rademacher random variable.
- We have $\mathbb{E}[Q^k] = 0$ for odd k, and $\mathbb{E}[Q^k] = B^k$ for even k. Therefore, we have

$$\mathbb{E}[e^{\lambda Q}] = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} B^{2k} \preceq e^{\frac{\lambda^2 B^2}{2}},$$

  implying that Q is sub-Gaussian with parameter $V = \sigma^2 B^2$.
- Now suppose B is a symmetric random matrix, independent of ε, that satisfies $\|B\|_{\mathrm{op}} \leq b$.
- Then $\Pi_Q(\lambda) \preceq \frac{\lambda^2 b^2}{2} I_d$, implying that Q is sub-Gaussian with parameter $V = b^2 I_d$.

Sub-exponential random matrices and Bernstein condition

Definition

A zero-mean random matrix is sub-exponential with parameters (V, b) if its cumulant function $\Phi_Q(\lambda) \preceq \frac{\lambda^2 V}{2}$ for all $|\lambda| \leq 1/b$.

The following Bernstein condition for random matrices provides one useful way of certifying the sub-exponential condition.

Definition

A zero-mean symmetric random matrix Q satisfies a Bernstein condition with parameter $b > 0$ if

$$\mathbb{E}[Q^j] \preceq \frac{1}{2} j!\, b^{j-2}\ \mathrm{Var}(Q) \quad \text{for } j = 3, 4, \ldots.$$

Similar to the scalar case, the Bernstein condition holds whenever Q has a bounded operator norm, $\|Q\|_{\mathrm{op}} \leq b$. In this case, $\mathbb{E}[Q^j] \preceq b^{j-2}\ \mathrm{Var}(Q)$.

Bernstein condition implies sub-exponential condition

Lemma

For any symmetric zero-mean random matrix satisfies the Bernstein condition, we have

$$\Phi_Q(\lambda) \preceq \frac{\lambda^2 \operatorname{Var}(Q)}{1 - b\,|\lambda|} \quad \textit{for all } |\lambda| \leq \tfrac{1}{b}.$$

Proof is similar to the scalar case.

Matrix-Chernoff approach

Lemma

Let Q be a zero-mean symmetric random matrix whose cumulant function Φ_Q exists in an open interval $(-a, a)$. Then for any $\delta > 0$, we have

$$\mathbb{P}[\gamma_{\max}(Q) \geq \delta] \leq \mathrm{Tr}\left(e^{\Phi_Q(\lambda)}\right) e^{-\lambda\delta} \quad \textit{for all } \lambda \in [0, a).$$

Similarly, we have

$$\mathbb{P}[\|Q\|_{op} \geq \delta] \leq 2\ \mathrm{Tr}\left(e^{\Phi_Q(\lambda)}\right) e^{-\lambda\delta} \quad \textit{for all } \lambda \in [0, a).$$

Proof is similar to the scalar case.

Cumulant function of sum of independent matrices

The cumulant function of sum of independent matrices does not decompose additively, because **matrix products need not commute.**

Fortunately, for independent random matrices, it is possible to establish an upper bound in terms of the trace of the cumulant generating functions.

Lemma

Let $Q_1, \ldots, Q_n$ be independent symmetric random matrices whose cumulant functions exists for all $\lambda \in I$. Then the sum $S_n = \sum_{i=1}^n Q_i$ satisfies

$$\mathrm{Tr}\left(e^{\Phi_{S_n}(\lambda)}\right) \leq \mathrm{Tr}\left(e^{\sum_{i=1}^n \Phi_{Q_i}(\lambda)}\right) \quad \textit{for all } \lambda \in I.$$

A proof uses Lieb's theorem: for any fixed $H \in \mathcal{S}^{d\times d}$, the following function is concave:

$$A \mapsto \mathrm{Tr}\left(e^{H+\log(A)}\right).$$

Tail bounds for sub-Gaussian matrices

Theorem (Hoeffding bound for random matrices)

Let $Q_1, \ldots, Q_n$ be independent symmetric random matrices that are sub-Gaussian with parameters $V_1, \ldots, V_n$. Then for any $\delta > 0$, we have

$$\mathbb{P}\Big[\| \sum_{i=1}^{n} Q_i \|_{op} \geq \delta\Big] \leq 2\, d\, e^{-\frac{n\delta^2}{2\sigma^2}},$$

where $\sigma^2 = \| n^{-1} \sum_{i=1}^{n} V_i \|_{op}$.

This inequality also implies an analogous bound for general independent but potentially non-symmetric and/or non-square matrices in $\mathbb{R}^{d_1 \times d_2}$, with d replaced by $d_1 + d_2$ (why?).

Example: Looseness/Sharpness of leading factor d

- Let $n = d$, and E_i denote the diagonal matrix with 1 in position (i, i) and 0 elsewhere.
- Let $D_i = g_i E_i$ where g_i are i.i.d. sub-Gaussian with parameter 1.
- We showed D_i is sub-Gaussian with $V_i = E_i$, and hence $\sigma^2 = \|d^{-1} I_d\|_{\text{op}} = 1/d$. Therefore,

$$\mathbb{P}\left[\|\frac{1}{d}\sum_{i=1}^{d} Q_i\|_{\text{op}} \geq \delta\right] \leq 2\, d\, e^{-\frac{d\delta^2}{2\sigma^2}},$$

  implying $\|\frac{1}{d}\sum_{i=1}^{d} Q_i\|_{\text{op}} \leq \frac{\sqrt{2\log(2d)}}{d}$ with high probability.
- On the other hand, if g_i are Rademacher variables, then $\|\frac{1}{d}\sum_{i=1}^{d} Q_i\|_{\text{op}} = \frac{1}{d}$ and the concentration inequality is off by the order d; if g_i are standard Gaussians, then $\|\frac{1}{d}\sum_{i=1}^{d} Q_i\|_{\text{op}} \approx \frac{\sqrt{2\log(2d)}}{d}$ and the inequality cannot be improved.

Bernstein-type bounds for random matrices

Theorem (Matrix Bernstein concentration inequality)

Let $Q_1, \ldots, Q_n$ be a sequence of independent, zero-mean, symmetric random matrices that satisfy the Bernstein condition with parameter $b > 0$. Then

$$\mathbb{P}\Big[\| \sum_{i=1}^{n} Q_i \|_{op} \geq \delta \Big] \leq 2\, d\, \exp\Big\{ - \frac{n\delta^2}{2(\sigma^2 + b\delta)} \Big\},$$

where $\sigma^2 = \| n^{-1} \sum_{i=1}^{n} \mathrm{Var}(Q_i) \|_{op}$.

This inequality can also be generalized to non-symmetric matrices $A_i \in \mathbb{R}^{d_1 \times d_2}$, as long as we use

$$\sigma^2 = \max \Big\{ \| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[A_i A_i^T] \|_{\mathrm{op}},\ \| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[A_i^T A_i] \|_{\mathrm{op}} \Big\},$$

and replace d by $d_1 + d_2$.

Example: Tail bounds in matrix completion

- Consider an i.i.d. sequence of matrices of the form $A_i = \xi_i X_i \in \mathbb{R}^{d\times d}$.
- ξ_i is symmetric around zero, satisfying Bernstein condition with parameter b and variance ν^2.
- X_i is independent from ξ_i, with a single entry d in a position chosen uniformly at random from all d^2 entries.
- Define a symmetric version

$$Q_i = \begin{bmatrix} 0_{d\times d} & A_i \\ A_i^T & 0_{d\times d} \end{bmatrix}$$

- $\|\sum_{i=1}^n A_i\|_{\text{op}} = \|\sum_{i=1}^n Q_i\|_{\text{op}}$, Q_i satisfies the Bernstein condition with parameter bd, and $\sigma^2 = \nu^2 d$.
- Then we have

$$\mathbb{P}\Big[\|\sum_{i=1}^n A_i\|_{\text{op}} \geq \delta\Big] \leq 4\,d\,\exp\Big\{-\frac{n\delta^2}{2d(\nu^2+b\delta)}\Big\}.$$

Example: Tail bounds in matrix completion

Now we try to reduce the symmetric assumption on the distribution of ξ_i. We achieve this via the symmetrization technique:

$$\begin{aligned}
\mathbb{E}\Big[\exp\Big\{\lambda\gamma_{\max}(\sum_{i=1}^{n} Q_i)\Big\}\Big] &= \mathbb{E}\Big[\exp\Big\{\lambda \sup_{\|u\|_2=1} u^T\Big(\sum_{i=1}^{n} Q_i\Big)u\Big\}\Big] \\
&\le \mathbb{E}\Big[\exp\Big\{2\lambda \sup_{\|u\|_2=1} u^T\Big(\sum_{i=1}^{n} \varepsilon_i Q_i\Big)u\Big\}\Big] \\
&= \mathbb{E}\Big[\exp\Big\{2\lambda\gamma_{\max}(\sum_{i=1}^{n} \varepsilon_i Q_i)\Big\}\Big],
\end{aligned}$$

where ε_i are i.i.d. Rademacher variables, and the second step follows by the symmetrization theorem with $\Phi(t) = e^{\lambda t}$.

Therefore, we may consider the symmetrized version $\varepsilon_i Q_i$ with the loss of a constant factor.

Applications to covariance matrices

Corollary (Sample Covariance concentration)

Let X_i be i.i.d. zero-mean random vectors with covariance Σ, such that $\|x_i\|_2 \leq \sqrt{b}$ almost surely. Then for all $\delta > 0$,

$$\mathbb{P}\big[\|\widehat{\Sigma} - \Sigma\|_{op} \geq \delta\big] \leq 2d \exp\Big(-\frac{n\delta^2}{2b(\|\Sigma\|_{op} + \delta)}\Big).$$

Proof: Apply matrix Bernstein concentration inequality to $Q_i = x_i x_i^T - \Sigma$.

$$\|Q_i\|_{\mathrm{op}} \leq \|x_i\|_2^2 + \|\Sigma\|_{\mathrm{op}} \leq 2b.$$

Moreover,

$$\mathrm{Var}(Q_i) \leq \mathbb{E}[(x_i x_i^T)^2] \preceq b\Sigma.$$

Example: Random vectors uniform on sphere

x_i are chosen uniformly from the sphere $S^{d-1}(\sqrt{d})$, so that $\|x_i\|_2 = \sqrt{d}$.

By construction, $\mathbb{E}[x_i x_i^T] = \Sigma = I_d$, and $\|\Sigma\|_{\mathsf{op}} = 1$. Therefore,

$$\mathbb{P}\big[\|\widehat{\Sigma} - \Sigma\|_{\mathsf{op}} \geq \delta\big] \leq 2d \exp\Big(-\frac{n\delta^2}{2d(1+\delta)}\Big),$$

which implies the high probability bound

$$\|\widehat{\Sigma} - \Sigma\|_{\mathsf{op}} \lesssim \sqrt{\frac{d \log d}{n}} + \frac{d \log d}{n}.$$

This bound is off by a factor of $\log d$, since we can directly apply the matrix sub-Gaussian concentration inequality (x_i is sub-Gaussian with parameter c for some universal constant $c > 0$).

Example: “Spiked” random vectors

x_i is uniformly chosen from $\{\sqrt{d}e_1, \ldots, \sqrt{d}e_d\}$, where $e_j \in \mathbb{R}^d$ is the canonical basis vector with 1 in position j.

As before, we have $\|x_i\|_2 = \sqrt{d}$, and $\mathbb{E}[x_i x_i^T] = I_d$. Therefore, the same bound applies:

$$\|\widehat{\Sigma} - \Sigma\|_{\mathsf{op}} \lesssim \sqrt{\frac{d \log d}{n}} + \frac{d \log d}{n}.$$

This time, this bound is sharp (up to constant factors).