Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 17

Yun Yang

- Random matrices and covariance estimation

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

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

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

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

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

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

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

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 $\sigma^2 = \|d^{-1} I_d\|_{\mathsf{op}} = 1/d$. Therefore,

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

  implying $\|\frac{1}{d}\sum_{i=1}^{d} Q_i\|_{\mathsf{op}} \leq \frac{\sqrt{2\log(2d)}}{d}$ with high probability.
- On the other hand, if g_i are Rademacher variables, then $\|\frac{1}{d}\sum_{i=1}^{d} Q_i\|_{\mathsf{op}} = \frac{1}{d}$ and the concentration inequality is off by the order d; if g_i are standard Gaussians, then $\|\frac{1}{d}\sum_{i=1}^{d} Q_i\|_{\mathsf{op}} \approx \frac{\sqrt{2\log(2d)}}{d}$ 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:

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

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 $\delta > 0$,

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

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

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

Moreover,

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

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 $\|\Sigma\|_{\text{op}} = 1$. Therefore,

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

which implies the high probability bound

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

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).

Structured covariance estimation: sparsity and thresholding

- Suppose Σ is known to be sparse, but the positions of non-zero entires are unknown.
- Motivates estimators based thresholding.
- Given a tuning parameter $\lambda > 0$, define the *hard thresholding operator* $T_\lambda : \mathbb{R} \to \mathbb{R}$ by

$$T_\lambda(u) = u\,\mathbb{I}[|u| > \lambda].$$

- For a matrix M, we define $T_\lambda(M)$ by applying T_λ to each element.
- We will study the property of the estimator $T_{\lambda_n}(\widehat{\Sigma})$, where $\lambda_n > 0$ is a suitably chosen parameter.

Sparsity and thresholding

- Let $A \in \mathbb{R}^{d \times d}$ denote the adjacency matrix, where $A_{ij} = \mathbb{I}(\Sigma_{ij} \neq 0)$.
- $\|A\|_{\text{op}}$ provides a measure of sparsity: if Σ has at most s non-zero entries per row, then $\|A\|_{\text{op}} \leq s$.

Theorem

x_i are independent zero-mean sub-Gaussian with parameter at most σ^2. If $n \geq \log d$, then for any $\delta > 0$ and $\lambda_n/\sigma^2 = 8\sqrt{\frac{\log d}{n}} + \delta$,

$$\mathbb{P}\left[\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\|_{op} \geq 2\|A\|_{op}\lambda_n\right] \leq 8e^{-\frac{n}{16}\min\{\delta, \delta^2\}}.$$

Corollary

Suppose Σ has at most s non-zero entries per row, then

$$\mathbb{P}\left[\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\|_{op}/\sigma^2 \geq 16s\sqrt{\frac{\log d}{n}} + 2\delta\right] \leq 8e^{-\frac{n}{16}\min\{\delta, \delta^2\}}.$$

Example: Sparsity and adjacency matrices

- In certain cases, the two bounds discussed before coincide.
- Consider any graph with maximum degree $s - 1$ that contains a s-clique
- For any such graph, we have

$$|||A|||_{\text{op}} = s - 1.$$

- In general, the bound with $|||A|||_{\text{op}}$ can be substantially sharper.
- Consider a hub-and-spoke graph, in which one central node known as the hub is connected to s of the remaining $d - 1$ node.
- For this graph, we have

$$|||A|||_{\text{op}} = \sqrt{s}.$$