Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 13

Yun Yang

- Gaussian comparison inequality
- Random matrices and covariance estimation

Gaussian comparison inequality

Suppose that we are given a pair of Gaussian vectors $\{X_j, j = 1, \ldots, N\}$ and $\{Y_j, j = 1, \ldots, N\}$ of the same dimension. Gaussian comparison inequalities compare the two Gaussian vectors in terms of the expected value of some real-valued function F defined on $\mathbb{R}^n$.

Theorem (Sudakov-Fernique)

Given a pair of centered Gaussian vectors $\{X_j, j = 1, \ldots, N\}$ and $\{Y_j, j = 1, \ldots, N\}$, suppose that

$$\mathbb{E}(X_i - X_j)^2 \leq \mathbb{E}(Y_i - Y_j)^2 \quad \textit{for all pair } (i,j) \in N^2.$$

Then $\mathbb{E}[\max_{j=1,\ldots,N} X_j] \leq \mathbb{E}[\max_{j=1,\ldots,N} Y_j]$.

The results can be extended for comparing two Gaussian processes, by taking limits of maxima over finite subsets.

Sudakov's lower bound

The following theorem provides a lower bound on the expected supremum of Gaussian process.

Theorem (Sudakov minoration)

Let X_θ be a zero-mean Gaussian process defined on non-empty set $\mathcal{T}$. Then

$$\mathbb{E}\big[\sup_{\theta\in\mathcal{T}} X_\theta\big] \geq \sup_{\varepsilon>0} \frac{\varepsilon}{2}\sqrt{\log M(\varepsilon,\, \mathcal{T},\, \rho_X)},$$

where $\rho_X(\theta,\, \theta') = \sqrt{\mathit{Var}(X_\theta - X_{\theta'})}$.

Proof: For any $\varepsilon > 0$, let $\{\theta^1, \ldots, \theta^M\}$ be an ε-packing of $\mathcal{T}$. Let $Y_i = X_{\theta^i}$. Define $X_i \overset{iid}{\sim} \mathcal{N}(0, \varepsilon^2/2)$. Then

$$\mathbb{E}[(Y_i - Y_j)^2] \geq \varepsilon^2 = \mathbb{E}[(X_i - X_j)^2].$$

Therefore, $\mathbb{E}\big[\sup_{\theta\in\mathcal{T}} X_\theta\big] \geq \mathbb{E}\big[\max_i Y_i\big] \geq E\big[\max_i X_i\big] \geq \frac{\varepsilon}{2}\sqrt{\log M}$.

Example: Gaussian complexity of ℓ_2-ball

We have proved previously that

$$\mathcal{G}(\mathbb{B}_2^d) \leq \sqrt{d}.$$

Now we apply the Sudakov minoration to capture a $\mathcal{O}(\sqrt{d})$ lower bound. We proved that

$$\log N(\varepsilon, \mathbb{B}_2^d, \|\cdot\|_2) \geq d \log(1/\varepsilon).$$

Therefore, the Sudakov bound implies

$$\mathcal{G}(\mathbb{B}_2^d) \geq \sup_{\varepsilon>0} \left\{ \frac{\varepsilon}{2} \sqrt{d \log(1/\varepsilon)} \right\} \geq \frac{\sqrt{\log 2}}{4} \sqrt{d},$$

by choosing $\varepsilon = 1/2$.

Example: Metric entropy of ℓ_1-ball

Recall that we have the Gaussian complexity upper bound

$$\mathcal{G}(\mathbb{B}_1^d) \le \sqrt{2\log d}.$$

Now we apply Sudakov's minoration to get an upper bound on the metric entropy,

$$\log N(\varepsilon,\, \mathbb{B}_1^d,\, \|\cdot\|_2) \le c\,(1/\varepsilon)^2\, \log d.$$

This bound is tight in ε and d, suggesting that the ℓ_1-ball is much smaller than the ℓ_2 ball when d is large.

Example: Lower bounds on maximum singular value

Recall that for a standard Gaussian random matrix $W \in \mathbb{R}^{n \times d}$, we can write

$$\mathbb{E}[\|W\|_{\mathsf{op}}] = \mathbb{E}\big[\sup_{\Theta \in \mathbb{M}} \langle\langle W,\, \Theta \rangle\rangle\big],$$

where $\mathbb{M} = \big\{\Theta \in \mathbb{R}^{n \times d} :\ \mathrm{Tr}(\Theta) = 1,\ \mathrm{rank}(\Theta) = 1\big\}$.
It can be shown that there exists some universal constant $c > 0$ such that

$$\log N(\varepsilon,\, \mathbb{M},\, \|\cdot\|_{\mathsf{F}}) \geq c\,(n+d)\,\log(1/\varepsilon).$$

This implies

$$\frac{1}{\sqrt{n}}\,\mathbb{E}[\|W\|_{\mathsf{op}}] \geq c'\left(1 + \sqrt{\frac{d}{n}}\right).$$

Covariance estimation: Notation and preliminaries

- Denote the set of all symmetric $d \times d$ matrices by $\mathcal{S}^{d\times d} = \{Q \in \mathbb{R}^{d\times d} : \, Q = Q^T\}$.
- Set of positive semidefinite matrices $\mathcal{S}_+^{d\times d} = \{Q \in \mathcal{S}^{d\times d} : \, Q \succeq 0\}$.
- We use $\gamma(Q)$ to denote the vector of its eigenvalues, ordered as

$$\gamma_{\max}(Q) = \gamma_1(Q) \geq \gamma_2(Q) \geq \cdots \geq \gamma_d(Q) = \gamma_{\min}(Q).$$

- Rayleigh-Ritz variational characterization of the minimum and maximum eigenvalues:

$$\gamma_{\max}(Q) = \max_{v \in \mathcal{S}^{d-1}} v^T Q v, \quad \text{and} \quad \gamma_{\min}(Q) = \min_{v \in \mathcal{S}^{d-1}} v^T Q v,$$

  where $\mathcal{S}^{d-1} = \{v \in \mathbb{R}^d : \, \|v\|_2 = 1\}$.
- For $Q \in \mathcal{S}^{d\times d}$, its ℓ_2-operator norm is

$$\|Q\|_{\text{op}} = \max\{\gamma_{\max}(Q), \, |\gamma_{\min}(Q)|\} = \max_{v \in \mathcal{S}^{d-1}} |v^T Q v|.$$

Covariance estimation: Setup

- Let $\{x_1, \ldots, x_n\}$ be a collection of n independent and identically distributed samples from a distribution in $\mathbb{R}^d$ with zero mean, and covariance matrix $\Sigma = \mathrm{Cov}(x) \in \mathcal{S}^{d \times d}$.
- A standard estimator of Σ is the *sample covariance matrix*

$$\widehat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T = \frac{1}{n} X^T X,$$

  where the ith row of $X \in \mathbb{R}^{n \times d}$ is x_i^T.
- $\widehat{\Sigma}$ is an unbiased estimator of Σ, and our goal is to obtain bounds on the error $\widehat{\Sigma} - \Sigma$ measured in the ℓ_2-operator norm.
- By Weyl's theorem, such a bound implies an error bound on the eigenvalues via

$$\max_{j=1,\ldots,d} |\gamma_j(\widehat{\Sigma}) - \gamma_j(\Sigma)| \leq \|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}.$$

Wishart matrices

- Assume x_i is drawn i.i.d. from a multivariate $\mathcal{N}(0, \Sigma)$ distribution.
- We say X is drawn from a Σ-Gaussian ensemble.
- The sample covariance $\widehat{\Sigma}$ follow a multivariate Wishart distribution.

Theorem (Concentration of Gaussian random matrices)

For each $\delta > 0$, the maximum singular value satisfies

$$\mathbb{P}\Big[\frac{\gamma_{\max}(X)}{\sqrt{n}} \geq \gamma_{\max}(\sqrt{\Sigma})(1+\delta) + \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}\Big] \leq e^{-n\delta^2/2}.$$

Moreover, if $n \geq d$, then the minimum singular value satisfies

$$\mathbb{P}\Big[\frac{\gamma_{\min}(X)}{\sqrt{n}} \leq \gamma_{\min}(\sqrt{\Sigma})(1-\delta) - \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}\Big] \leq e^{-n\delta^2/2}.$$

Example: Operator norm bounds for the standard Gaussian ensemble

Consider a random matrix $W \in \mathbb{R}^{n \times d}$ with i.i.d. $\mathcal{N}(0,1)$ entries.

This corresponds to $\Sigma = I_d$. The theorem implies that when $n \geq d$,

$$\frac{\gamma_{\max}(W)}{\sqrt{n}} \leq 1 + \delta + \sqrt{\frac{d}{n}}, \quad \text{and} \quad \frac{\gamma_{\min}(W)}{\sqrt{n}} \geq 1 - \delta - \sqrt{\frac{d}{n}}$$

holds with probability at least $1 - 2e^{-n\delta^2/2}$.

These bounds implies that

$$\|\frac{1}{n} W^T W - I_d\|_{\mathsf{op}} \leq 2\varepsilon + \varepsilon^2, \quad \varepsilon = \delta + \sqrt{\frac{d}{n}},$$

with the same probability.

Example: Gaussian covariance estimation

We reduce the problem to the standard Gaussian ensemble by writing $X = W\sqrt{\Sigma}$, where $W \in \mathbb{R}^{n\times d}$ has i.i.d. $\mathcal{N}(0,1)$ entries.

$$\begin{aligned}
\|\frac{1}{n}X^T X - \Sigma\|_{\text{op}} &= \|\Sigma^{1/2}(\frac{1}{n}W^T W - I_d)\Sigma^{1/2}\|_{\text{op}} \\
&\le \|\Sigma\|_{\text{op}}\,\|\frac{1}{n}W^T W - I_d\|_{\text{op}}.
\end{aligned}$$

Consequently,

$$\frac{\|\widehat{\Sigma} - \Sigma\|_{\text{op}}}{\|\Sigma\|_{\text{op}}} \le 2\delta + 2\sqrt{\frac{d}{n}} + \Big(\delta + \sqrt{\frac{d}{n}}\Big)^2$$

holds with probability at least $1 - 2e^{-n\delta^2/2}$.

Proof: Concentration of Gaussian random matrices

We only prove the upper bound. The proof consists of two steps. Recall $X =$, where $W \in \mathbb{R}^{n\times d}$ has i.i.d. $\mathcal{N}(0,1)$ entries.

Step one: we use concentration inequalities to argue that the random singular value is close to its expectation with high probability.

Consider the mapping $W \mapsto \gamma_{\max}(W\sqrt{\Sigma})/\sqrt{n}$. It is Lipschitz w.r.t. the Euclidean norm with parameter at most $L = \gamma_{\max}(\sqrt{\Sigma})/\sqrt{n}$. Therefore,

$$\mathbb{P}\left[\gamma_{\max}(X) \geq \mathbb{E}[\gamma_{\max}(X)] + \sqrt{n}\,\gamma_{\max}(\sqrt{\Sigma})\,\delta\right] \leq e^{-n\delta^2}.$$

Proof: Concentration of Gaussian random matrices

Step two: we use Gaussian comparison inequalities to bound the expected value

$$\mathbb{E}[\gamma_{\max}(X)] \leq \sqrt{n}\,\gamma_{\max}(\sqrt{\Sigma}) + \sqrt{\operatorname{Tr}(\Sigma)}.$$

We use the variational characterization

$$\gamma_{\max}(X) = \max_{u \in \mathcal{S}^{n-1}} \max_{v \in \mathcal{S}^{d-1}(\Sigma^{-1})} \underbrace{u^T W v}_{Z_{u,v}},$$

where $\mathcal{S}^{d-1}(\Sigma^{-1}) = \{v \in \mathbb{R}^d : \|\Sigma^{-1/2} v\|_2 = 1\}$ is an ellipsoid. $\gamma_{\max}(X)$ is the supremum of the zero-mean GP $Z_{u,v}$.

It can be verified that

$$\mathbb{E}[(Z_{u,v} - Z_{\tilde{u},\tilde{v}})^2] = \|uv^T - \tilde{u}\tilde{v}^T\|_{\mathsf{F}}^2 \leq \gamma_{\max}^2(\sqrt{\Sigma})\,\|u-\tilde{u}\|_2^2 + \|v-\tilde{v}\|_2^2.$$

Proof: Concentration of Gaussian random matrices

Define another GP $Y_{u,v}$ by

$$Y_{u,v} = \gamma_{\max}(\sqrt{\Sigma})\langle g,\, u\rangle + \langle h,\, v\rangle,$$

where $g \sim \mathcal{N}(0,\, I_n)$ and $h \sim \mathcal{N}(0,\, I_d)$. Then

$$\mathbb{E}[(Z_{u,v} - Z_{\tilde{u},\tilde{v}})^2] \leq \mathbb{E}[(Y_{u,v} - Y_{\tilde{u},\tilde{v}})^2].$$

We may apply the Sudakov-Fernique bound to obtain

$$\begin{aligned}
\mathbb{E}[\gamma_{\max}(X)] &\leq \mathbb{E}[\max_{u \in \mathcal{S}^{n-1}} \max_{v \in \mathcal{S}^{d-1}(\Sigma^{-1})} Y_{u,v}] \\
&= \gamma_{\max}(\sqrt{\Sigma})\, \mathbb{E}[\|g\|_2] + \mathbb{E}[\|\sqrt{\Sigma}\, h\|_2] \\
&\leq \sqrt{n}\, \gamma_{\max}(\sqrt{\Sigma}) + \sqrt{\mathrm{Tr}(\Sigma)}.
\end{aligned}$$