Statistics 210B Lecture 16 Notes

Daniel Raban

March 15, 2022

1 Concentration of Sample Covariance of Sub-Gaussian and Bounded Random Vectors

1.1 Concentration of sample covariance of sub-Gaussian vectors

Last time, we were talking about concentration of sub-Gaussian sample covariance. If we have $X_i \overset{\text{iid}}{\sim} \mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$ and covariance matrix $\mathbb{E}[X_i X_i^\top] = \Sigma \in S_+^{d\times d}$. Then we can estimate Σ by the sample covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top = \frac{1}{n} X^\top X \in S_+^{d\times d}$.

Definition 1.1. We say a mean 0 random variable $x \in \mathbb{R}^d$ is **sub-Gaussian**(σ) if

$$\mathbb{E}[e^{\lambda\langle v, x\rangle}] \le e^{\lambda^2 \|v\|_2^2 \sigma^2/2} \qquad \forall \lambda \in \mathbb{R}, v \in \mathbb{R}^d.$$

A sufficient condition for $X \in \mathbb{R}^d$ to be $\text{sG}(\sigma)$ is that X_i are independent with $X_i \sim \text{sG}(\sigma)$.

Theorem 1.1. *Let $(X_i)_{i\in[n]}$ be independent, mean 0 $\text{sG}(\sigma)$. Then wth probability at least $1-\delta$, we have*

$$\|\widehat{\Sigma} - \Sigma\|_{\text{op}} \le C\sigma^2 \left(\sqrt{\frac{d + \log(1/\delta)}{n}} + \frac{d + \log(1/\delta)}{n}\right).$$

Proof. Here is the high level intuition of the proof:

We can represent

$$\begin{aligned}\|\widehat{\Sigma} - \Sigma\|_{\text{op}} &= \sup_{v \in S^{d-1}} |\langle v, (\widehat{\Sigma} - \Sigma)v\rangle| \\ &= \sup_{v\in S^{d-1}} \left|\frac{1}{n}\sum_{i=1}^n (\langle X_i, v\rangle^2 - \mathbb{E}[\langle X_i, v\rangle^2])\right|.\end{aligned}$$

(a) Fix v. Then $\frac{1}{n}\sum_{i=1}^n (\langle X_i, v\rangle^2 - \mathbb{E}[\langle X_i, v\rangle^2])$ has a sub-exponential tail bond.

(b) If we let $|\Omega_\varepsilon = N|$ be the size of an ε-cover of the sphere, then we get the metric entropy bound (instead of a union bound over all the points on the sphere)

$$\sup_{v \in \Omega_\varepsilon} \left| \frac{1}{n} \sum_{i=1}^{n} (\langle X_i, v \rangle^2 - \mathbb{E}[\langle X_i, v \rangle^2]) \right| \lesssim \sqrt{\frac{\log(N_\varepsilon/\delta)}{n}} + \frac{\log(N_\varepsilon/\delta)}{n}.$$

(c) Show that $N_\varepsilon \asymp d$.

(d) Last, show that the discretization error is multiplicative.

Now for the actual proof:

Let $\Omega_\varepsilon = \{v^1, \ldots, v^{N_\varepsilon}$ be an ε-covering of S^{d-1} in the $\|\cdot\|_2$ norm. Then $|\Omega_\varepsilon| \leq (1+2/\varepsilon)^d$. We claim that for every matrix $A \in \mathbb{R}^{d \times d}$,

$$\|A\|_{\text{op}} \leq \frac{1}{1 - 2\varepsilon - \varepsilon^2} \sup_{v \in \Omega_\varepsilon} |\langle , Av \rangle|.$$

This claim holds because

$$\|A\|_{\text{op}} = \sup_{v \in S^{d-1}} v, Av\rangle.$$

Then for all $v \in S^{d-1}$, there is a $v^j \in \Omega_\varepsilon$ such that $\|v - w\|_2 \leq \varepsilon$. We can then compare

$$\langle v, Av \rangle = \langle w, Aw \rangle + 2\langle v - w, Aw \rangle + \langle v - w, A(v - w) \rangle.$$

Using this algebra, we get the bound

$$\sup_{v \in S^{d-1}} |\langle v, Av \rangle| \leq \sup_{w \in \Omega_\varepsilon} |\langle w, Aw \rangle| + (2\varepsilon + \varepsilon^2)\|A\|_{\text{op}}.$$

Rearranging this gives the claim:

$$\|A\|_{\text{op}} \leq \frac{1}{1 - 2\varepsilon - \varepsilon^2} \sup_{v \in \Omega_\varepsilon} |\langle v, Av \rangle|.$$

Take $\varepsilon = 1/8$, so we have a covering with $|\Omega_\varepsilon| \leq 17^d$. Then

$$\|\widehat{\Sigma} - \Sigma\|_{\text{op}} \leq 2 \sup_{v \in \Omega_{1/8}} |\langle v, (\widehat{\Sigma} - \Sigma)v \rangle|.$$

Now look at the tail bound of $\|\langle v, (\widehat{\Sigma} - \Sigma)v \rangle|$ for fixed v. Then

$$|\langle v, (\widehat{\Sigma} - \Sigma)v \rangle| = \left| \frac{1}{n} \sum_{i=1}^{n} (\langle v, X_i \rangle^2 - \mathbb{E}[\langle v, X_i \rangle^2] \right|.$$

By assumption, $\langle v, X_i\rangle/\sigma$ is sG(1), so $((\langle v, X_i\rangle^2 - \mathbb{E}[\langle v, X_i\rangle^2])/\sigma^2$ is sE$(1,1)$. Therefore, $(\frac{1}{n}\sum_{i=1}^n(\langle v, X_i\rangle^2 - \mathbb{E}[\langle v, X_i\rangle^2])/\sigma^2$ is sE$(1/\sqrt{n}, 1/n)$.

Thus, we get the sub-exponential tail bound

$$\mathbb{P}(|\langle v, (\widehat{\Sigma} - \Sigma)v\rangle| \geq \sigma^2 t) \leq 2\exp(-n\min(t^2, t).$$

Using a union bound, we get

$$\mathbb{P}(|\langle v, (\widehat{\Sigma} - \Sigma)v\rangle| \geq \sigma^2 t) \leq 2\exp(-n\min(t^2, t) + d\log 17).$$

Now pick $t = C\max\{\sqrt{\frac{d+\log(1/\delta)}{n}}, \frac{d+\log(1/\delta)}{n}\}$, so we get

$$\mathbb{P}\left(\sup_{v\in\Omega_{1/8}} |\langle v, (\widehat{\Sigma} - \Sigma)v\rangle| \leq C\sigma^2 \max\left\{\sqrt{\frac{d+\log(1/\delta)}{n}}, \frac{d+\log(1/\delta)}{n}\right\}\right) \geq 1-\delta.$$

That is, with high probability,

$$\|\widehat{\Sigma} - \Sigma\|_{\text{op}} \leq C\sigma^2 \max\left\{\sqrt{\frac{d+\log(1/\delta)}{n}}, \frac{d+\log(1/\delta)}{n}\right\}. \qquad \square$$

1.2 Concentration of sample covariance of bounded random vectors

Theorem 1.2. *Let $X_i \overset{\text{iid}}{\sim} X \in \mathbb{R}^d$, and let the covariance matrix $\mathbb{E}[XX^\top] = \Sigma$. Suppose that $\|X\|_2^2 \leq b$ almost surely. Then with probability $1-\delta$,*

$$\|\widehat{\Sigma} - \Sigma\|_{\text{op}} \lesssim \sqrt{\frac{b\|\Sigma\|_2 \log(d/\delta)}{n}} + \frac{b}{n}\log(d/\delta).$$

Example 1.1. Let $X \sim \text{Unif}(S^{d-1}(\sqrt{d}))$, and let $\Sigma = \mathbb{E}[XX^\top] = \text{Id}$. Then we have $b = d$, so the theorem gives

$$\|\widehat{\Sigma} - \Sigma\|_2 \lesssim \sqrt{\frac{d\log d}{n}} + \frac{d}{n}\log d.$$

The proof of this theorem follows from a matrix Bernstein inequality, which we will now prove.

1.3 Matrix Hoeffding/Bernstein inequality

In general, let $X_1, X_2, \ldots, X_n \in \mathbb{R}$ be independent sG(σ) random variables. Then the scalar Hoeffding inequality says

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2\sigma^2}\right).$$

The matrix Hoeffding inequality says

Theorem 1.3. *Let $Q_1, Q_2, \ldots, Q_n \in S^{d\times d}$ be independent* $\mathrm{sG}(V)$*, where $V \in S_+^{d\times d}$. Then*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(Q_i - \mathbb{E}[Q_i])\right\|_{\mathrm{op}} \geq t\right) \leq 2d\exp\left(-\frac{nt^2}{2\|V\|_{\mathrm{op}}^2}\right).$$

We get an extra factor of d in the bound. Notice that when $d = 1$, notice that this reduces to the scalar Hoeffding inequality. Let's review the proof of the scalar Hoeffding inequality:

Use the scalar Chernoff inequality to get

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq t\right) \leq \inf_{\lambda\geq 0}\frac{\mathbb{E}[e^{\lambda\sum_{i=1}^{n}(X_i-\mathbb{E}[X_i])}}{e^{\lambda tn}}$$

Using the scalar tensorization of the MGF,

$$= \inf_{\lambda\geq 0}\frac{\prod_{i=1}^{n}\mathbb{E}[e^{\lambda(X_i-\mathbb{E}[X_i])}]}{e^{\lambda t^n}}$$

Now we use a scalar MGF bound from the sub-Gaussian definition.

$$\leq \inf_{\lambda\geq 0}\prod_{i=1}^{n} e^{\lambda^2\sigma^2/2}e^{-\lambda t^n}$$

$$= e^{-\frac{nt^2}{2\sigma^2}}.$$

The proof of the scalar Bernstein inequality is similar.

1.3.1 Matrix Chernoff inequality

Here is a Matrix Chernofff inequality:

Lemma 1.1. *Let $Q \in S^{d\times d}$ be a symmetric matrix. Then*

$$\mathbb{P}(\lambda_{\max}(Q) \geq t) \leq \inf_{\lambda\geq 0}\frac{\mathbb{E}[\mathrm{tr}(e^{\lambda Q})]}{e^{\lambda t}}.$$

Let $Q \in S^{d\times d}$ be a symmetric matrix with eigendecomposition $Q = U\Lambda U^\top$. If we let $f : \mathbb{R} \to \mathbb{R}$, we define $f(Q) := U\,\mathrm{diag}(f(\lambda_1, \ldots, f(\lambda_d))U^\top \in S^{d\times d}$, so $e^Q = U\,\mathrm{diag}(e^{\lambda_1}, \ldots, e^{\lambda_d})U^\top$. If f is an analytic function with Taylor expansion $f(x) = \sum_{i=1}^{\infty}\frac{f^{(i)}(0)}{i!}x^i$, then

$$f(Q) = \sum_{i=1}^{\infty}\frac{f^{(i)}(0)}{i!}Q^i.$$

In particular,

$$e^Q = \sum_{i=0}^{\infty}\frac{1}{i!}Q^i.$$

Proof. For $\lambda \geq 0$,

$$\begin{aligned}\mathbb{P}(\lambda_{\max}(Q) \geq t) &= \mathbb{P}(e^{\lambda\lambda_{\max}(Q)} \geq e^{\lambda t}) \\ &= \mathbb{P}(\lambda_{\max}(e^{\lambda Q}) \geq e^{\lambda t})\end{aligned}$$

Use Markov's inequality.

$$\leq \frac{\mathbb{E}[\lambda_{\max}(e^{\lambda Q})]}{e^{\lambda t}}$$

The largest eigenvalue of a positive definite matrix is upper bounded by its trace.

$$\leq \frac{\mathbb{E}[\operatorname{tr}(e^{\lambda Q})]}{e^{\lambda t}}.$$

□

1.3.2 Sub-Gaussian and sub-exponential matrices

Definition 1.2. A matrix $Q \in S^{d\times d}$ with $\mathbb{E}[Q] = 0$ is **sub-Gaussian**(V) for $V \in S_+^{d\times d}$ if

$$\Phi_Q(\lambda) = \mathbb{E}[e^{\lambda Q}] \preceq e^{\lambda^2 V/2} \qquad \forall \lambda \in \mathbb{R}.$$

This is not equivalent to the definition we have given for vectors.

Example 1.2. Let $Q = \varepsilon B$, where $B \in S^{d\times d}$ and $\varepsilon \sim \text{Unif}(\{\pm 1\})$. Then

$$\begin{aligned}\mathbb{E}[e^{\lambda Q}] &= \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}[Q^k] \\ &= \sum_{i=0}^{\infty} \frac{\lambda^{2i}}{(2i)!} \mathbb{E}[Q^{2i}] \\ &= \sum_{i=1}^{\infty} \frac{\lambda^{2i}}{(2i)!} B^{2i} \\ &\prec \sum_{i=1}^{\infty} \frac{1}{i!} \left(\frac{\lambda^2 B^2}{2}\right)^i \\ &= e^{\lambda^2 B^2/2}.\end{aligned}$$

Similarly, we can define sub-exponential matrices.

Definition 1.3. A matrix $Q \in S^{d\times d}$ with $\mathbb{E}[Q] = 0$ is **sub-exponential**(V, α) for $V \in S_+^{d\times d}$ and $\alpha \in \mathbb{R}_{\geq 0}$ if

$$\Phi_Q(\lambda) = \mathbb{E}[e^{\lambda Q}] \preceq e^{\lambda^2 V/2} \qquad \forall |\lambda| \leq \frac{1}{\alpha}.$$

Here is a sufficient condition: Define $\operatorname{Var}(Q) = \mathbb{E}[Q^2] - (\mathbb{E}[Q])^2 \in S_+^{d\times d}$. If $\mathbb{E}[Q] = 0$ and $\|Q\|_{\text{op}} \leq b$ a.s., then $Q \sim \text{sE}(\operatorname{Var}(Q), b)$. This is proved in Wainwright's textbook.

Example 1.3. Let $\|X_i\|_2 \le \sqrt{b}$, so $\mathbb{E}[X_i X_i^\top] = \Sigma$. Then if we let $Q = X_i X_i^\top - \Sigma$, then $\|Q\|_{\text{op}} \le b$. This gives

$$\text{Var}(Q) = \mathbb{E}[(X_i X_i^\top - \Sigma)^2] \preceq b\Sigma,$$

so $Q \sim \text{sE}(b\Sigma, b)$.

1.3.3 Tensorization of the matrix MGF

Now we know how to give an upper bound of the matrix MGF. The last step is to see the tensorization of the matrix MGF. The scalar MGF tensorizes as

$$\mathbb{E}[e^{\lambda \sum_{i=1}^n X_i}] = \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}].$$

This is not true for matrices:

$$\mathbb{E}[e^{\lambda \sum_{i=1}^n Q_i}] \neq \prod_{i=1}^n \mathbb{E}[e^{\lambda Q_i}],$$

since $e^{A+B} \neq e^A e^B$. However, this lemma solves the problem:

Lemma 1.2. *Let $Q_1, \dots, Q_n$ be independent. Then*

$$\text{tr}(\mathbb{E}[e^{\lambda \sum_{i=1}^n Q_i}]) \le \text{tr}(e^{\sum_{i=1}^n \log \mathbb{E}[e^{\lambda Q_i}]}).$$

To prove this, we use the following general matrix inequality:

Lemma 1.3 (Lieb's inequality, 1973)**.** *Let $H \in S^{d\times d}$. Then the function $f : S_+^{d\times d} \to \mathbb{R}$ sending $A \mapsto \text{tr}(e^{H+\log A})$ is concave.*

This inequality was originally proven for the use of quantum information theory. Using Lieb's inequality, the lemma is just the repeated application of this concavity and Jensen's inequality. Now we we can prove the matrix Hoeffding inequality:

Proof. Let Q_i be independent $\text{sG}(V_i)$ random matrices with $\mathbb{E}[Q_i] = 0$. Use the matrix Chernoff inequality to get

$$\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n Q_i\right) \ge t\right) \le \inf_{\lambda \ge 0} \mathbb{E}[\text{tr}(e^{\lambda \sum_{i=1}^n Q_i})] e^{-\lambda n t}$$

Using the Matrix tensorization of the MGF,

$$\le \inf_{\lambda \ge 0} \text{tr}(e^{\sum_{i=1}^n \log \mathbb{E}[e^{\lambda Q_i}]}) e^{-\lambda n t}$$

Now apply the matrix sub-Gaussian upper bound and the inequality $\log A \preceq \log B$ if $A \prec B$ (which is not true in general for every monotone function) to get

$$\begin{aligned}
&\inf_{\lambda \geq 0} \operatorname{tr}(e^{\sum_{i=1}^n (\lambda^2/2) V_i}) e^{-\lambda n t} \\
&\leq d \inf_{\lambda \geq 0} e^{(\lambda^2/2) n \|V\|_{\mathrm{op}}} e^{-\lambda n t} \\
&= d e^{-\frac{nt^2}{2\|V\|_{\mathrm{op}}}}.
\end{aligned}$$

□

This gives the matrix Hoeffding and matrix Bernstein inequalities:

Theorem 1.4 (Matrix Hoeffding inequality)**.** *Let* $Q_i \overset{\text{ind}}{\sim} \mathrm{sG}(V_i)$ *with* $\mathbb{E}[Q_i] = 0$*. Then*

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^n Q_i \right\|_{\mathrm{op}} \geq t \right) \leq 2d \exp\left(-\frac{nt^2}{2\sigma^2} \right),$$

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

Theorem 1.5 (Matrix Bernstein inequality)**.** *Let* $Q_i \overset{\text{ind}}{\sim} \mathrm{sE}(V_i, \alpha_i)$ *with* $\mathbb{E}[Q_i] = 0$*. Then*

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^n Q_i \right\|_{\mathrm{op}} \geq t \right) \leq 2d \exp\left(-n \min\left\{ \frac{t^2}{2\sigma^2}, \frac{t}{2\alpha_*} \right\} \right),$$

where $\sigma^2 = \|\frac{1}{n} \sum_{i=1}^n V_i\|_{\mathrm{op}}$ *and* $\alpha_* = \max_{i \in [n]} \alpha_i$*.*

Remark 1.1. These are symmetric versions of these inequalities. We can prove non-symmetric versions by taking $A \in \mathbb{R}^{n \times d}$ and considering

$$Q = \begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix} \in \mathbb{R}^{(n+d) \times (n+d)}.$$

The singular values of A are related to the eigenvalues of Q.

Going back to the sample covariance, we have $\|X_i\|_2^2 \leq b$ and $\mathbb{E}[X_i X_i^\top] = \Sigma$. Then $\widehat{\Sigma} - \Sigma \sim \mathrm{sE}(b\Sigma, b)$, which gives us the matrix Bernstein bound

$$\mathbb{P}(\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} \geq t) \leq 2d \exp\left(-n \min\left\{ \frac{t^2}{2b\|\Sigma\|_{\mathrm{op}}}, \frac{t}{2b} \right\} \right).$$

So with high probability,

$$\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} \lesssim \sqrt{\frac{b\|\Sigma\|_{\mathrm{op}} \log(d/\delta)}{n}} + \frac{b}{n} \log(d/\delta).$$