In this chapter, we focus on the upper bound of the supremum of Sub-Gaussian process on some metric space T, i.e.,

$$\mathbb{E}\sup_{\theta\in T}|X_\theta| \quad \text{or} \quad \mathbb{E}\sup_{\theta\in T} X_\theta.$$

Firstly, we motivate the Chapter by two important examples

Example 1 Gaussian / Rademacher complexity

$$\mathcal{G}(T) = \mathbb{E}\sup_{\theta\in T}\langle\theta,\varepsilon\rangle,\ \varepsilon\sim N(0,I_n)$$

$$\mathcal{R}(T) = \mathbb{E}\sup_{\theta\in T}\langle\theta,\varepsilon\rangle,\ \varepsilon\sim \mathrm{Unif}\{-1,1\}^n,$$

where $\langle\theta,\varepsilon\rangle$ can be viewed as a stochastic process on $T\subset\mathbb{R}^n$.

Now we give an example of complexity: Maximum singular value of random matrix.

Let $W\in M^{n\times d}(\mathbb{R})$ be a random matrix with zero-mean iid entries W_{ij}, the operator norm on $M^{n\times d}(\mathbb{R})$:

$$\|W\| = \sup_{v\in S^{d-1}}\sqrt{v^T W^T W v}$$

$$= \sup_{v\in S^{d-1},\,u\in S^{n-1}} u^T W v$$

$$= \sup_{v,u} \sum_{i,j} u_i W_{ij} v_j$$

$$= \sup_{v,u} \langle\langle W, uv^T \rangle\rangle,$$

where $\langle\langle A, B \rangle\rangle = \sum_{i,j} A_{ij} B_{ij}$, note that

$$\{uv^T : v \in S^{d-1}, u \in S^{n-1}\}$$
$$= \{A \in M^{n\times d}(\mathbb{R}) : \operatorname{rank}(A) = 1, \sqrt{\langle\langle A, A \rangle\rangle} := \|A\| = 1\}$$
$$=: M^{n\times d}_{1,1}(\mathbb{R})$$

$$\Rightarrow \|W\| = \sup_{A \in M^{n\times d}_{1,1}(\mathbb{R})} \langle\langle A, W \rangle\rangle,$$

where $\langle\langle A, W \rangle\rangle$ can be viewed as the stochastic process on $M^{n\times d}_{1,1}(\mathbb{R})$.

Example 2 Learning process

For regression problem, we assume that $X_1, \dots, X_n \overset{iid}{\sim} (\mathcal{X}, \Sigma, P)$, and

$$Y_i = T(X_i) + \varepsilon_i, \text{ where } X_i \perp \varepsilon_i$$

and $\varepsilon_i \perp \varepsilon_j$, $i \neq j$, $\mathbb{E}\varepsilon_i = 0$.

Given a class of functions on $\mathcal{X}$: $\mathcal{F}$, since T may not belong to $\mathcal{F}$, then

$$f^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}(f(X) - T(X))^2$$

is the optimal function we can learn. With help from predictor Y, note that

$$\mathbb{E}(f(X)-Y)^2 = \mathbb{E}(f(X)-T(X)-\varepsilon)^2$$
$$= \mathbb{E}(f(X)-T(X))^2 + \mathbb{E}\varepsilon^2,$$

if $\mathbb{E}\varepsilon^2 < \infty$, then

$$f^* = \operatorname{argmin}_{f\in\mathcal{F}} \mathbb{E}(f(X)-Y)^2,$$

which implies that we can use M-estimator of f^*:

$$\hat{f}_n = \operatorname{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(f(X_i)-Y_i)^2,$$

for approximation of f^*.

To ensure $\hat{f}_n$ is a good estimator for f^*, we need to evaluate the excess risk:

$$R(\hat{f}_n) - R(f^*),$$

where $R(\hat{f}_n) = \mathbb{E}(\hat{f}_n(X)-T(X))^2$, $R(f^*) = \mathbb{E}(f^*(X)-T(X))^2$.

Let $R_n(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(X_i)-T(X_i))^2$, note that

$$R(\hat{f}_n) - R(f^*) \le R(\hat{f}_n) - R_n(\hat{f}_n)$$
$$+ \underbrace{R_n(\hat{f}_n) - R_n(f^*)}_{\le 0} + R_n(f^*) - R(f^*)$$

$\leq 2 \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|$.

Now, Let $\tilde{\mathcal{F}} = \{(f-T)^2 ; f \in \mathcal{F}\}$, define:

$$Pf =: R(f) = \mathbb{E} f(X),$$
$$P_n f =: R_n(f) = \frac{1}{n} \sum_{i=1}^n f(X_i),$$
$$P_n^\varepsilon f = \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i), \quad \varepsilon_i \sim \mathrm{Unif}\{-1, 1\} \perp X_i.$$

Then

Lemma 1

$$\mathbb{E} \|P_n - P\|_{\tilde{\mathcal{F}}} \leq 2 \mathbb{E} \|P_n^\varepsilon\|_{\tilde{\mathcal{F}}},$$ where $\|Q - P\|_{\tilde{\mathcal{F}}} = \sup_{f \in \tilde{\mathcal{F}}} |Qf - Pf|$, Q, P are two laws on $\mathcal{X}$.

$$Pf: \quad |P_n f - Pf| = |P_n f - \mathbb{E} P_n' f| X|$$

(where $P_n' f = \frac{1}{n} \sum_{i=1}^n f(X_i')$, $X_i' \perp X_i$)

$$\overset{\text{Jesen}}{\leq} \mathbb{E} |P_n f - P_n' f| | X$$

$$\Rightarrow \mathbb{E} \|P_n - P\|_{\tilde{\mathcal{F}}} \leq \mathbb{E} \sup_{f \in \tilde{\mathcal{F}}} \mathbb{E} |P_n f - P_n' f| | X$$

$$\leq \mathbb{E} \|P_n - P_n'\|_{\tilde{\mathcal{F}}}, \text{ note that}$$

$$(\varepsilon_i (f(X_i) - f(X_i')) =_d f(X_i) - f(X_i'))$$

$$= \mathbb{E} \|P_n^\varepsilon - P_n'^{,\varepsilon}\|_{\tilde{\mathcal{F}}}$$
$$\leq 2 \mathbb{E} \|P_n^\varepsilon\|_{\tilde{\mathcal{F}}}.$$

□

Maximum of Gaussian sequence

If $X \sim N(0, \sigma^2)$. the moment generating function of X:

$$\mathbb{E} e^{\lambda X} = \int_{\mathbb{R}} e^{\lambda x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx$$

$$= e^{\frac{\lambda^2\sigma^2}{2}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\lambda\sigma^2)^2}{2\sigma^2}} dx$$

$$= e^{\lambda^2\sigma^2/2}.$$

Lemma 2 If $X_i \sim N(0, \sigma^2)$, then

$$\mathbb{E} \max_{i \le n} X_i \le \sqrt{2\sigma^2 \log n}.$$

Pf: $e^{\lambda \mathbb{E} \max_i X_i} \overset{\text{Jesen}}{\le} \mathbb{E} e^{\lambda \max_i X_i}$

$$\le n \, \mathbb{E} e^{\lambda X_i}, \forall \lambda > 0$$

$$\Rightarrow \mathbb{E} \max_i X_i \le \log n / \lambda + \frac{\lambda\sigma^2}{2}, \forall \lambda > 0$$

$$\Rightarrow \mathbb{E} \max_i X_i \le \sqrt{2\sigma^2 \log n}. \quad \square$$

Sub-Gaussian variable

X is called a sub-Gaussian iff $\exists\, Y \sim N(0, \tau^2)$ s.t.

$$\mathbb{P}(|X| \geq t) \leq \mathbb{P}(|Y| \geq t).$$

Remark: ① X sub-Gaussian $\Leftrightarrow$ $-X$ sub-Gaussian

② $$\mathbb{P}(|Y| \geq t) = \sqrt{\frac{2}{\tau^2 \pi}} \int_t^{+\infty} e^{-\frac{x^2}{2\tau^2}} dx$$

$$= o\left(e^{-\frac{t^2}{2}}\right)$$

$\Rightarrow \exists\, C > 0$ s.t. $\mathbb{P}(|X| \geq t) \leq C e^{-\frac{t^2}{2}}$

$$=: 2e^{-\frac{t^2}{2\sigma^2}}$$

the minimal of σ is determined by X, we mark that

$$\|X\|_\psi = \sigma, \quad \text{(Why?)}$$

indeed, $\|\cdot\|_\psi$ is a norm of X, and

X sub-Gaussian $\Leftrightarrow \|X\|_\psi < \infty$ (Why?)

③ If X sub-Gaussian, then $\exists\, \sigma \geq 0$ s.t.

$$\mathbb{E}e^{\lambda X} \leq e^{\lambda^2 \sigma^2/2} \quad \text{and} \quad \sigma \geq \|X\|_\psi.$$

Moreover,

$$\mathbb{E}e^{\lambda X} \leq e^{\lambda^2 \sigma^2/2} \Leftrightarrow \mathbb{P}(|X| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

(why?)

Lemma 3 X_i sub-Gaussian, s.t.

$$\mathbb{E} e^{\lambda X_i} \le e^{\lambda^2 \sigma^2 / 2}, \quad \forall i \le n.$$

then

$$\mathbb{E} \max_{i \le n} |X_i| \lesssim \sqrt{\sigma^2 \log n}$$

Pf:

$$e^{\lambda \mathbb{E} \max_{i \le n} |X_i|}$$

$$\overset{\text{Jesen}}{\le} \mathbb{E} e^{\lambda \max_{i \le n} |X_i|}$$

$$= \mathbb{E} \max_i e^{\lambda |X_i|}$$

$$\le n \, \mathbb{E} e^{\lambda |X_i|}$$

note that

$$\mathbb{E} e^{\lambda |X_i|} \le \mathbb{E}(e^{-\lambda X_i} + e^{\lambda X_i})$$

$$\le 2 e^{\lambda^2 \sigma^2 / 2}$$

$$\Rightarrow e^{\lambda \mathbb{E} \max_i |X_i|} \le 2n e^{\lambda^2 \sigma^2 / 2}$$

$$\Rightarrow \mathbb{E}\max_i |X_i| \le \frac{\log 2n}{\lambda} + \frac{\lambda \sigma^2}{2}, \quad \forall \lambda > 0$$

$$\Rightarrow \mathbb{E}\max_i |X_i| \le \sqrt{4 \log 2n \cdot \frac{\sigma^2}{2}}$$

$$= \sqrt{2 \log(2n)\, \sigma^2}$$

$$\lesssim \sqrt{\sigma^2 \log n} \qquad \square$$

Sub-Gaussian process

$\{X_\theta, \theta \in T\}$ is called sub-Gaussian process iff

$$\mathbb{E}\, e^{\lambda (X_{\theta_1} - X_{\theta_2})} \le e^{\frac{\lambda^2 (d(\theta_1, \theta_2))^2}{2}}$$

Remark: $\{X_\theta, \theta \in T\}$ sub-Gaussian

$$\Leftrightarrow \|X_{\theta_1} - X_{\theta_2}\|_\psi \lesssim d(\theta_1, \theta_2).$$

Example 1 Gaussian / Rademacher complexity

Let $X_\theta = \langle \theta, \varepsilon \rangle$, $\varepsilon \sim N(0, I_n)$, then

$$X_{\theta_1} - X_{\theta_2} = \langle \theta_1 - \theta_2, \varepsilon \rangle \sim N(0, \|\theta_1 - \theta_2\|^2)$$

$$\Rightarrow \mathbb{E} e^{\lambda (X_{\theta_1} - X_{\theta_2})} = e^{\frac{\lambda^2 \|\theta_1 - \theta_2\|^2}{2}}$$

If $\varepsilon_i \sim \mathrm{Unif}\{-1, 1\}^n$,

$$\|X_{\theta_1} - X_{\theta_2}\|_\psi$$

$$= \left\| \sum_{i=1}^n (\theta_{1(i)} - \theta_{2(i)}) \varepsilon_i \right\|_\psi$$

$$\leq \sum_{i=1}^n |\theta_{1(i)} - \theta_{2(i)}| \, \|\varepsilon_i\|_\psi$$

$$\lesssim \|\theta_1 - \theta_2\|_\infty \quad \text{(Why?)}$$

If $\|\varepsilon_i\|_\psi < 6$, then

$$\|X_{\theta_1} - X_{\theta_2}\| \lesssim \|\theta_1 - \theta_2\|_\infty \leq \|\theta_1 - \theta_2\|,$$

$\|\cdot\|$ is an arbitrary norm of $\mathbb{R}^n$.

Example 2 Learning process

Recall $P_n^{\varepsilon} \check{f} = \frac{1}{n} \sum_{i=1}^{n} \check{f}(X_i) \varepsilon_i$, assume that

$$R = \sup_{f \in \mathcal{F}} \sqrt{\mathbb{E} f(X_1)^2} < \infty .$$

Remark:

(Hoeffding) $\varepsilon_i \overset{iid}{\sim} \mathrm{Unif}\{-1, 1\}$

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} \alpha_i \varepsilon_i \right| > t \right) \leq 2 e^{-\frac{t^2}{2\|\alpha\|^2}}$$

Let $X_f = \frac{1}{n} \sum_{i=1}^{n} \check{f}(X_i) \varepsilon_i$, then

$$\mathbb{P}\left(|X_f - X_g| > t \mid X \right)$$

$$= \mathbb{P}\left(\frac{1}{n} \left| \sum_{i=1}^{n} (\check{f}(X_i) - \check{g}(X_i)) \varepsilon_i \right| > t \mid X \right)$$

$$\leq 2 e^{-\frac{nt^2}{2 \sum_{i=1}^{n} (\check{f}(X_i) - \check{g}(X_i))^2}}$$

$$\Rightarrow \mathbb{P}\left(|X_f| > t \right) \leq 2 e^{-\frac{t^2}{2 d(\check{f}, \check{g})}},$$

(why?)

where $d(\check{f}, \check{g}) = \sqrt{\mathbb{E}(\check{f}(X) - \check{g}(X))^2}$.

Upper bound by one-step discretization

If T is totally bounded, let T_δ be the δ-covering for T, then $\forall \theta \in T$,

$$\exists t \in T_\delta, \quad d(\theta, t) \le \delta,$$

$$X_\theta = X_\theta - X_t + X_t$$

$$\Rightarrow \mathbb{E} \sup_{\theta \in T} |X_\theta| \le \mathbb{E} \sup_{d(\theta_1, \theta_2) \le \delta} |X_{\theta_1} - X_{\theta_2}| + \mathbb{E} \max_{t \in T_\delta} |X_t|,$$

If $\delta \downarrow 0^+$,

$\sup_{d(\theta_1, \theta_2) \le \delta} |X_{\theta_1} - X_{\theta_2}| = o_p(1)$, if X_θ is sub-Gaussian process, (Why?) which implies that if δ is small enough, then

$\mathbb{E} \max_{t \in T_\delta} |X_t|$ is a tight upper bound for $\mathbb{E} \sup_{\theta \in T} |X_\theta|$. Similarly, if

$\mathbb{E} X_\theta = 0$, $\forall \theta \in T$, then

$$\mathbb{E}\sup_{\theta\in T} X_\theta = \mathbb{E}\sup_{\theta\in T} X_\theta - X_{\theta_0}$$

$$\le \mathbb{E}\sup_{\theta_1,\theta_2\in T} |X_{\theta_1} - X_{\theta_2}|$$

And

$$|X_{\theta_1} - X_{\theta_2}| \le |X_{t_1} - X_{t_2}| + |X_{t_1} - X_{\theta_1}| + |X_{t_2} - X_{\theta_2}|,\ t_1, t_2 \in T_\delta$$

$$\Rightarrow \mathbb{E}\sup_{\theta_1,\theta_2\in T} |X_{\theta_1} - X_{\theta_2}| \lesssim$$

$$\mathbb{E}\sup_{d(\theta_1,\theta_2)\le\delta} |X_{\theta_1} - X_{\theta_2}| + o_p(1)$$

$$\mathbb{E}\sup_{t_1,t_2\in T_\delta} |X_{t_1} - X_{t_2}|$$

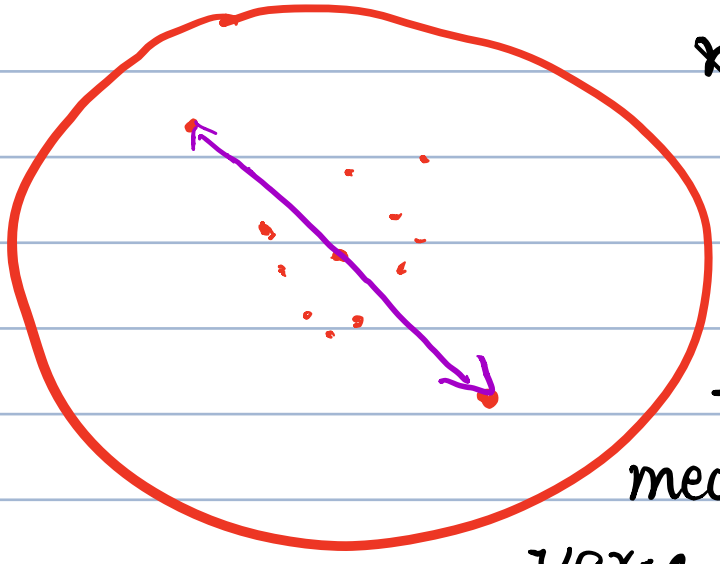
If $D = \mathrm{diam}(T) < \infty$, then (Lemma 3)

$$\mathbb{E}\sup_{\theta_1,\theta_2\in T} |X_{\theta_1} - X_{\theta_2}| \lesssim o(1) + \sqrt{D^2 \log N_\delta},$$

where $N_\delta =: N(\delta; T, d)$.

Chaining

Somehow, $\sqrt{D^2 \log N_\delta}$ is not a tight bound for $\mathbb{E}\sup_{t_1,t_2\in T_\delta}|X_{t_1}-X_{t_2}|$, and



note that $N_\delta \uparrow +\infty$ as $\delta \downarrow 0$, but $\mathbb{E}\sup_{t_1,t_2\in T_\delta}|X_{t_1}-X_{t_2}|$ tents to decrease, which means that if δ is very small, the upper bound for $\mathbb{E}\sup_{\theta_1,\theta_2\in T_\delta}|X_{\theta_1}-X_{\theta_2}|$ is not suitable, i.e.,

one chooses δ so as to achieve the optimal trade-off between two terms.

Assume the $|T|<\infty$ and $\operatorname{diam}(T) = D < \infty$. Let $\delta_0 = D$,

$$\delta_k \downarrow 0, \quad k \to \infty.$$

Given $L > 0$, T_k is the minimal δ_k-covering of T, $k = 0, \cdots, L$.

Theorem 1 $\{X_\theta, \theta \in T\}$ is zero-mean process and $\operatorname{diam}(T) := D < \infty$, then

$$\mathbb{E} \sup_{\theta_1, \theta_2 \in T} |X_{\theta_1} - X_{\theta_2}| \lesssim J(T, d)$$

where $J(\delta; T) := \int_0^D \sqrt{\log N(\varepsilon; T, d)}\, d\varepsilon$.

Pf: $\forall \theta_1, \theta_2 \in T$, $T_{L+1} := T$, and

$$\mathbb{E} \max_{\theta_1, \theta_2 \in T} |X_{\theta_1} - X_{\theta_2}|$$

(assume that $|T| < \infty$)

$$\lesssim \mathbb{E} \max_{t_1 \in T_{L+1}, t_2 \in T_L, d(t_1, t_2) \le \delta_L} |X_{t_1} - X_{t_2}|$$

$$+ \mathbb{E} \max_{t_1, t_2 \in T_L} |X_{t_1} - X_{t_2}|$$

$$\lesssim \sum_{i=1}^{L} \mathbb{E} \max_{t_1 \in T_{i+1}, t_2 \in T_i, d(t_1, t_2) \le \delta_i} |X_{t_1} - X_{t_2}|$$

$$\lesssim \sum_{i=1}^{L} \sqrt{\delta_i^2 \log\left(N(\delta_{i+1}; T, d)\, N(\delta_i, T, d)\right)}$$

$$\lesssim \sum_{i=1}^{L} \sqrt{\log N(\delta_{i+1}; T, d)}\, \delta_i$$

(Let $\delta_{i+1} = \delta_i / 2$)

$$\leq \sum_{i=1}^{L} \sqrt{\log N(\delta_{i+1}; T, d)}\ \delta_{i+1}/2$$

$$\leq \int_0^D \sqrt{\log N(\delta; T, d)}\, d\delta$$

$$=: J(T, d)$$

Entropy Integral

□

Remark:

If $\{X_\theta, \theta \in T\}$ is the sub-Gaussian, $\operatorname{diam}(T) < \infty$, then

$$\mathbb{E} \sup_{\theta \in T} X_t \leq J(T, d)$$

Example 1

$$\mathbb{E} \sup_{\theta \in B} \langle \theta, \epsilon \rangle \leq \int_0^1 \sqrt{\log N(\delta; B, \|\cdot\|)}\, d\delta$$

$$\leq \int_0^1 \sqrt{\log C(1+\tfrac{2}{\delta})^p}\, d\delta$$

$$\leq \sqrt{p} \int_0^1 \sqrt{\log C(1+\tfrac{2}{\delta})}\, d\delta \leq \sqrt{p}$$

Example 2

If $\mathcal{F}$ is the class of functions on X, and $0 \in \mathcal{F}$, $D =: \operatorname{diam}(\mathcal{F})$

Remark: If $\exists \theta_0 \in T$ s.t.

$$X_{\theta_0} = 0 \quad a.s.$$

Then

$$\mathbb{E} \sup_{\theta \in T} |X_\theta| = \mathbb{E} \sup_{\theta \in T} |X_\theta - X_{\theta_0}|$$
$$\leq \mathbb{E} \sup_{\theta_1, \theta_2} |X_{\theta_1} - X_{\theta_2}|$$
$$\Rightarrow \mathbb{E} \sup_{\theta \in T} |X_\theta| \leq J(T, d)$$

Note that $\phi: f \mapsto (f-T)^2$ is continuous since

$$P((f-T)^2 - (g-T)^2)^2 = P(f-g)^2(f+g-T)^2$$
$$\leq P(f-g)^2$$

Then $\overline{\mathcal{F}}$ is also bounded and $\exists C > 0$ s.t.

$$N(\varepsilon; \overline{\mathcal{F}}, d) \leq N(C\varepsilon; \mathcal{F}, d)$$

According to Theorem 1:

$$R(\hat{f}_n) - R(f^*) \leq \mathbb{E} \|P_n^\varepsilon\|_{\overline{\mathcal{F}}}$$

$$= \frac{\mathbb{E} \|\sqrt{n} P_n^\varepsilon\|_{\overline{\mathcal{F}}}}{\sqrt{n}}$$

$$\lesssim \frac{\int_0^{+\infty} \sqrt{\log N(\delta; \mathcal{F}, d)}\, d\delta}{\sqrt{n}}$$

$$\lesssim \frac{\int_0^{+\infty} \sqrt{\log N(\delta; \mathcal{F}, d)}\, d\delta}{\sqrt{n}}$$

If $V(\mathcal{F}) =: v < \infty$, then

$$R(\hat{f}_n) - R(f^*) \lesssim \frac{\int_0^{+\infty} \sqrt{v \log \frac{1}{\delta}}\, d\delta}{\sqrt{n}}$$

$$\asymp \sqrt{\frac{v}{n}}$$

Example 3

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \le x)$$

$$= P f_x, \text{ where } f_x(y) = \mathbb{1}(y \le x)$$

Then 正 $\|\hat{F}_n - F\|_\infty \lesssim \sqrt{\frac{1}{n}}$

$\Rightarrow \|\hat{F}_n - F\|_\infty = O_p(1)$.