

# **SDS 384 11: Theoretical Statistics**

## **Lecture 14: Uniform Law of Large Numbers- Covering number**

---

Purnamrita Sarkar  
Department of Statistics and Data Science  
The University of Texas at Austin

# A parametric class

## Example

For any fixed  $\theta$ , define the real-valued function  $f_\theta(x) := 1 \exp(-\theta|x|)$ , and consider the function class

$$\mathcal{F} = \{f_\theta : [0, 1] \rightarrow \mathcal{R} \mid \theta \in [0, 1]\}$$

Using the uniform norm as a metric, i.e.

$\|f - g\|_\infty := \sup_{x \in [0, 1]} |f(x) - g(x)|$ . Prove that

$$\left\lfloor \frac{1 - 1/e}{2\delta} \right\rfloor + 1 \leq N(\delta; \mathcal{F}, \|\cdot\|_\infty) \leq \frac{1}{2\delta} + 2.$$

## Proof-upper bound

- First note that  $|f(\theta) - f(\theta')| \leq |\theta - \theta'|$
- For any  $\delta \in (0, 1)$ , let  $T = \lfloor \frac{1}{2\delta} \rfloor$
- Consider  $S = \{\theta^0, \dots, \theta^{T+1}\}$  where  $\theta^i = 2\delta i$  for  $i \leq T$  and  $\theta^{T+1} = 1$ .
- $\{f_{\theta^i} : \theta^i \in S\}$  is a  $\delta$  cover for  $\mathcal{F}$ .
- For any  $\theta \in [0, 1]$  we can find  $\theta^i \in S$  such that  $|\theta^i - \theta| \leq \delta$
- Indeed we have,

$$\begin{aligned}\|f_{\theta^i} - f_{\theta}\|_{\infty} &= \sup_{x \in [0, 1]} |\exp(-\theta^i |x|) - \exp(-\theta |x|)| \\ &\leq |\theta^i - \theta| \leq \delta\end{aligned}$$

$$\text{So } N(\delta; \mathcal{F}, \|\cdot\|_{\infty}) \leq 2 + T \leq 2 + \frac{1}{\delta}$$

## Proof-lower bound

- We will do a  $\delta$  packing.
- Let  $\theta^i = -\log(1 - i\delta)$  for  $i = 0, \dots, T$
- $-\log(1 - T\delta) = 1$ , and so the largest integral value is  $T = \lfloor \frac{1 - 1/e}{\delta} \rfloor$
- So  $M(\delta; \mathcal{F}, \|\cdot\|_\infty) \geq 1 + \lfloor \frac{1 - 1/e}{\delta} \rfloor$
- $N(\delta; \mathcal{F}, \|\cdot\|_\infty) \geq M(2\delta; \mathcal{F}, \|\cdot\|_\infty) \geq 1 + \lfloor \frac{1 - 1/e}{2\delta} \rfloor$

## Make a comparison

- Recall that for a  $L$  Lipschitz continuous functions supported on  $[0, 1]$  with  $f(0) = 0$ , the metric entropy was  $L/\delta$
- Also recall that for a  $L$  Lipschitz continuous functions supported on  $[0, 1]^d$  with  $f(0) = 0$ , the metric entropy was  $(L/\delta)^d$
- However for a given function class like the last one the metric entropy is  $\log(1/\delta)$
- Recall that for Unit hypercubes in  $d$  dimensions the metric entropy is  $d \log(1 + 1/\delta)$
- Note that for Lipschitz continuous functions the dependence on  $d$  is exponential. This is a much richer class of functions, so the size is considerably larger and scales poorly with  $d$ .

# A Stochastic Process

- Consider a set  $\mathcal{T} \subseteq \mathcal{R}^d$ .
- The family of random variables  $\{X_\theta : \theta \in \mathcal{T}\}$  define a Stochastic process indexed by  $\mathcal{T}$ .
- We are often interested in the behavior of this process given its dependence on the structure of the set  $\mathcal{T}$ .
- In the other direction, we want to know the structure of  $\mathcal{T}$  given the behavior of this process.

## Definition

A canonical Gaussian process indexed by  $\mathcal{T}$  is defined as:

$$G_{\theta} := \langle z, \theta \rangle = \sum_k z_k \theta_k,$$

where  $z_k \stackrel{\text{iid}}{\sim} N(0, 1)$ . The supremum  $\mathcal{G}(\mathcal{T}) := E_z[\sup_{\theta \in \mathcal{T}} G_{\theta}]$  is the Gaussian complexity of  $\mathcal{T}$ .

# Rademacher complexity

- Replacing the iid standard normal variables by iid Rademacher random variables gives a Rademacher process  $\{R_\theta, \theta \in \mathcal{T}\}$ , where

$$R_\theta := \langle \epsilon, \theta \rangle = \sum_k \epsilon_k \theta_k, \quad \text{where } \epsilon_k \stackrel{\text{iid}}{\sim} \text{Uniform}\{0, 1\}$$

- $\mathcal{R}(\mathcal{T}) := E_\epsilon[\sup_{\theta \in \mathcal{T}} R_\theta]$  is called the Rademacher complexity of  $\mathcal{T}$ .



# How does this relate to the former notions of Rademacher complexity?

- Recall that

$$\mathcal{R}_{\mathcal{F}} := E[\sup_{f \in \mathcal{F}} |\sum_i \epsilon_i f(X_i)|] = E[E[\sup_{f \in \mathcal{F}} |\sum_i \epsilon_i f(X_i)| | X_1, \dots, X_n]]$$

- Now the inner expectation can be upper bounded by

$$E_{\epsilon} \sup_{\theta \in \mathcal{T} \cup -\mathcal{T}} \sum_i \epsilon_i \theta_i, \text{ where } \mathcal{T} \subseteq \mathbb{R}^n \text{ can be written as}$$

$$\mathcal{T} = \{(f(X_1), \dots, f(X_n)) | f \in \mathcal{F}\}$$

## Theorem

For  $\mathcal{T} \in \mathbb{R}^d$ ,

$$\mathcal{R}(\mathcal{T}) \leq \sqrt{\frac{\pi}{2}} \mathcal{G}(\mathcal{T}) \leq c \sqrt{\log d} \mathcal{R}(\mathcal{T})$$

- This is showing that there can be there are some sets where the Gaussian complexity can be substantially larger than the Rademacher complexity.
- We will in fact give an example.

## Proof (of first inequality)

$$\begin{aligned} E[\mathcal{G}(\mathcal{T})] &= E \sup_{\theta \in \mathcal{T}} \sum_i z_i \theta_i \\ &= E \sup_{\theta \in \mathcal{T}} \sum_i \epsilon_i |z_i| \theta_i \\ &= E_\epsilon E_Z \sup_{\theta \in \mathcal{T}} \sum_i \epsilon_i |z_i| \theta_i \\ &\geq E_\epsilon \sup_{\theta \in \mathcal{T}} \sum_i \epsilon_i E |z_i| \theta_i \\ &= \sqrt{\frac{2}{\pi}} \mathcal{R}(\mathcal{T}) \end{aligned}$$

# Example

## Example

Consider the  $L_1$  ball in  $\mathcal{R}^d$  denoted by  $B_1^d$ .

$$\mathcal{R}(B_1^d) = 1, \mathcal{G}(B_1^d) \leq \sqrt{2 \log d}$$

- $\mathcal{R}(B_1^d) = E\left[\sup_{\|\theta\|_1 \leq 1} \sum_i \theta_i \epsilon_i\right] = E[\|\epsilon\|_\infty] = 1$
- Similarly,  $\mathcal{G}(B_1^d) = E[\|z\|_\infty]$

## Recall the finite class lemma?

### Theorem

*Consider  $z$  with independent sub-gaussian components.*

$$E \max_{a \in A} \langle z, a \rangle \leq \max_{a \in A} \|a\| \sqrt{2 \log |A|}$$

- In our case,  $A = \{e_i, i \in [d]\}$ ,  $|A| = d$  and  $\max_{a \in A} \|a\| = 1$ .

## Theorem

Consider a random matrix  $M = (\xi_{ij})_{i,j \in [n]}$  where  $\xi_{ij}$  are standard normal random variables.

$$P(\|M\|_{op} \geq A\sqrt{n}) \leq C \exp(-cAn)$$

where  $c, C$  are absolute constants and  $A \geq C$ .

- This works for symmetric wigner ensembles and hermitian matrices as well.

# Operator norm

- Let  $S_n := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$
- $\|M\|_{op} := \sup_{x \in \mathbb{R}^n} \|Mx\|$
- First note that we have

$$P(\|Mx\| \geq A\sqrt{n}) \leq C \exp(-cAn)$$

- This is because for each row  $M_i$ , we have

$$M_i^T x \sim \text{Subgaussian}(1), (M_i^T x)^2 - 1 \sim \text{Subexponential}(2, 4)$$

- $\|Mx\|^2 - n \sim \text{Subexponential}(2\sqrt{n}, 4)$
- So  $P(\|Mx\| \geq A\sqrt{n}) \leq C \exp(-cAn)$

## Can I just use an Union bound?

- Not really.
- But I can form a  $1/2$  cover of  $S_n$ .