THE UNIVERSITY OF TEXAS AT AUSTIN
Department of Statistics and Data Sciences
College of Natural Sciences

SDS 384 11: Theoretical Statistics

Lecture 15: Uniform Law of Large Numbers-

Rademacher and Gaussian Complexity

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

A parametric class

Example

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

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

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

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

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

Proof-upper bound

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

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

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

Proof-lower bound

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

Make a comparison

- Recall that for a L Lipschitz continuous functions supported on $[0, 1]$ with $f(0) = 0$, the metric entropy was L/δ
- 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 $\mathcal{T} \subseteq \mathcal{R}^{\boldsymbol{d}}$.
- 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.

Gaussian and Rademacher processes

Definition

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

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

where $z_k \overset{\text{iid}}{\sim} N(0,1)$. 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

$$R_\theta := \langle \epsilon, \theta \rangle = \sum_k \epsilon_k \theta_k, \qquad \text{where } \epsilon_k \overset{\text{iid}}{\sim} \mathit{Uniform}\{-1, 1\}$$

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

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

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

- Now the inner expectation can be upper bounded by
  $E_\epsilon \sup_{\theta\in\mathcal{T}\bigcup -\mathcal{T}} \sum_i \epsilon_i\theta_i$, where $\mathcal{T}\subseteq\mathbb{R}^n$ can be written as

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

Relationship

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)

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

Example

Example

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

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

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

Recall the finite class lemma?

Theorem

Consider z with independent sub-gaussian components.

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

- In our case, $A = \{e_i, i \in [d]\}$, $e_i(j) = \pm 1(j = i)$, $|A| = 2d$ and $\max_{a \in A} \|a\| = 1$.
- This gives a weaker bound on the Gaussian complexity.

Application-Random matrix singular value

Theorem

Consider a random matrix $M = (\xi_{ij})_{i,j \in [n]}$ where ξ_{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

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

- $\|Mx\|^2 - n \sim Subexponential(2\sqrt{n}, 4)$

Recall sub-exponential random variables?

Theorem

Let X be a sub-exponential random variable with parameters (ν, b). Then,

$$P(X \geq \mu + t) \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}} & \textit{if } 0 \leq t \leq \frac{\nu^2}{b} \\ e^{-\frac{t}{2b}} & \textit{if } t \geq \frac{\nu^2}{b} \end{cases}$$

- $P(\|Mx\|^2 - n \geq Cn) = \leq e^{-Cn/8}$, $C > 1$.

Can I just use an Union bound?

- Not really.
- But I can form a $1/2$ cover of S_n.
- Find $\mathcal{C} = \{x^1, \ldots, x^N\}$ such that for all $x \in S_n$, $\exists x^i \in S$ $\|x - x^i\| \leq 1/2$.
- Consider $y \in S$ such that $\|My\| = \|M\|_{op}$. Let x^i be a member of the $1/2$ cover s.t. $\|y - x^i\| \leq 1/2$
- So $\|M(y - x^i)\| \leq \|M\|_{op}/2$ and $\|M(y - x^i)\| \geq \|My\| - \|Mx^i\| \geq \|M\|_{op}$.
- Hence $\|Mx^i\| \geq \|M\|_{op}/2$

Using the covering number

$$
\begin{aligned}
P(\|M\|_{op} \geq \sqrt{(C+1)n}) &\leq P(\exists x^i \in \mathcal{C}, \|Mx^i\| \geq \sqrt{(C+1)n}/2) \\
&\leq |\mathcal{C}| P(\|Mx^i\| \geq \sqrt{(C+1)n}/2) \\
&\leq |\mathcal{C}| P(\|Mx^i\|^2 - n \geq (C-3)n/4) \\
C > 7 \text{ gives } (C-3)n/4 \geq \nu^2/b \qquad &\leq |\mathcal{C}| \exp(-(C-3)n/8)
\end{aligned}
$$

- ϵ covering number of the unit ball in n dimensions is bounded by $(1+2/\epsilon)^n$

$$
\begin{aligned}
P(\|M\|_{op} \geq \sqrt{(C+1)n}) &\leq 5^n \exp(-(C-3)n/8) \\
&\leq \exp(-n((C-3)/8 - 1.6))
\end{aligned}
$$

- So C will have to be something like 19!!

Kernel density estimation

Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples of random variable with density f on the real line with support $[0, 1]$. A standard estimate of f is the kernel density estimate

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where $K : \mathbb{R} \to [0, \infty]$ is a kernel function satisfying $\int_{-\infty}^{\infty} K(t)dt = 1$, and h is a bandwidth parameter. Also assume that $|K(x) - K(y)| \leq L|x - y|$. Let $K(x) \leq K(0)$.

We are interested in the quantity $\sup_{x \in [0,1]} |\hat{f}(x) - E[\hat{f}(x)]|$

Kernel Density Estimation

- First do a ϵ cover of x by $\mathcal{C} := \{x^1, \ldots, x^N\}$.
- Let $\tilde{K}((x - X_i)/h) = K(.) - EK(.)$
- Similarly $\tilde{f}(.) = \hat{f}(.) - E[\hat{f}(.)]$
- The Lipschitz condition gives $\left|\tilde{K}\left(\frac{x - X_i}{h}\right) - \tilde{K}\left(\frac{y - X_i}{h}\right)\right| \leq \frac{2L|x - y|}{h}$
- So $|\tilde{f}(x) - \tilde{f}(x^i)| \leq \frac{2L|x - x^i|}{h^2}$
- So this gives a $2L\epsilon/h^2$ cover for the $\tilde{f}$ values.

Kernel Density Estimation

- Let y be the point where $\sup_{x \in [0,1]} |\tilde{f}(x)|$ is achieved.
- There exists a i such that $|\tilde{f}(y) - \tilde{f}(x^i)| \leq 2L\epsilon/h^2$
- So $\exists i, |\tilde{f}(x^i)| \geq \sup_{x \in [0,1]} |\tilde{f}(x)| - 2L\epsilon/h^2$
- Finally

$$
\begin{aligned}
P\left(\sup_{x \in [0,1]} |\tilde{f}(x)| \geq \delta\right) &\leq P(\exists i \in \mathcal{C}, |\tilde{f}(x^i)| \geq \sup_{x \in [0,1]} |\tilde{f}(x)| - 2L\epsilon/h^2) \\
&\leq |\mathcal{C}| P\left(|\tilde{f}(x^i)| \geq \delta - 2L\epsilon/h^2\right)
\end{aligned}
$$

- Set $\delta = 4L\epsilon/h^2$, the RHS can be obtained using Hoeffding.
- Can you derive it?