Metric Entropy

Group Study and Seminar Series (Summer 20)

Presented by: Jiancong Xiao

The Chinese University of Hong Kong, Shenzhen, China

August 6, 2020

Mainly based on:
Wainwright, M. J. (2019). High-dimensional statistics: A non-asymptotic viewpoint (Vol. 48). Chapter 5.
MIT IDS.160 / 18.S998 / 9.521 Spring 20. Mathematical Statistics: A Non-Asymptotic Approach. Lecture 16-19.

Outline

Suprema of Subgaussian Processes

Definition 1.
Stochastic process $(U_\theta)_{\theta \in \Theta}$, indexed by $\theta \in \Theta$, is a collection of random variables on a common probability space.

- The index θ can be 'time'.
- We are interested in the case that Θ has some metric structure.
- We will be interested in the behavior of

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta$$

Subgaussian process

To understand this object, we need

Definition 2.

Stochastic process $(U_\theta)_{\theta \in \Theta}$ is sub-Gaussian with respect to a metric d on θ if U_θ is zero-mean and

$$\forall \theta, \theta' \in \Theta, \lambda \in \mathbb{R}, \quad \mathbb{E} \exp\left\{\lambda(U_\theta - U_{\theta'})\right\} \leq \exp\left\{\lambda^2 d(\theta, \theta')^2/2\right\}$$

- $U_\theta - U_{\theta'}$ is subgaussian with $\sigma = d(\theta, \theta')$
- The main examples have a linearly parametrized form

Gaussian process and Rademacher process

- **Gaussian process**
  Let $G_\theta = \langle g, \theta \rangle$, $g = (g_1, \cdots, g_n)^T$, $g_i \sim N(0,1)$ i.i.d. Take $d(\theta, \theta') = \|\theta - \theta'\|$. Then

$$G_\theta - G_{\theta'} = \langle g, \theta - \theta' \rangle \sim N(0, \|\theta - \theta'\|^2)$$

  is trivially subgaussian with respect to the Euclidean distance on Θ.
- **Rademacher process**
  Let $R_\theta = \langle \epsilon, \theta \rangle$, $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T$, ϵ i.i.d. Rademacher. Again, take $d(\theta, \theta') = \|\theta - \theta'\|$. Then

$$R_\theta - R_{\theta'} = \langle \epsilon, \theta - \theta' \rangle$$

  is subgaussian.

Relationship between Gaussian and Rademacher Process

Definition 3.
We will call $\hat{\mathcal{R}}(\Theta) = \mathbb{E}\sup_{\theta\in\Theta} R_\theta = \mathbb{E}\sup_{\theta\in\Theta}\langle\epsilon,\theta\rangle$ the (empirical) Rademacher averages of Θ. The corresponding expected supremum of the Gaussian process will be called the Gaussian averages or the Gaussian width of Θ and denoted by $\hat{\mathcal{G}}(\Theta)$.

- Rademacher complexity of Θ is $\frac{1}{n}\hat{\mathcal{R}}(\Theta)$ (Qingyan's present on July 16th)
- **Property 1**
  $\forall\Theta \subset \mathbb{R}^n$, we have

$$\hat{\mathcal{R}}(\Theta) \lesssim \hat{\mathcal{G}}(\Theta) \lesssim \sqrt{\log n}\,\hat{\mathcal{R}}(\Theta)$$

Relationship between Gaussian and Rademacher Process

Proof of Property 1(a): $\hat{\mathcal{R}}(\Theta) \overset{a}{\lesssim} \hat{\mathcal{G}}(\Theta) \overset{b}{\lesssim} \sqrt{\log n}\hat{\mathcal{R}}(\Theta)$

$$
\begin{aligned}
\hat{\mathcal{G}}(\Theta) &= \mathbb{E} \sup_{\theta \in \Theta} \sum_{i=1}^{n} g_i \theta_i \\
&= \mathbb{E}_\epsilon \mathbb{E}_g \sup_{\theta \in \Theta} \sum_{i=1}^{n} \epsilon_i |g_i| \theta_i \\
&\geq \mathbb{E}_\epsilon \sup_{\theta \in \Theta} \sum_{i=1}^{n} \epsilon_i \mathbb{E}|g_i| \theta_i \\
&= \sqrt{\frac{2}{\pi}} \mathbb{E}_\epsilon \sup_{\theta \in \Theta} \sum_{i=1}^{n} \epsilon_i \theta_i \\
&= \sqrt{\frac{2}{\pi}} \hat{\mathcal{R}}(\Theta)
\end{aligned}
$$

Relationship between Gaussian and Rademacher Process

Proof of Property 1(b): $\hat{\mathcal{R}}(\Theta) \overset{a}{\lesssim} \hat{\mathcal{G}}(\Theta) \overset{b}{\lesssim} \sqrt{\log n}\hat{\mathcal{R}}(\Theta)$

$$
\begin{aligned}
\hat{\mathcal{G}}(\Theta) &= \mathbb{E} \sup_{\theta \in \Theta} \sum_{i=1}^{n} g_i \theta_i \\
&= \mathbb{E}_\epsilon \mathbb{E}_g \sup_{\theta \in \Theta} \sum_{i=1}^{n} \epsilon_i |g_i| \theta_i \\
&= \mathbb{E}_g \hat{\mathcal{R}}(|g| \cdot \Theta) \\
&\leq \mathbb{E}_g \max_i |g_i| \hat{\mathcal{R}}(\Theta) \quad \text{(Lipschitz Property, week 5)} \\
&\leq \sqrt{2 \log 2n}\hat{\mathcal{R}}(\Theta) \quad \text{(Page 56, week 2)}
\end{aligned}
$$

A few examples

- **Example 1**: $\Theta = \mathbb{B}_2^n$
  Consider the Rademacher and Gaussian complexity of Euclidean ball $\mathbb{B}_2^d = \{\theta | \|\theta\|_2 \leq 1\}$, by Cauchy-Schwartz inequality, it is easy to have

$$\hat{\mathcal{R}}(\mathbb{B}_2^n) = \mathbb{E} \sup_{\|\theta\|_2 \leq 1} \langle \epsilon, \theta \rangle = \mathbb{E}\|\epsilon\|_2 = \sqrt{n}$$

  and

$$\hat{\mathcal{G}}(\mathbb{B}_2^n) = \mathbb{E} \sup_{\|\theta\|_2 \leq 1} \langle g, \theta \rangle = \mathbb{E}\|g\|_2 \leq \sqrt{\mathbb{E}\|g\|_2^2} = \sqrt{n}$$

- Actually $\mathbb{E}\|g\|_2 \asymp \sqrt{n}$
- This is an example for the left inequality $\hat{\mathcal{R}}(\Theta) \lesssim \hat{\mathcal{G}}(\Theta)$

A few examples

- **Example 2**: $\Theta = \mathbb{B}_1^n$
  Consider the Rademacher and Gaussian complexity of $\mathbb{B}_1^d = \{\theta | \|\theta\|_1 \leq 1\}$, again by holder's inequality, we have
  $$\hat{\mathcal{R}}(\mathbb{B}_1^n) = \mathbb{E} \sup_{\|\theta\|_1 \leq 1} \langle \epsilon, \theta \rangle = \mathbb{E}\|\epsilon\|_\infty = 1$$
  and
  $$\hat{\mathcal{G}}(\mathbb{B}_1^n) = \mathbb{E} \sup_{\|\theta\|_1 \leq 1} \langle g, \theta \rangle = \mathbb{E}\|g\|_\infty \leq \sqrt{2\log(2n)}$$
- The last inequality is from Page 56 (week 2)
- This is an example for the left inequality $\hat{\mathcal{G}}(\Theta) \lesssim \sqrt{\log n}\hat{\mathcal{R}}(\Theta)$

Outline

finite-class lemma

Recap: How to bound the Gaussian or Rademacher complexity?

- finite-class: Massart lemma.
- infinite-class: build the ϵ-net and use the covering number.

Lemma 4 (Massart).

Let d be a metric on Θ and assume (U_θ) is a subgaussian process. Then for any finite subset $A \subseteq \Theta \times \Theta$,

$$\mathbb{E} \max_{(\theta,\theta')\in A} U_\theta - U_{\theta'} \leq \max_{(\theta,\theta')\in A} d(\theta - \theta')\sqrt{2 \log \mathit{card}(A)}$$

Definition 5 (covering number).

Let (Θ, d) be a metric space. A set $\theta_1, \cdots, \theta_N \in \Theta$ is a cover of Θ at scale ϵ for any θ there exists $j \in [N]$ such that $d(\theta, \theta_j) \leq \epsilon$. The covering number of Θ at scale ϵ is the size of the smallest cover, denoted by $\mathcal{N}(\Theta, d, \epsilon)$.

finite-class lemma

A simple consequence of Lemma 5 is

Lemma 6 (Single scaled upper bound).

If $(U_\theta)_{\theta\in\Theta}$ is subgaussian with respect to d on Θ, then for any $\delta > 0$,

$$\mathbb{E}\sup_{\theta\in\Theta} U_\theta \leq 2\mathbb{E}\sup_{d(\theta,\theta')\leq\delta}(U_\theta - U_{\theta'}) + 2\mathit{diam}(\Theta)\sqrt{\log\mathcal{N}(\Theta,d,\delta)}$$

Proof:

$$\mathbb{E}\sup_{\theta\in\Theta} U_\theta = \mathbb{E}\sup_{\theta\in\Theta} U_\theta - U_{\theta'} \leq \mathbb{E}\sup_{\theta,\theta'\in\Theta} U_\theta - U_{\theta'}$$

Let $\hat{\Theta}$ be a δ-cover of Θ. Then

$$U_\theta - U_{\theta'} = U_\theta - U_{\hat\theta} + U_{\hat\theta} - U_{\hat\theta'} + U_{\hat\theta'} - U_{\theta'} \leq 2\sup_{d(\theta,\theta')\leq\delta}(U_\theta - U_{\theta'}) + \sup_{\hat\theta,\hat\theta'\in\hat\Theta}(U_{\hat\theta} - U_{\hat\theta'})$$

Example

Lemma 7 is not the best. Let us go back to example 1 that $\Theta \subset \mathbb{B}_2^n$. Then

- the first term

$$2\mathbb{E} \sup_{d(\theta,\theta')\leq\delta} (U_\theta - U_{\theta'}) = 2\mathbb{E} \sup_{\|\theta,\theta'\|\leq\delta} \langle g, \theta - \theta' \rangle \leq 2\delta\sqrt{n}$$

- the second term

$$2\text{diam}(\Theta)\sqrt{\log \mathcal{N}(\Theta, d, \delta)} \leq 2\sqrt{d \log(1 + \frac{2}{\delta})}$$

- Suppose Θ lies in a d dimensional subspace with $d < n$
- Remark: the covering number of $\mathbb{B}_2^d$ (We will prove it later by packing number)

$$\mathcal{N}(\Theta, \|\cdot\|_2, \delta) \leq \left(1 + \frac{2}{\delta}\right)^d$$

Example

Continue:

- Take $\delta = \sqrt{d/n}$

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta \leq 2\delta\sqrt{n} + 2\sqrt{d \log(1 + \frac{2}{\delta})} \leq \mathcal{O}(\sqrt{d \log(n/d)})$$

- We have already show that

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta \leq \mathcal{O}(\sqrt{d})$$

- Single scale upper bound is not the best
- the second term can be improved by chaining

Chaining

Definition 7 (δ-truncated Dudley's entropy integral).

Define D be the diameter of Θ. The δ-truncated Dudley's entropy integral is defined as

$$\mathcal{J}(\delta, D) = \int_{\delta}^{D/2} \sqrt{\log \mathcal{N}(\Theta, d, \epsilon)} d\epsilon$$

Theorem 8 (Dudley's entropy upper bound).

If $(U_\theta)_{\theta \in \Theta}$ is subgaussian with respect to d on Θ. Let $D = diam(\Theta)$, then for any $\delta \in [0, D]$,

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta \leq 2\mathbb{E} \sup_{d(\theta, \theta') \leq \delta} (U_\theta - U_{\theta'}) + 8\sqrt{2}\mathcal{J}(\delta/4, D)$$

Chaining

Proof: Do Lemma 6 in multiple steps. (Week 5, Theorem 11)

The best upper bound is

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta \leq \inf_{\delta \in [0,D]} \left[2\mathbb{E} \sup_{d(\theta,\theta') \leq \delta} (U_\theta - U_{\theta'}) + 8\sqrt{2}\mathcal{J}(\delta/4, D) \right]$$

- It is computational intractable.
- Simply take $\delta = 0$ we have

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta \leq 8\sqrt{2}\mathcal{J}(0, D)$$

- In example 1:

$$\mathbb{E} \sup_{\theta \in \Theta} U_\theta \leq 8\sqrt{2} \int_0^{D/2} \sqrt{\log \mathcal{N}(\Theta, \|\cdot\|, \epsilon)} d\epsilon = 8\sqrt{2} \int_0^{D/2} \sqrt{d \log(1 + \frac{2}{\epsilon}) } d\epsilon \leq \mathcal{O}(\sqrt{d})$$

Outline

Definition 9 (Packing number).

A δ-packing of a set θ with respect to a metric d is a set $\{\theta_1, \cdots, \theta_M\}$ such that $d(\theta_i, \theta_j) > \delta$ for all distinct $i, j \in \{1, 2, ..., M\}$.The δ-packing number $\mathcal{M}(\Theta, d, \delta)$ is the cardinality of the largest δ-packing.

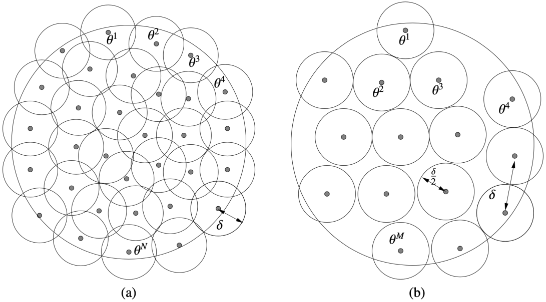

Figure: (a): covering and (b): packing

Covering and Packing

Lemma 10.
Let (Θ, d) *be a metric space, then*

$$\mathcal{M}(\Theta, d, 2\delta) \overset{a}{\leq} \mathcal{N}(\Theta, d, \delta) \overset{b}{\leq} \mathcal{M}(\Theta, d, \delta)$$

Proof of (a):

- Suppose there exists a 2δ-packing $\{y_1, \cdots, y_M\}$ and a δ-covering $\{x_1, \cdots, x_N\}$ with $M \geq N + 1$.
- By pigeonhole principle, $\exists i, j$ and k, s.t. $y_i, y_j \in B(x_k, \delta)$
- then $d(y_i, y_j) \leq 2\delta$
- contradiction $\square$

Covering and Packing

Continue,

$$\mathcal{N}(\Theta, d, \delta) \overset{b}{\leq} \mathcal{M}(\Theta, d, \delta)$$

Proof of (b):

- Suppose $E = \{\theta_1, \cdots, \theta_M\}$ is a maximal δ-packing.
- Then $\forall \theta \in \Theta \backslash E$, $\exists j$ s.t. $d(\theta, \theta_j) \leq \delta$
- (Otherwise, we can add θ to E to form a better packing)
- E is a δ-covering of Θ.

□

Sudakov's lower bound

Theorem 11 (Sudakov Minoration).

Let $(G_\theta)_{\theta \in \Theta}$ be a zero mean Gaussian process defined on Θ. Then

$$\mathbb{E} \sup_{\theta \in \Theta} G_\theta \geq \sup_{\delta > 0} \frac{\delta}{2} \sqrt{\log \mathcal{M}(\Theta, \|\cdot\|, \delta)}$$

- $\mathcal{M}(\Theta, \|\cdot\|, \delta)$ can be replaced by $\mathcal{N}(\Theta, \|\cdot\|, \delta)$ by lemma 10
- in example 1

$$\mathbb{E} \sup_{\theta \in \Theta} G_\theta \geq \sup_{\delta > 0} \frac{\delta}{2} \sqrt{d \log(1/\delta)} \gtrsim \sqrt{d}$$

metric entropy of unit balls

- Let B be the unit norm ball and d be the metric induced by the norm
- It lefts to show

$$(\frac{1}{\delta})^d \overset{a}{\leq} \mathcal{N}(B,d,\delta) \leq \mathcal{M}(B,d,\delta) \overset{b}{\leq} (1+\frac{2}{\delta})^d$$

- Proof of (a): Let $\{\theta_1, \cdots, \theta_N\}$ be a δ-covering of B, then $B \subset \cup_{j=1}^N [\theta_j + \delta B]$
- Then $vol(B) \leq N\delta^d vol(B)$
- Proof of (b): Let $\{\theta_1, \cdots, \theta_M\}$ be a δ-packing of B, then

$$Mvol(B\delta/2) \leq vol(B + B\delta/2)$$

- it is $M(\delta/2)^d vol(B) \leq (\delta/2)^d (1+2/\delta)^d vol(B)$ □

Sudakov's lower bound

Before we give the proof of Theorem 11, we state two fact (without proof).

- **fact 1: Sudakov-Fernique inequality**
  Given a pair of zero-mean Gaussian vectors $(X_1, \cdots, X_N)$ and $(Y_1, \cdots, Y_N)$ such that

  $$\mathbb{E}[(X_i - X_j)^2] \leq \mathbb{E}[(Y_i - Y_j)^2] \quad \forall i, j$$

  Then $\mathbb{E}[\max X_i] \leq \mathbb{E}[\max Y_i]$
- **fact 2:** If $X_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. then

  $$\sigma\sqrt{(1/2)\log N} \leq \mathbb{E}[\max X_i] \leq \sigma\sqrt{2\log N}$$

Sudakov's lower bound

Proof of Theorem 11:

Let $E = \{\theta_1, \cdots, \theta_M\}$ be a maximal δ-packing of Θ. Consider the sequence $Y_i = G_{\theta_i}$, we have

$$\mathbb{E}[(Y_i - Y_j)^2] = \|\theta_i - \theta_j\|^2 > \delta^2$$

Then we define $X_i \sim \mathcal{N}(0, \delta^2/2)$ i.i.d for $i = 1, \cdots, M$, we have

$$\mathbb{E}[(X_i - X_j)^2] = \delta^2$$

Then

$$\mathbb{E} \sup_{\theta \in \Theta} G_\theta \geq \mathbb{E} \max_{i=1,\cdots,M} Y_i \geq \mathbb{E} \max_{i=1,\cdots,M} X_i \geq \frac{\delta}{2}\sqrt{\log M}$$

□

illustration of upper bound and lower bound

- Combine the upper bound and lower bound, for Gaussian process, we have

$$C_1 \sup_{\delta>0} \delta \sqrt{\log \mathcal{N}(\Theta, \|\cdot\|, \delta)} \le \mathbb{E} \sup_{\theta\in\Theta} G_\theta \le C_2 \int_0^{D/2} \sqrt{\log \mathcal{N}(\Theta, \|\cdot\|, \epsilon)} d\epsilon$$

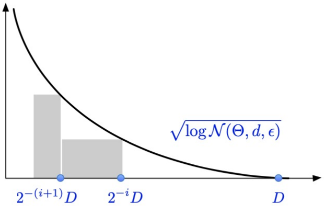

Figure: Illustration of upper bound and lower bound

Illustration of upper bound and lower bound

- Notice that the upper bound is for any subgaussian process
- While lower bound is only for gaussian process
- Recap:

$$\sqrt{\frac{2}{\pi}}\hat{\mathcal{R}}(\Theta) \le \hat{\mathcal{G}}(\Theta) \le \sqrt{2\log 2n}\,\hat{\mathcal{R}}(\Theta)$$

- the lower bound for Rademacher average is

$$\frac{C_3}{\sqrt{\log 2n}} \sup_{\delta>0} \delta \sqrt{\log \mathcal{M}(\Theta, \|\cdot\|, \delta)} \le \mathbb{E} \sup_{\theta\in\Theta} R_\theta$$

Outline

Function class and metric

- In Machine Learning Theory, we are interested in the complexity of Function class
- Given a set of functions $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$, a probability measure P on $\mathcal{X}$
- we define
$$\|f\|_{L^2(P)}^2 = \mathbb{E}f(X)^2$$
- Similarly, given a set of sample $X_1, \cdots, X_n$, we define a pseudometric
$$\|f\|_{L^2(P_n)}^2 = \frac{1}{n}\sum_{i=1}^{n} f(X_i)^2$$
- the ε-covering number and packing number is
$$\mathcal{N}(\mathcal{F}, L^2(P), \varepsilon) \quad \text{and} \quad \mathcal{M}(\mathcal{F}, L^2(P), \varepsilon)$$
- Remark: pseudometric: $d(x,y) = 0 \nRightarrow x = y$

Upper bound and lower bound of Rademacher complexity

- As before, Let $U_\theta = \langle \epsilon, \theta \rangle$, $\Theta = \frac{1}{\sqrt{n}} \mathcal{F}|_{x_1, \cdots, x_n}$, Then by Dudley's upper bound

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{n}} \sum_{i-1}^{n} \epsilon_i f(x_i) = \mathbb{E} \sup_{\theta \in \Theta} U_\theta \leq 2\delta\sqrt{n} + 8\sqrt{2}\mathcal{J}(\delta/4, D)$$

- Move the $\sqrt{n}$ to the left hand side, replace $\delta/4$ by δ the empirical Rademacher complexity is

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i-1}^{n} \epsilon_i f(x_i) \leq \inf_{\delta \geq 0} \left[8\delta + \frac{12}{\sqrt{n}} \int_{\delta}^{D/2} \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(P), \varepsilon)} d\varepsilon \right]$$

- By Sudakov's Lower bound

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i-1}^{n} \epsilon_i f(x_i) \geq \frac{c}{\log 2n} \sup_{\delta \geq 0} \sqrt{\frac{\log \mathcal{M}(\mathcal{F}, L^2(P), \varepsilon)}{n}}$$

Metric Entropy

Thank you!