Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 9

Yun Yang

- Uniform laws of large numbers via metric entropy

Covering and packing numbers

A way to measure the "size" of a set with infinitely many elements. Recall:

Definition

A metric space $(\mathbb{T}, \rho)$ consists of a non-empty set $\mathbb{T}$ equipped with a mapping $\rho : \mathbb{T} \times \mathbb{T} \to [0, \infty)$ satisfying:

1. $\rho(\theta, \theta') = 0$ if and only if $\theta = \theta'$;
2. It is symmetric: $\rho(\theta, \theta') = \rho(\theta', \theta)$;
3. Triangle inequality: $\rho(\theta, \theta'') \leq \rho(\theta, \theta') + \rho(\theta', \theta'')$.

If the first property is replaced with $\rho(\theta, \theta) = 0$, then $(\mathbb{T}, \rho)$ is called a pseudometric space.

Examples: Euclidean space $(\mathbb{R}^n, \|\cdot\|_2)$, function space $(L^2[0, 1], \|\cdot\|_\infty)$, function space with pseudometric $\rho(f, g) = \|f - g\|_n = \sqrt{n^{-1} \sum_{i=1}^n [f(x_i) - g(x_i)]^2}$.

Covering number

Definition

An ε-cover of a set $\mathbb{T}$ w.r.t. a metric ρ is a set $\{\theta^1, \ldots, \theta^N\} \subset \mathbb{T}$ such that for each $\theta \in \mathbb{T}$, there exists some $i \in \{1, \ldots, N\}$, $\rho(\theta, \theta^i) \leq \varepsilon$. The **$\varepsilon$-covering number** $N(\varepsilon, \mathbb{T}, \rho)$ is the smallest cardinality of all ε-covers.

A set $\mathbb{T}$ is **totally bounded** if for all $\varepsilon > 0$, $N(\varepsilon, \mathbb{T}, \rho) < \infty$ (compact?).

The function $\varepsilon \mapsto \log N(\varepsilon, \mathbb{T}, \rho)$ is the **metric entropy** of $\mathbb{T}$ w.r.t. ρ.

$N(\varepsilon, \mathbb{T}, \rho)$ is non-increasing in ε. Often interested in the growth of metric entropy as $\varepsilon \to 0_+$. If $\lim_{\varepsilon \to 0_+} \log N(\varepsilon) / \log(1/\varepsilon)$ exists, it is called the **metric dimension**.

Example: Covering number of unit cubes

Example

Consider interval $[-1, 1]$ in $\mathbb{R}$, equipped with the Euclidean metric $|\cdot|$. Then we have

$$N(\varepsilon, [-1, 1], |\cdot|) \leq \frac{1}{\varepsilon} + 1, \quad \text{for all } \varepsilon > 0.$$

More generally, for the d-dim cube $[-1, 1]^d$, we have $N(\varepsilon, [-1, 1]^d, \|\cdot\|_\infty) \leq \left(\frac{1}{\varepsilon} + 1\right)^d$, and its metric dimension is d.

Packing number

Definition

An ε-packing of a set $\mathbb{T}$ w.r.t. a metric ρ is a set $\{\theta^1, \ldots, \theta^M\} \subset \mathbb{T}$ such that $\rho(\theta^i, \theta^j) > \varepsilon$ for all distinct pairs $(i, j) \in \{1, \ldots, M\}^2$. The **$\varepsilon$-packing number** $M(\varepsilon, \mathbb{T}, \rho)$ is the largest cardinality of all ε-packings.

Covering and packing relation

Theorem

For all $\varepsilon > 0$, the packing and covering numbers are related by:

$$M(2\varepsilon, \mathbb{T}, \rho) \leq N(\varepsilon, \mathbb{T}, \rho) \leq M(\varepsilon, \mathbb{T}, \rho).$$

Thus, the scalings of the covering and packing numbers are the same.

Example: Packing number of unit cubes

Example

Consider interval $[-1,\, 1]$ in $\mathbb{R}$, equipped with the Euclidean metric $|\cdot|$. Then we have

$$M(2\varepsilon,\, [-1,\, 1],\, |\cdot|) \geq \left\lfloor \frac{1}{\varepsilon} \right\rfloor, \quad \text{for all } \varepsilon > 0.$$

Therefore, from the previous theorem, we can conclude

$$\log N(\varepsilon,\, [-1,\, 1],\, |\cdot|) \asymp \log \frac{1}{\varepsilon}, \quad \text{for all } \varepsilon > 0.$$

More generally, for the d-dim cube $[-1,\, 1]^d$, we have $\log N(\varepsilon,\, [-1,\, 1]^d,\, \|\cdot\|_\infty) \asymp d \log(1/\varepsilon)$.

Volume ratios and metric entropy

Theorem

Consider a pair of norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on $\mathbb{R}^d$, Let $\mathbb{B}_1$ and $\mathbb{B}_2$ be the corresponding unit balls. The the ε-covering number of $\mathbb{B}_1$ in the $\|\cdot\|_2$ norm satisfies

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\textit{vol}\,(\mathbb{B}_1)}{\textit{vol}\,(\mathbb{B}_2)} \leq N(\varepsilon,\, \mathbb{B},\, \|\cdot\|_2) \leq \frac{\textit{vol}\,(\frac{2}{\varepsilon}\,\mathbb{B}_1 + \mathbb{B}_2)}{\textit{vol}\,(\mathbb{B}_2)}.$$

In particular, if $\|\cdot\|_1 = \|\cdot\|_2 = \|\cdot\|$, then

$$\left(\frac{1}{\varepsilon}\right)^d \leq N(\varepsilon,\, \mathbb{B},\, \|\cdot\|) \leq \left(\frac{2}{\varepsilon} + 1\right)^d.$$

Example: smoothly parameterized functions

- Let $\mathcal{F}$ be a parameterized class of functions

$$\mathcal{F} = \{ f_\theta(\cdot) : \theta \in \Theta \}.$$

- Let $\| \cdot \|_\Theta$ be a norm on Θ and $\| \cdot \|_\mathcal{F}$ be a norm on $\mathcal{F}$.
- Suppose the mapping $\theta \mapsto f_\theta(\cdot)$ is L-Lipschitz,

$$\| f_\theta - f_{\theta'} \|_\mathcal{F} \leq L \, \| \theta - \theta' \|_\Theta.$$

  Then $N(\varepsilon, \mathcal{F}, \| \cdot \|_\mathcal{F}) \leq N(\varepsilon / L, \Theta, \| \cdot \|_\Theta)$.
- For example: $f_\theta = 1 - e^{-\theta |x|}$, $\theta \in [0, 1]$ and $x \in [0, 1]$; $\| \cdot \|_\theta = | \cdot |$ and $\| \cdot \|_\mathcal{F} = \| \cdot \|_\infty$. Then $N(\varepsilon, \mathcal{F}, \| \cdot \|_\infty) \leq \lfloor 2/\varepsilon \rfloor + 1$.
- A function class with a metric entropy that scales as $\log(1/\varepsilon)$ when $\varepsilon \to 0$ is relatively small.

Example: Lipschitz functions on the unit interval

Consider the class of Lipschitz functions

$$\mathcal{F}_L = \left\{ g : [0,1] \to \mathbb{R} \,\middle|\, g(0) = 0,\ g \text{ is } L\text{-Lipschitz} \right\}.$$

Property

The metric entropy of $\mathcal{F}_L$ w.r.t. the sup-norm scales as

$$\log N(\varepsilon, \mathcal{F}_L, \|\cdot\|_\infty) \asymp L/\varepsilon, \quad \text{as } \varepsilon \to 0.$$

More generally, for d-dimensional L-Lipschitz (w.r.t. the sup-norm) function class $\mathcal{F}_L([0,1]^d)$, then

$$\log N(\varepsilon, \mathcal{F}_L([0,1]^d), \|\cdot\|_\infty) \asymp \big(L/\varepsilon\big)^d, \quad \text{as } \varepsilon \to 0.$$

It has exponential dependence on the dimension d (curse of dimensionality).

Example: Higher-order smoothness classes

For some integer α and parameter $\gamma \in (0, 1]$, consider the class $\mathcal{F}_{\alpha,\gamma}$ of functions $f : [0, 1] \to \mathbb{R}$ such that

$$|f^{(j)}(x)| \leq C, \quad \text{for all } x \in [0, 1], j = 0, 1, \ldots, \alpha, \text{ and}$$
$$|f^{(\alpha)}(x) - f^{(\alpha)}(y)| \leq L\,|x - y|^{\gamma}, \quad \text{for all } x, y \in [0, 1].$$

Property

The metric entropy of $\mathcal{F}_{\alpha,\gamma}$ w.r.t. the sup-norm scales as

$$\log N(\varepsilon, \mathcal{F}_L, \|\cdot\|_\infty) \asymp \left(1/\varepsilon\right)^{\frac{1}{\alpha+\gamma}}, \quad \text{as } \varepsilon \to 0.$$

More generally, we can similarly define d-dimensional class $\mathcal{F}_{\alpha,\gamma}([0, 1]^d)$, and

$$\log N(\varepsilon, \mathcal{F}_L([0, 1]^d), \|\cdot\|_\infty) \asymp \left(1/\varepsilon\right)^{\frac{d}{\alpha+\gamma}}, \quad \text{as } \varepsilon \to 0.$$

Example: Infinite dimensional ellipsoids in $\ell^2(\mathbb{N})$

Given a sequence of non-negative real numbers $\mu_1 \geq \mu_2 \geq \cdots$ such that $\sum_{j=1}^{\infty} \mu_j < \infty$, consider the ellipsoid

$$\mathcal{E} = \left\{ (\theta_j)_{j=1}^{\infty} \;\middle|\; \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq 1 \right\} \subset \ell^2(\mathbb{N}).$$

More concretely, focusing on $\mu_j = j^{-2\alpha}$ for $j = 1, 2, \ldots$ and some $\alpha > 1/2$.

Property

$$\log N(\varepsilon, \mathcal{E}, \|\cdot\|_2) \asymp \left(\frac{1}{\varepsilon}\right)^{1/\alpha} \quad \text{for sufficiently small } \varepsilon > 0.$$

Canonical Rademacher and Gaussian processes

Definition

Fix a set $\mathcal{T} \subset \mathbb{R}^n$.

1. The **canonical Gaussian process** is the stochastic process $\{G_\theta : \theta \in \mathcal{T}\}$, where

$$G_\theta = \langle g, \theta \rangle = \sum_{i=1}^{n} g_i \theta_i, \quad g_i \overset{iid}{\sim} \mathcal{N}(0, 1).$$

2. The **canonical Rademacher process** is the stochastic process $\{R_\theta : \theta \in \mathcal{T}\}$, where

$$R_\theta = \langle \varepsilon, \theta \rangle = \sum_{i=1}^{n} \varepsilon_i \theta_i, \quad g_i \overset{iid}{\sim} \text{uniform over } \{-1, +1\}.$$

Canonical Rademacher and Gaussian processes

Recall the Gaussian complexity of $\mathcal{T}$ is $\mathcal{G}(\mathcal{T}) = \mathbb{E}[\sup_{\theta\in\mathcal{T}} G_\theta]$, and the Rademacher complexity of $\mathcal{T}$ is $\mathcal{R}(\mathcal{T}) = \mathbb{E}[\sup_{\theta\in\mathcal{T}} R_\theta]$.

Properties

1. (Relation) for $\mathcal{T} \subset \mathbb{R}^d$,

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

2. (Finite Lemma) $g = (g_1, \ldots, g_d)$ has sub-Gaussian components with parameters σ^2. If $\mathcal{A} \subset \mathbb{R}^d$ has finite size, then

$$\mathbb{E} \max_{a\in\mathcal{A}} \langle g,\, a\rangle \leq \sigma \max_{a\in\mathcal{A}} \|a\|_2 \sqrt{2\log|\mathcal{A}|}.$$

Proof: Left as a homework problem.