Metric Entropy

① Phisics: The degree of disorder or randomness in a system.
② Information Theory: The randomness of a distribution.

In this chapter, Metric Entropy represents a similar meaning:

size of a class of functions

Learning theory

To motivate the definition of metric entropy, we introduce some fundemtal knowledge about learning theory.

We assume that $(\mathcal{X}, \Sigma, P)$ is a probability space, and $X_1, \cdots, X_n$ are iid sampled from law P, we mark that

$$X_1, \cdots, X_n \sim (\mathcal{X}, \Sigma, P)$$

$X_1, \cdots, X_n \sim (\mathcal{X}, \Sigma, P)$ and

$$Y_i = f_0(X_i) + \varepsilon_i, \quad f_0 \in \mathcal{F},$$

where $\mathcal{F}$ is a class of functions: $\mathcal{X} \mapsto \mathbb{R}$.

If we observe $(X_i, Y_i)_{i=1,\cdots,n}$, we want to estimate the ture f_0, usually, given loss function $L: \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$,

$$f_0 := \operatorname{argmin}_{f \in \mathcal{F}} \underline{\mathbb{E} L(Y_1, f(X_1))},$$

Risk function

Example

$\mathcal{X} = \mathbb{R}^p$.

① $\mathcal{F} = \mathbb{R}^p$, $L(y_1, y_2) = (y_1 - y_2)^2$, $\operatorname{Var} \varepsilon_i = \sigma^2 < \infty \Leftrightarrow$ Least square.

② $\mathcal{F} = \{\beta \in \mathbb{R}^p : \|\beta\|_0 \le s\} \Leftrightarrow$ Best subset selection.

③ $\mathcal{F} = \{f \in \mathbb{R}^p \to \mathbb{R}, f^{(k)}$ exists and $\|f^{(k)}\|_q \le C\} \Leftrightarrow$ smoothing spline.

④ RKHS / SVM / DL.

We can define so called M-estimator of f_0 by emprical risk:

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i))$$

We need to answer some questions:

① How to ensure $\hat{f}_n \to f_0$ in some sense. (Asymptotic)

② How to ensure $\mathbb{P}(d(f_0, \hat{f}_n) > \varepsilon)$ is small enough for $\varepsilon > 0$, where d is a metric on $\mathcal{F}$. (Non-asymptotic)

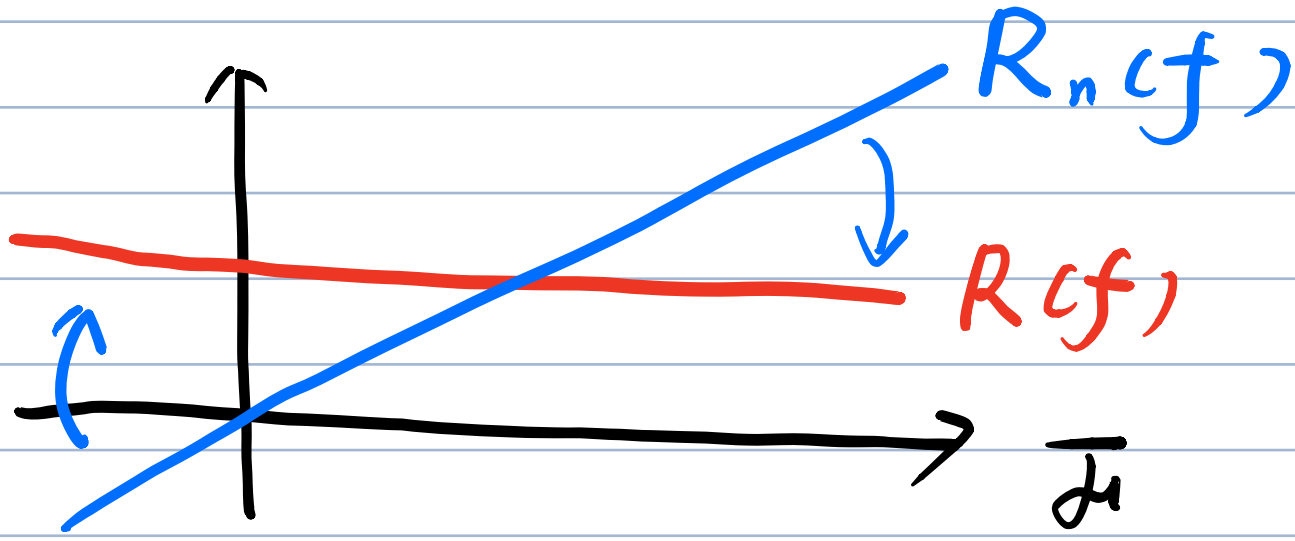
Learning Process

$$R(f) = \mathbb{E} L(Y_1, f(X_1)),$$
$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)),$$

(Formally, $PL(*, f(\cdot))$ and $P_n L(*, f(\cdot))$)

Notice that $R_n(f) \to_{a.s.} R(f)$, for fixed f, but:

$$R_n(\hat{f}_n) = \inf_{f \in \mathcal{F}} R_n(f) \not\to \inf_{f \in \mathcal{F}} R(f),$$



A key condition for

$$\inf_{f\in\mathcal{F}} R_n(f) \to \inf_{f\in\mathcal{F}} R(f) \quad (1)$$

is the "compactness" of $\mathcal{F}$. Moreover, if $|\mathcal{F}| < \infty$, (1) holds, which implies the "magnitude" of $\mathcal{F}$ seem to play an important role on the convergence of learning process.

In this chapter, we focus on the definition of the "magnitude" of a given class of functions, to evaluate the possible risk for a learning procedure.

Setting

① $X_1, \cdots, X_n \overset{iid}{\sim} (\mathcal{X}, \Sigma, P)$.

② $\mathcal{F}$ is a class of functions:

$$\mathcal{X} \mapsto \mathbb{R}.$$

③ $\mathcal{F}$ is a metric space equipped with metric d.

If $|\mathcal{F}| = \infty$, in order to quantify the size of $|\mathcal{F}|$, it's common to get the "finite version" of $\mathcal{F}$ by the covering number of $\mathcal{F}$.

Covering Number

$\{f_i\} \subset \mathcal{F}$ is said to be a δ-cover of $\mathcal{F}$ if

$$\mathcal{F} \subset \cup_i \bar{B}(f_i; \delta).$$

Mark $C_\delta(\mathcal{F})$ be the collection of all δ-cover of $\mathcal{F}$, then

$$N(\delta; \mathcal{F}, d) := \inf_{A \in C_\delta(\mathcal{F})} |A|.$$

Remark:

① If $N(\varepsilon; \mathcal{F}, d) < \infty$, $\forall \varepsilon > 0$, we call $\mathcal{F}$ is totally bounded.

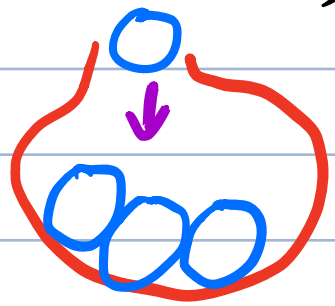
② The quantity $\log N(\varepsilon; \mathcal{F}, d)$ is called as the metric entropy of $\mathcal{F}$.

Normally, the covering number of $\mathcal{F}$ is hard to calculate, since it lacks efficient algorithms to get the minimal cardinality of $C_\varepsilon(\mathcal{F})$. Additionally, we don't need to calculate the specific form of $N(\varepsilon; \mathcal{F}, d)$, to be concise, note that

$$N(\varepsilon; \mathcal{F}, d) \uparrow, \text{ if } \varepsilon \downarrow 0,$$

we use the speed of incresing of $N(\varepsilon; \mathcal{F}, d)$ as $\varepsilon \downarrow 0$ to quantify the "size" of $\mathcal{F}$.

So we need to create a "equivalent amount" of covering number, which is easy to calculate.



Packing Number

$\{g_j\} \subset \mathcal{F}$ is said to be a δ-packing of $\mathcal{F}$ if

$$d(g_j, g_i) > \delta \text{ if } i \neq j.$$

Similarly, $P_\delta(\mathcal{F})$ is the collection of all δ-packing of $\mathcal{F}$,

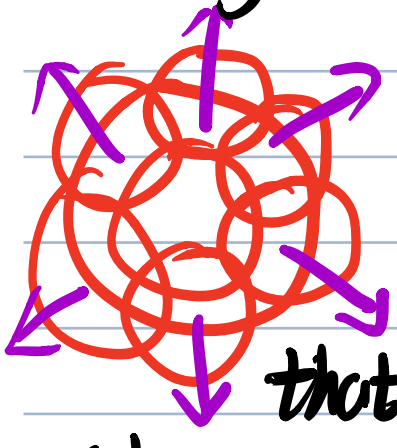
$$P(\delta; \mathcal{F}, d) := \sup_{B \in P_\delta(\mathcal{F})} |B|$$

Lemma 1 $\forall \delta > 0$, we have

$$P(2\delta; \mathcal{F}, d) \leq N(\delta; \mathcal{F}, d) \leq P(\delta; \mathcal{F}, d)$$

Pf: $\forall \{f_i\} \in P_{2\delta}(\mathcal{F}), \forall \{g_j\} \in C_\delta(\mathcal{F})$,

For $\forall i$, $\exists g_{n_i} \in \{g_j\}$ s.t. $d(f_i, g_{n_i}) \leq \delta$.



Define $\phi(f_i) = g_{n_i}$, claim that ϕ is injection, since if $\exists f_j \in \{f_i\}$ s.t. $d(f_j, g_{n_i}) \leq \delta$

$\Rightarrow d(f_j, f_i) \leq 2\delta$, contradiction.

$\Rightarrow |\{f_i\}| \leq |\{g_j\}|$.

$\Rightarrow P(2\delta; \mathcal{F}, d) \leq N(\delta; \mathcal{F}, d)$.

For $N(\delta; \mathcal{F}, d) \leq P(\delta; \mathcal{F}, d)$, Let $\{f_i\}$ be the maximal δ-packing of $\mathcal{F}$, i.e.,

$$d(f_i, f_j) > \delta, \forall i \neq j.$$

Then $\{f_i\}$ is a δ-cover of $\mathcal{F}$, since if $\exists f \in \mathcal{F}$ s.t.

$$d(f, f_i) > \delta, \forall f_i \in \{f_i\},$$

$\Rightarrow \{f_i\} \cup f \in P_\delta(\mathcal{F})$, contradiction.

$\Rightarrow N(\delta; \mathcal{F}, d) \leq P(\delta; \mathcal{F}, d)$.

□

Example 1

Lemma 2 If $\mathcal{F} = \mathbb{R}^p$ equipped with norm $\|\cdot\|$, define

$$\mathbb{B} = \{\theta \in \mathbb{R}^p : \|\theta\| \leq 1\},$$

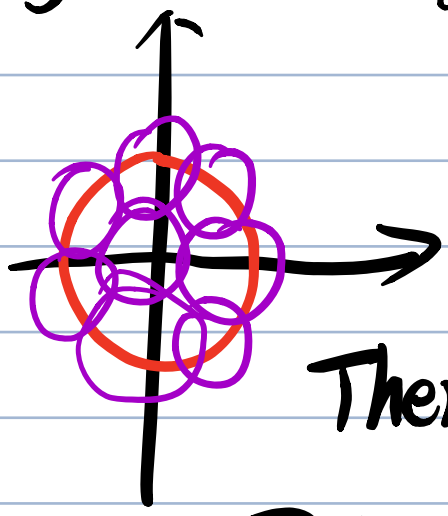
Then

$$\left(\frac{1}{\delta}\right)^p \leq N(\delta; \mathbb{B}, \|\cdot\|) \leq \left(\frac{2}{\delta} + 1\right)^p$$

Pf: Let $\{\theta_i\} \in C_\delta(B)$, i.e.,

$$B \subset \cup_i \bar{B}(\theta_i; \delta)$$

and $|\{\theta_i\}| = N(\delta; B, \|\cdot\|)$,



Then

$$\mathrm{Vol}(B) \le \mathrm{Vol}(\cup_i \bar{B}(\theta_i; \delta))$$

$$\le \sum_i \mathrm{Vol}(\bar{B}(\theta_i; \delta))$$

$$= N(\delta; B, \|\cdot\|)\, \mathrm{Vol}(\bar{B}(0; \delta))$$

$$= \delta^p N(\delta; B, \|\cdot\|)\, \mathrm{Vol}(B)$$

$$\Rightarrow \left(\frac{1}{\delta}\right)^p \le N(\delta; B, \|\cdot\|).$$

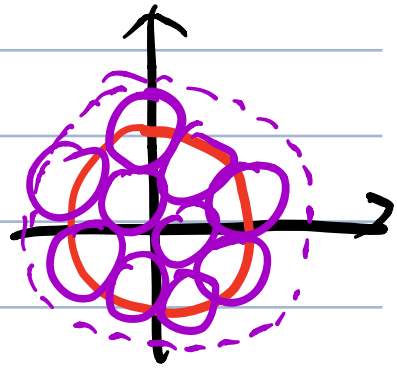
Note that $N(\delta; B, \|\cdot\|) \le P(\delta; B, \|\cdot\|)$

and $P(\delta; B, \|\cdot\|) \cdot \mathrm{Vol}(\bar{B}(0, \frac{\delta}{2}))$

$$\le \mathrm{Vol}(B + \bar{B}(0, \tfrac{\delta}{2}))$$

$$\Rightarrow P(\delta; B, \|\cdot\|)$$

$$\le (1 + \tfrac{\delta}{2})^p / (\tfrac{\delta}{2})^p = (\tfrac{2}{\delta} + 1)^p$$



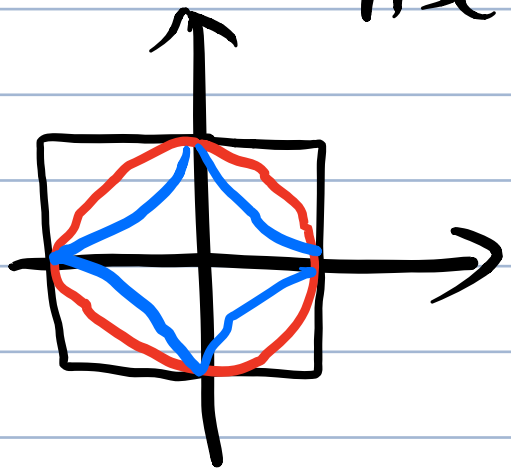
□

(Covering Numbers of unit ball

p-norm of $\mathbb{R}^n$ is defined as:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}, \quad \forall x \in \mathbb{R}^n,$$

Unit ball in $\mathbb{R}^n$: $\mathbb{B}_p = \{x \in \mathbb{R}^n : \|x\|_p \le 1\}$,



Notice that

$$N(\varepsilon; \mathbb{B}_\infty, \|\cdot\|_\infty) \le \left(\left\lceil \frac{1}{\varepsilon} \right\rceil\right)^p$$

and

$$N(\varepsilon; \mathbb{B}_\infty, \|\cdot\|_\infty) \ge P(2\varepsilon; \mathbb{B}_\infty, \|\cdot\|_\infty) \ge \left(\left\lfloor \frac{1}{\varepsilon} \right\rfloor\right)^p$$

$$\Rightarrow \quad N(\varepsilon; \mathbb{B}_\infty, \|\cdot\|_\infty) \simeq \left(\frac{1}{\varepsilon}\right)^p$$

Example 2

(Lipschitz function)

In learning theory, the continuous functions usually are the functions we want to learn, but as the Lemma in Functional Analysis:

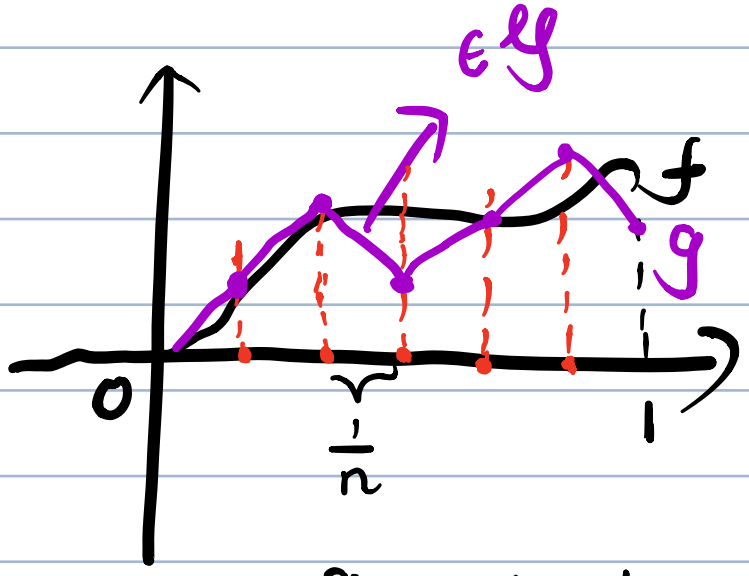
The bound set of infinite dimensional vector space is not totally bounded.

We usually need to "imagine" some regularisations for continuous functions space, for example, special convex bound set. Define:

$$\mathcal{F}_L := \{ g : [0,1] \to \mathbb{R},\ g(0) = 0 \text{ and } |g(x) - g(x')| \leq L|x - x'|, \forall x, x' \in [0,1] \},$$ where $L > 0$.

$\mathcal{F}_L$ equids with $\|\cdot\|_{sup}$:

$\|f - g\|_{sup} = \sup_{x \in [0,1]} |f(x) - g(x)|$, $f, g \in \mathcal{F}_L$. Then:



For $\forall f \in F_L$,
$\exists g \in \mathcal{G}$ s.t.
$\|f-g\|_{sup} \le \frac{L}{n}$,

and if $g_1 \neq g_2 \in \mathcal{G}$

$\|g_1 - g_2\|_{sup} \ge \frac{2L}{n}$,

Given $\epsilon > 0$, let $n = \lceil \frac{L}{\epsilon} \rceil$

$\Rightarrow |\mathcal{G}| = 2^n = 2^{\lceil \frac{L}{\epsilon} \rceil}$

$\Rightarrow \log N(\epsilon; F_L, \|\cdot\|_{sup}) \le \frac{L}{\epsilon}$.

Let $n = \lfloor \frac{L}{\epsilon} \rfloor \Rightarrow |\mathcal{G}| = 2^{\lfloor \frac{L}{\epsilon} \rfloor}$

$\Rightarrow \log N(\epsilon; F_L, \|\cdot\|_{sup}) \gtrsim \frac{L}{\epsilon}$.

Remark:

① For p-dimensional Lipschiz function, it's easy to show:

$\log N(\epsilon; F_L, \|\cdot\|_{sup}) \lesssim (\frac{L}{\epsilon})^p$,

which implies the curse of dimension.

② If $f(0)$ is not fixed but

$$|f(x)| \leq C$$

$\log N(\delta; \mathcal{F}_L, \|\cdot\|_{sup}) \simeq \frac{L}{\delta}$ still holds.