Metric Entropy

Tianwei Gao

January 18, 2025

1 Metric Entropy

In this section, we discuss about metric entropy.

1.1 Covering and Packing

Definition 1.1. *Throughout the definition, we fix a metric space (X, ρ) and suppose we are given a $\delta > 0$,*

1. *A δ-cover of (X, ρ) is a subset U of X, s.t. , for each $x \in X$, $\exists y \in U$, $\rho(x, y) \leq \delta$*
2. *The covering number $N(X, \rho, \delta)$ is smallest cardinality of a δ-cover.*
3. *A δ-packing of (X, ρ) is a subset T of X, such that for each pair of points $x, y \in U$, $x \neq y$ implies $\rho(x, y) > \delta$.*
4. *The δ-packing number $M(X, \rho, \delta)$ is the largest cardinality of a δ-packing.*

Remark 1.1. *We have the following observations:*

1. *A maximal δ-packing is a δ-covering, so that $M(X, \rho, \delta) \geq N(X, \rho, \delta)$.*
2. *Suppose $N(X, \rho, \delta) < \infty$, then $M(X, \rho, 2\delta) \leq N(X, \rho, \delta)$, otherwise, assume U is a minimal δ-covering, for each $u \in U$, denote by $B(u, \delta)$ is δ-neighborhood, for a set T with $|T| > |U|$, there exists two distinct points of T lies in one $B(u, \delta)$, these two points' distance is smaller that 2δ, thus T cannot be a 2δ-packing.*

Here we present some examples of estimation of covering number and packing number.

Example 1.1. *Consider the boolean hypercube $\mathbb{H}^d := \{0, 1\}^d$ equipped with scaled hamming metric, i.e.*

$$\rho(x, y) := \frac{1}{d} \sum_{i=1}^{d} \mathbb{I}(x_i \neq y_i).$$

Now we bound its packing number $M(X, \rho, \delta)$ for given $\delta > 0$. The idea is, δ-packing implies the $\frac{\delta}{2}$-neighborhood of each points are disjoint, thus we have

$$Vol(B(0, \frac{\delta}{2})) \times \#(\delta\text{-packing}) \leq Vol(X).$$

For each $x \in \mathbb{H}^d$, the cardinality of $B(x, \frac{\delta}{2})$ is

$$|B(x, \frac{\delta}{2})| = 2^d P(Z_d \leq \frac{\delta}{2} d)$$

where $Z_d := \sum_{i=1}^{d} X_i$ and X_i are i.i.d random variables satisfying $X_1 \sim B(\frac{1}{2})$. Note that, for a δ-packing T, the δ-neighborhood of the points of T are mutually disjoint, i.e., they are disjoint subsets of $\mathbb{H}^d$, which implies:

$$|T| \times |B(0, \frac{\delta}{2})| \leq |\mathbb{H}^d| = 2^d.$$

This shows

$$|T| \leq (P(Z_d \leq \frac{\delta}{2} d))^{-1}.$$

Exploiting the result of following lemma, we have

$$\log |T| \leq d\mathbb{D}(\frac{\delta}{2}||\frac{1}{2}) + \log(d+1)$$

Lemma 1.1 (Lower tail bound for binomial sums)**.** *Suppose X_i are i.i.d following binomial distribution with $p \in (0, \frac{1}{2}]$, for $1 \leq i \leq n$,*

$$Z_n := \sum_{i=1}^{n} X_i,$$

then

$$P(Z_n \leq \delta n) \geq \frac{1}{n+1} \exp(-n\mathbb{D}(\delta||p)).$$

Example 1.2 (Ball)**.**

Claim 1. *Consider a pair of norms in $\mathbb{R}^d$, $\|\cdot\|$ and $\|\cdot\|'$, $\mathbb{B}, \mathbb{B}'$ the corresponding unit ball of the norm, then for $\delta > 0$, we have*

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

With this claim, consider the special case where the two norm are identical, then we have for each norm $\|\cdot\|$,

$$(\frac{1}{\delta})^d \leq N(B(0,1), \|\cdot\|, \delta) \leq (1 + \frac{2}{\delta})^d$$

Now we discuss metric entropy in infinite dimensional space(function space),

Example 1.3. *Consider the space X all the $1-$Lipschitz function defined on $[0,1]$, such that $f(0) = 0$, we want to approximate its covering number.*

1. ***Lower bound*** *For a given δ, consider a $M := [\frac{1}{\delta}]$ step symmetric random walk, $\{X_t\}_{t=1}^M$. Based on a path, we construct the follwing piecewise linear function by*

$$f(x) = \begin{cases} 0, & \text{if } d = 0, \\ f(d\delta) + (x - d\delta) \cdot X_d, & \text{if } x \in [(d-1)\delta, d\delta] \\ f(M\delta), & \text{otherwise}, \end{cases},$$

   Note that for different path, their distance is at least 2δ. Hence the set S of all such functions form a 2δ-packing of X,

   thus

$$N(X, \|\cdot\|_{L^\infty}, \delta) \geq M(X, \|\cdot\|_{L^\infty}, 2\delta) \geq 2^M,$$

   i.e.

$$\log N(X, \|\cdot\|_{L^\infty}, \delta) \geq \frac{1}{\delta} \log 2.$$

2. ***Upper bound*** *We show that S also forms a 3δ-covering. For a general function g in x, we can construct a $f \in X$, such that*

$$|f(i\delta) - g(i\delta)| \leq \delta, \quad \text{for } 0 \leq i \leq M,$$

   we can choose path $\{X_i\}_{i=1}^M$ inductively: if $g((i+1)\delta) \geq f(i\delta)$, we choose $X_{i+1} = 1$, otherwise, we choose -1. We check the construction satisfies the above condition inductively, WLOG, $X_i = 1$, then

$$g((i+1)\delta) - f((i+1)\delta) \geq f(i\delta) - f((i+1)\delta) = -\delta,$$

$$g((i+1)\delta) - f((i+1)\delta) \leq g(i\delta) + \delta - f((i+1)\delta) \leq \delta.$$

   Note that

$$|f(i\delta) - g(i\delta)| \leq \delta, \quad \text{for } 0 \leq i \leq M,$$

implies that

$$|f(x) - g(x)| \leq 3\delta,$$

this shows S forms a 3δ-covering, hence

$$\log N(X, \|\cdot\|_{L^\infty}, \delta) \leq \frac{3\log 2}{\delta}.$$

3. ***Conclusion*** *From the above discussion, we see*

$$\log N(X, \|\cdot\|_{L^\infty}, \delta) \asymp \frac{1}{\delta}.$$

1.2 Gaussian and Rademacher Complexity

Now we consider sets in $\mathbb{R}^d$ and Euclidean metric exclusively.

Definition 1.2. *1. Suppose T is a set in $\mathbb{R}^d$, random variable $g \sim \mathcal{N}(0, I_d)$, then the Gaussian complexity $\mathcal{G}(T)$ is defined as*

$$\mathcal{G}(T) = \mathbb{E}_g[\sup_{\theta\in T}\langle\theta, g\rangle].$$

2. Suppose T is a set in $\mathbb{R}^d$, random variable ε follows d-dimensional Rademacher distribution, then the Rademacher complexity $\mathcal{R}(T)$ is defined as

$$\mathcal{R}(T) = \mathbb{E}_\varepsilon[\sup_{\theta\in T}\langle\theta, \varepsilon\rangle].$$

Remark 1.2. *Note that both complexity are not translation invariant.*

Lemma 1.2. *There exists positive constant C independent of d, such that*

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

Proof. 1. **Lower bound** Define the argmax function π as follows,

$$\begin{aligned}\pi : \{-1, 1\}^d &\to T\\ \varepsilon &\mapsto \arg\max_\theta\langle\theta, \varepsilon\rangle.\end{aligned}$$

Then

$$\begin{aligned}\mathcal{G}(T) &\geq \mathbb{E}_g[\langle\pi(\mathbf{sgn}(g)), g\rangle]\\ &= \sum_{i=1}^d \mathbb{E}_g[\pi_i(\mathbf{sgn}(g))g_i]\\ &= \sum_{i=1}^d \mathbb{E}_g[\pi_i(\mathbf{sgn}(g)) \cdot |g_i| \cdot (\mathbf{sgn}(g))_i]\\ &= \sum_{i=1}^d \mathbb{E}_{\mathbf{sgn}(g)}\left[\pi_i(\mathbf{sgn}(g))(\mathbf{sgn}(g))_i\mathbb{E}[|g_i|\big|\mathbf{sgn}(g)]\right]\\ &= \sqrt{\frac{2}{\pi}}\sum_{i=1}^d \mathbb{E}_{\mathbf{sgn}(g)}\left[\pi_i(\mathbf{sgn}(g))(\mathbf{sgn}(g))_i\right]\\ &= \sqrt{\frac{2}{\pi}}\mathbb{E}_{\mathbf{sgn}(g)}[\langle\pi(\mathbf{sgn}(g)), \mathbf{sgn}(g)\rangle]\\ &= \sqrt{\frac{2}{\pi}}\mathbb{E}_\varepsilon[\langle\pi(\varepsilon), \varepsilon\rangle]\\ &= \sqrt{\frac{2}{\pi}}\mathcal{R}(T).\end{aligned}$$

2. **Upper bound** Now we set π to be the argmax function for scaled G, i.e.

$$\begin{aligned}\pi : [-1,1]^d &\to T\\ g &\mapsto \arg\max_{\theta\in T}\langle\theta, g\rangle.\end{aligned}$$

Furthermore, since $\frac{x}{\|x\|_\infty}$ lies in the unit cube $[-1,1]^d$, which is the convex hull of $\mathbb{H}^d := \{-1,1\}^d$, we can define the following function $\{c_h\}_{h\in\mathbb{H}^d}$, satisfying the following condition:

$$\begin{aligned}c_h :\ [-1,1]^d &\to [0,1],\\ \sum_{h\in\mathbb{H}^d} c_h(x) &= 1,\\ \sum_{h\in\mathbb{H}^d} c_h(x)h &= x.\end{aligned}$$

Indeed, $c_h(x)$ are the coefficient of h when x is written in the form of a convex combination of h.

Now, denote by

$$s(g) := \frac{g}{\|g\|_\infty},$$

we have,

$$\begin{aligned}\mathcal{G}(T) &= \mathbb{E}[\sup_{\theta\in T}\langle\theta, g\rangle]\\ &= \mathbb{E}[\langle\pi(s(g)), s(g)\rangle\cdot\|g\|_\infty]\\ &= \mathbb{E}[\langle\pi(s(g)), \sum_{h\in\mathbb{H}^d} c_h(s(g))h\rangle\cdot\|g\|_\infty]\\ &= \mathbb{E}[\langle\pi(s(g)), \sum_{h\in\mathbb{H}^d} c_h(s(g))h\rangle]\mathbb{E}[\|g\|_\infty]\\ &= \mathbb{E}[\|g\|_\infty]\sum_{h\in\mathbb{H}^d}\mathbb{E}[c_h(s(g))\langle\pi(s(g)), h\rangle]\\ &\le \mathbb{E}[\|g\|_\infty]\sum_{h\in\mathbb{H}^d}\mathbb{E}[c_h(s(g))\langle\pi(h), h\rangle]\\ &= \mathbb{E}[\|g\|_\infty]\sum_{h\in\mathbb{H}^d}(\langle\pi(h), h\rangle)\mathbb{E}[c_h(s(g))] \quad \Big(\text{integrate on } g\Big)\\ &= 2^d\mathbb{E}[\|g\|_\infty]\mathbb{E}[c_h(s(g))]\mathcal{R}(T)\\ &= \mathbb{E}[\|g\|_\infty](\sum_{h\in\mathbb{H}^d}\mathbb{E}[c_h(s(g))])\mathcal{R}(T)\\ &= \mathbb{E}[\|g\|_\infty]\mathbb{E}[\sum_{h\in\mathbb{H}^d} c_h(s(g))]\mathcal{R}(T) \quad \Big(\sum_{h\in\mathbb{H}^d} c_h(s(g)) = 1\Big)\\ &= \mathbb{E}[\|g\|_\infty]\mathcal{R}(T)\end{aligned}$$

Since $\mathbb{E}\|g\|_\infty \approx 2\sqrt{\log d}$, we finished the proof.

□

Now we see some concrete examples,

Example 1.4. *1.* ***Complexity of*** $T = B_2^d$.

We have

$$\begin{aligned}\mathcal{G}(T) &= \mathbb{E}\|g\|_2 = \frac{\sqrt{2}\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} = \sqrt{d}(1-o(1)),\\ \mathcal{R}(T) &= \sqrt{d}.\end{aligned}$$

2. ***Complexity of*** $T = B_1^d$.

We have

$$\begin{aligned}\mathcal{G}(T) &= \mathbb{E}\|g\|_\infty = 2\sqrt{\log d}(1-o(1)),\\ \mathcal{R}(T) &= 1.\end{aligned}$$