# A NOTE ON Metric Entropy

Wei Xiong *

May 24, 2023

## 1 Introduction

We are interested in the use of metric entropy. This note is for the chapter 5 of Wainwright [2019].

## 2 Covering and Packing

We consider a metric space $(\mathrm{T}, \rho)$.

### 2.1 Definitions

**Definition 1** (Covering number)**.** *A $\delta$-cover of a set $\mathrm{T}$ w.r.t. $\rho$ is a set $\{\theta^1, \cdots, \theta^N\} \subset \mathrm{T}$ s.t. for each $\theta \in \mathrm{T}$, there exists some $i \in [N]$ s.t.*

$$\rho(\theta, \theta^i) \leq \delta.$$

*The covering number $N(\delta; \mathrm{T}, \rho)$ is the cardinality of the smallest $\delta$-cover.*

$N(\delta; \mathrm{T}, \rho)$ is called the *metric entropy*, which is non-increasing function of $\delta$.

- $T = [-1, 1], \rho(a, b) = |a - b|$:
$$N(\delta; T, \rho) \leq \frac{1}{\delta} + 1;$$
- $T = [-1, 1]^d, \rho(a, b) = |a - b|_\infty$:
$$N(\delta; T, \rho) \leq (\frac{1}{\delta} + 1)^d;$$
- $T = \{0, 1\}^d, \rho(a, b) = \frac{1}{d}\sum_{j=1}^d I(a_j \neq b_j)$:
$$2d(\frac{1}{2} - \delta)^2 \log N(\delta; T, \rho) \leq \log 2\lceil d(1 - \delta)\rceil;$$

*The Hong Kong University of Science and Technology; email: wxiongae@connect.ust.hk.

where $\delta \in (0, \frac{1}{2})$.

**Definition 2** (Packing number)**.** *A $\delta$-packing of a set* $\mathrm{T}$ *w.r.t. $\rho$ is a set $\{\theta^1, \cdots, \theta^M\} \subset \mathrm{T}$ s.t. for all distinct $i, j \in [M]$, we have*

$$\rho(\theta^i, \theta^j) > \delta.$$

*The packing number $M(\delta; \mathrm{T}, \rho)$ is the cardinality of the largest $\delta$-cover.*

## 2.2 Estimate covering number via packing number

The covering number and the packing number provide essentially the smae measure of the massiveness of a set, as summarized in the following lemma:

**Lemma 1.** *For all $\delta > 0$, we have*

$$M(2\delta; \mathrm{T}, \rho) \leq N(\delta; \mathrm{T}, \rho) \leq M(\delta; \mathrm{T}, \rho).$$

A direct application is for $[-1, 1]^d$ and $\|\cdot\|_\infty$. We can observe that in $[-1, 1]$, $\{\theta^i = -1 + 2(i-1)\delta : i = 1, 2, \cdots \lfloor \frac{1}{\delta} \rfloor + 1\}$ is a $2\delta$-packing. Therefore, we have

$$\log N\left(\delta; [0, 1]^d, \|\cdot\|_\infty\right) \asymp d \log(1/\delta).$$

## 2.3 Estimate covering number via volume ratio

Covering is defined in terms of the number of balls, each of which is of a fixed radius and hence volume. The covering number is connected to the volume, stated in the following lemma.

**Lemma 2** (Volume ratios and metric entropy)**.** *Consider a pair of norms $\|\cdot\|$ and $\|\cdot\|'$ on $\mathbb{R}^d$ and let $\mathbb{B}$ and $\mathbb{B}'$ be their corresponding unit balls. Then, the $\delta$-covering number of $\mathbb{B}$ in the $\|\cdot\|'$ satisfies*

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

We have following immediate results.

- If $\mathbb{B}' \subset \mathbb{B}$, the upper bound becomes

$$(\frac{2}{\delta} + 1)^d \mathrm{vol}(\mathbb{B});$$

- If we further take $\mathbb{B} = \mathbb{B}'$, we obtain

$$d \log \frac{1}{\delta} \leq \log N(\delta; \mathbb{B}, \|\|) \leq d \log(1 + \frac{2}{\delta});$$

- In particular, the unit ball in Euclidean norm can be covered by at most $(1 + 2/\delta)^d$ balls with radius $\delta$ in the norm $\|\cdot\|_2$.

### 2.4 Covering number of smooth functions

We consider $L$-Lipschitz functions on $[0,1]^d$, i.e,

$$|f(x) - f(y)| \leq L \|x - y\|_{\infty}, \forall x, y \in [0,1]^d.|$$

The set of all $L$-Lipschitz functions on $[0,1]^d$ is denoted as $\mathcal{F}_L([0,1]^d)$ and we have

$$\log N_{\infty}\left(\delta; \mathcal{F}_L\left([0,1]^d\right)\right) \asymp (L/\delta)^d$$

We note that the metric entropy has an exponential dependence on the dimension $d$, which is a dramatic manifestation of the curse of dimensionality.

## 3 Gaussian and Rademacher complexity

The metric entropy plays a fundamental role in understanding the behavior of stochastic processes. We consider a coolection of random variables

$$\{X_\theta : \theta \in \mathrm{T}\}.$$

In particular, we consider a set $\mathrm{T} \in \mathbb{R}^d$ and

$$\{G_\theta = \langle w, \theta \rangle : \theta \in \mathrm{T}\}$$

with $x_i \sim N(0,1)$ i.i.d., which is known as the *canonical Gaussian process* associated with T. Its expected supremum

$$\mathcal{G}(\mathrm{T}) := \mathbb{E} \sup_{\theta \in \mathrm{T}} \langle \theta, w \rangle$$

is known as the *Gaussian complexity* of T, which measures the size of T in a certain sense. Replacing the normal random variables with Rademacher random variables yields the Rademacher process:

$$\{R_\theta : \theta \in \mathrm{T}\}$$

where

$$R_\theta := \langle \varepsilon, \theta \rangle = \sum_{i=1}^{d} \varepsilon_i \theta_i, \quad \text{with } \varepsilon_i \text{ uniform over } \{-1, +1\}, \text{ i.i.d. } .$$

Its expected supremum

$$\mathcal{R}(\mathrm{T}) := \mathbb{E} \sup_{\theta \in \mathrm{T}} \langle \theta, \epsilon \rangle .$$

We have the following lemma;

**Lemma 3.** *For any set* T*, we have*

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

*We also remark that there are sets for which the Gaussian complexity is substantially larger than the Rademacher complexity.*

We provide several examples.

- The Euclidean ball of unit norm $\mathbb{B}_2^d$:

$$\mathcal{R}(\mathbb{B}_2^d) = \sqrt{d} \text{ and } \mathcal{G}(\mathbb{B}_2^d)/\sqrt{d} = 1 - o(1);$$

- $\mathbb{B}_1^d$ is smaller than $\mathbb{B}_2^d$ because:

$$\mathcal{R}(\mathbb{B}_1^d) = 1 \text{ and } \mathcal{G}(\mathbb{B}_1^d)/\sqrt{2\log d} = 1 \pm o(1);$$

- $\mathbb{B}_0^d(s) := \{\theta \in \mathbb{R}^d : \|\theta\|_0 \le s\}$ and we consider $\mathbb{S}^d(s) := \mathbb{B}_0^d(s) \cap \mathbb{B}_2^d(1) = \Big\{\theta \in \mathbb{R}^d \mid \|\theta\|_0 \le s,$ and $\|\theta\|_2 \le 1\}$:

$$\mathcal{G}\left(\mathbb{S}^d(s)\right) \precsim \sqrt{s \log \frac{ed}{s}}$$

We can also study a function class via its image, i.e.,

$$\mathcal{F}(x_1^n) = \{(f(x_1), \cdots, f(x_n)) : f \in \mathcal{F}\} \subset \mathbb{R}^n.$$

If the function class $\mathcal{F}$ is uniformly bounded by $b$, then, we have

$$\mathcal{G}\left(\frac{\mathscr{F}(x_1^n)}{n}\right) = \mathbb{E}\left[\sup_{f \in \mathscr{F}} \sum_{i=1}^n \frac{w_i}{\sqrt{n}} \frac{f(x_i)}{\sqrt{n}}\right] \le b \frac{\mathbb{E}\left[\|w\|_2\right]}{\sqrt{n}} \le b.$$

## 4 Metric entropy and sub-Gaussian processes

We aim to bound a expected suprema involving some process, which has its applications in deriving upper bounds for Rademacher complexity.

**Definition 3** (Sub-Gaussian processes)**.** *A collection of zero-mean random variables $\{X_\theta : \theta \in \mathrm{T}\}$ is a sub-Gaussian process w.r.t. a metric $\rho_X$ on* T *if*

$$\mathbb{E}\left[e^{\lambda\left(X_\theta - X_{\tilde{\theta}}\right)}\right] \le e^{\frac{\lambda^2 \rho_X^2(\theta,\tilde{\theta})}{2}}, \qquad \forall \theta, \tilde{\theta} \in \mathrm{T}, \lambda \in \mathbb{R}.$$

By the Chernoff method, we obtain

$$\mathbb{P}\left[\left|X_\theta - X_{\tilde{\theta}}\right| \ge t\right] \le 2e^{-\frac{t^2}{2\rho_X^2(\theta,\tilde{\theta})}}.$$

Given a sub-Gaussian process, we use the notaiton $N_X(\delta; \mathrm{T})$ to denote the $\delta$-covering number of T w.r.t. $\rho_X$. We start with a basic idea: by approximating T up to some accuracy $\delta$, we may replace the supremum over T by a finite maximum over the $\delta$-covering set, plus an approximation error that scales proportionally with $\delta$. We denote the diameter of T as

$$D = \sup_{\theta_1, \theta_2 \in \mathrm{T}} \rho_X(\theta_1, \theta_2).$$

**Theorem 1** (Bound by one-step discretization.)**.** *For any $\delta \in [0, D]$ s.t. $N_X(\delta, \mathrm{T}) \geq 10$, we have*

$$\mathbb{E}\left[\sup_{\theta, \widetilde{\theta} \in \mathrm{T}} \left(X_\theta - X_{\widetilde{\theta}}\right)\right] \leq 2\mathbb{E}\left[\sup_{\substack{\gamma, \gamma' \in \mathrm{T} \\ \rho_X(\gamma, \gamma') \leq \delta}} (X_\gamma - X_{\gamma'})\right] + 4\sqrt{D^2 \log N_X(\delta; \mathrm{T})}.$$

*We remark that due to $X_{\theta_0}$ is mean-zero, we have*

$$\mathbb{E}\left[\sup_{\theta \in \mathrm{T}} X_\theta\right] = \mathbb{E}\left[\sup_{\theta \in \mathrm{T}} (X_\theta - X_{\theta_0})\right] \leq \mathbb{E}\left[\sup_{\theta, \widetilde{\theta} \in \mathrm{T}} \left(X_\theta - X_{\widetilde{\theta}}\right)\right].$$

*Proof.* The idea to approximate an infinite set with error is presented in this proof. For a given $\delta > 0$ and associated covering number $N = N_X(\delta; \mathrm{T})$, we let $\{\theta^1, \cdots, \theta^N\}$ be a $\delta$-cover of T. For any $\theta \in \mathrm{T}$, we can find some $\theta^i$ with $\rho_X(\theta, \theta^i) < \delta$ and

$$\begin{aligned} X_\theta - X_{\theta^1} &= (X_\theta - X_{\theta^i}) + (X_{\theta^i} - X_{\theta^1}) \\ &\leq \sup_{\substack{\gamma, \gamma' \in \mathrm{T} \\ \rho_X(\gamma, \gamma') \leq \delta}} (X_\gamma - X_{\gamma'}) + \max_{i=1,2,\ldots,N} |X_{\theta^i} - X_{\theta^1}| \end{aligned}$$

Similarly, we have

$$X_{\theta^1} - X_{\tilde{\theta}} \leq \sup_{\substack{\gamma, \gamma' \in \mathrm{T} \\ \rho_X(\gamma, \gamma') \leq \delta}} (X_\gamma - X_{\gamma'}) + \max_{i=1,2,\ldots,N} |X_{\theta^i} - X_{\theta^1}|.$$

Summing them up gives

$$X_\theta - X_{\tilde{\theta}} \leq 2 \sup_{\substack{\gamma, \gamma' \in \mathrm{T} \\ \rho_X(\gamma, \gamma') \leq \delta}} (X_\gamma - X_{\gamma'}) + 2 \max_{i=1,2,\ldots,N} |X_{\theta^i} - X_{\theta^1}|.$$

Since $\theta$ and $\tilde{\theta}$ are arbitrary, ew conclude that

$$\sup_{\theta, \tilde{\theta} \in \mathrm{T}} (X_\theta - X_{\tilde{\theta}}) \leq 2 \sup_{\substack{\gamma, \gamma' \in \mathrm{T} \\ \rho_X(\gamma, \gamma') \leq \delta}} (X_\gamma - X_{\gamma'}) + 2 \max_{i=1,2,\ldots,N} |X_{\theta^i} - X_{\theta^1}|.$$

34 □

We can further optimize w.r.t. $\delta$ to obtain the optimal bound. For instance, we consider the

Gaussian complexity:

$$\mathcal{G}(\mathrm{T}) \leq \min_{\delta \in [0,D]} \left\{ \mathcal{G}(\widetilde{\mathrm{T}}(\delta)) + 2\sqrt{D^2 \log N_2(\delta; \mathrm{T})} \right\}$$

where $N_2$ means Euclidean norm and

$$\widetilde{\mathrm{T}}(\delta) := \left\{ \gamma - \gamma' \mid \gamma, \gamma' \in \mathrm{T}, \|\gamma - \gamma'\|_2 \leq \delta \right\}.$$

The $\mathcal{G}(\widetilde{\mathrm{T}}$ is referred to as the *localized Gaussian complexity.* An analogous upper bound holds for the Rademacher complexity in terms of a localized Rademacher complexity. To be specific, we use Cauchy-Schwarz inequality to obtain

$$\mathcal{G}(\widetilde{\mathrm{T}}(\delta)) = \mathbb{E}\left[ \sup_{\theta \in \widetilde{\mathrm{T}}(\delta)} \langle \theta, w \rangle \right] \leq \delta \mathbb{E}\left[ \|w\|_2 \right] \leq \delta\sqrt{d}$$

which leads to

$$\mathcal{G}(\mathrm{T}) \leq \min_{\delta \in [0,D]} \left\{ \delta\sqrt{d} + 2\sqrt{D^2 \log N_2(\delta; \mathrm{T})} \right\}.‘ \tag{4.1}$$

We provide several examples here. In particular, we will consider the image of a function class so it is useful to know the following relations among metric entropies:

**Lemma 4.** *Let* $\|f-g\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^n \left(f(x_i) - g(x_i)\right)^2}$*. Then, we have*

$$\log N_2\left(\delta; \mathcal{F}(x_1^n)/\sqrt{n}\right) \leq \log N_\infty\left(\delta; \mathcal{F}(x_1^n)\right) \leq \log N\left(\delta; \mathcal{F}, \|\cdot\|_\infty\right).$$

*Proof.* This is because

$$\|f-g\|_n \leq \max_{i=1,\ldots,n} |f(x_i) - g(x_i)| \leq \|f-g\|_\infty.$$

Note that we are concerning the empirical sets (images) for the first two terms. □

We have

- We know that $\mathcal{G}(\mathbb{B}_2^d) = \sqrt{d}(1 - o(1))$. With the bound of entropy and the above result, we have
$$\mathcal{G}\left(\mathbb{B}_2^d\right) \leq \sqrt{d}\left\{\frac{1}{2} + 2\sqrt{2\log 5}\right\};$$
- $\mathcal{F}_L$: the set of L-Lipschitz functions on $[0,1]$:
$$\mathcal{G}\left(\mathscr{F}_L\left(x_1^n\right)/n\right) \leq \frac{1}{\sqrt{n}} \inf_{\delta \in (0,\delta_0)} \left\{ \delta\sqrt{n} + 3\sqrt{\frac{cL}{\delta}} \right\} \precsim n^{-1/3}$$
by setting $\delta = n^{-1/3}$.

## 40 5 Chaining and Dudley's entropy integral

The method used in last section only employs one-step discretization. The idea of chaining method is to decompose the supremum into a sum of finite maxima over sets that are successively refined so as to obtain tighter bounds. Let $\{X_\theta : \theta \in \mathrm{T}\}$ be a zero-mean sub-Gaussian process w.r.t $\rho_X$ and let $D = \sup_{\theta,\tilde{\theta}} \rho_X(\theta, \tilde{\theta})$. The $\delta$-truncated Dudley's integral is given by

$$\mathcal{J}(\delta; D) := \int_\delta^D \sqrt{\log N_X(u; \mathrm{T})} du$$

41 We then have

**Theorem 2** (Bound via Dudley's entropy integral). *For any $\delta \in [0, D]$, we have*

$$\mathbb{E}\left[\sup_{\theta,\vec{\theta} \in \mathbb{T}} \left(X_\theta - X_{\widetilde{\theta}}\right)\right] \leq 2\mathbb{E}\left[\sup_{\gamma,\gamma' \in \mathbb{T}} \left(X_\gamma - X_{\gamma'}\right)\right] + 32\mathcal{J}(\delta/4; D).$$

We can use it to derive bound for Rademacher complexity. Let $S = X_1^n$ and let $R_S(\mathcal{F})$ be the empirical Rademacher complexity. Then,

$$R_S(\mathcal{F}) \leq 4\alpha + 12 \int_\alpha^\infty \sqrt{\frac{\log N(\epsilon, \mathcal{F}, L_2(P_n))}{n}} d\epsilon, \tag{5.1}$$

where $\alpha \geq 0$ is arbitrary. If we further assume that $f$ is bounded in $[-1, 1]$, then we have

$$R_S(\mathcal{F}) \leq \inf_{\epsilon > 0} \left(\epsilon + \sqrt{\frac{2\log(N(\epsilon, \mathcal{F}, L_2(P_n))}{n}}\right),$$

42 where $L_2(P_n)(f, f') := \sqrt{\frac{1}{n}\sum_{i=1}^n (f(x_i) - f'(x_i))^2}$.

# References

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.