36-710: Advanced Statistical Theory II **Fall 2019**

Lecture 4: September 9

Lecturer: Matey Neykov *Scribes: Nil-Jana Akpinar*

4.1 Metric entropy and its uses (Chapter 5 in [W])

4.1.1 Covering and packing

Definition 4.1 (metric space) *A metric space is a tuple (T, ρ) where T is a non-empty set and $\rho : T \times T \to \mathbb{R}$ is a function such that for all $\theta, \tilde{\theta}, \hat{\theta} \in T$ the following conditions are satisfied:*

(i) (Non-negativity) $\rho(\theta, \tilde{\theta}) \geq 0$ with equality iff $\theta = \tilde{\theta}$.

(ii) (Symmetry) $\rho(\theta, \tilde{\theta}) = \rho(\tilde{\theta}, \theta)$.

(iii) (Triangle inequality) $\rho(\theta, \tilde{\theta}) \leq \rho(\theta, \hat{\theta}) + \rho(\hat{\theta}, \tilde{\theta})$.

Familiar examples include:

- Euclidean metric on $\mathbb{R}^d$: $\rho(\theta, \tilde{\theta}) = \left\|\theta - \tilde{\theta}\right\|_2$.
- Rescaled Hamming metric on discrete cube $\{0,1\}^d$: $\rho(\theta, \tilde{\theta}) = \frac{1}{d}\sum_{i=1}^d \mathbb{1}(\theta_i \neq \tilde{\theta}_i)$.
- Function space $\mathcal{L}^2(\mu, [0,1])$ with metric $\|f - g\|_2 = \left(\int_0^1 (f(x) - g(x))^2 d\mu(x)\right)^{1/2}$.
- Function space $\mathcal{C}([0,1])$ with metric $\|f - g\|_\infty = \sup_x |f(x) - g(x)|$.

Definition 4.2 (Covering number) *A δ-cover of a set T wrt to a metric ρ is a set $\{\theta^1, \theta^2, \dots, \theta^N\} \subseteq T$ such that for all $\theta \in T$ there exists an $i \in [N]$ such that $\rho(\theta, \theta^i) \leq \delta$. The δ-covering number $\mathcal{N}(\delta, T, \rho)$ is defined as the minimal cardinality of any δ-cover. We will assume that (T, p) is totally bounded which ensures that the covering number is finite for all δ.*

It is obvious that $\mathcal{N}(\delta', T, \rho) \leq \mathcal{N}(\delta, T, \rho)$ if $\delta \leq \delta'$.

In the situation of Theorem 4.2, we call $\log \mathcal{N}(\delta, T, \rho)$ *metric entropy.*

Definition 4.3 (Packing number) *A δ-packing number of a set T wrt ρ is a set $\{\theta^1, \theta^2, \dots, \theta^M\} \subseteq T$ such that $\rho(\theta^i, \theta^j) > \delta$ for all $i \neq j \in [M]$. The maximum of any δ-packing is called packing number and we write $\mathcal{M}(\delta, T, \rho)$.*

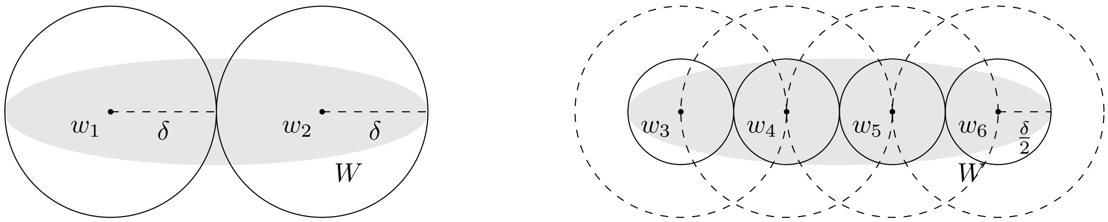

Figure 4.1: Visualization of δ-cover and δ-packing of $W \subseteq T$ in the metric space $(T, \rho) = (\mathbb{R}^2, \|\cdot\|_2)$. The set $P_1 = \{w_1, w_2\}$ is a maximum 2ε-packing of W (left). The set $P_2 = \{w_3, w_4, w_5, w_6\}$ is a maximum ε-packing of W and an ε-cover (right).

Covering and packing number are closely related as we can see in the next Lemma. An example is depicted in Figure 4.1.

Lemma 4.4 (Lemma 5.5 in [W]) *For $\delta > 0$, the packing and covering numbers satisfy*

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

Before stating the next Lemma, we define the *Minkowski sum* for two sets A and B by $A+B := \{a + b : a \in A, b \in B\}$ and $\{\alpha A\} := \{\alpha a : a \in A\}$.

Lemma 4.5 (Volume ratio bounds and metric entropy, Lemma 5.7 in [W]) *Let $\|\cdot\|, \|\cdot\|'$ a pair of norms and let B and B' be the corresponding unit balls in $\mathbb{R}^d$. Then,*

$$\left(\frac{1}{\delta}\right)^d \frac{Vol(B)}{Vol(B')} \leq \mathcal{N}(\delta, B, \|\cdot\|') \overset{(*)}{\leq} \frac{Vol((2/\delta)B + B')}{Vol(B')}.$$

If $B \subseteq B'$, $(*)$ can be simplified to $(2/\delta + 1)^d \frac{Vol(B)}{Vol(B')}$ since $Vol(\alpha S) = \alpha^d\, Vol(S)$.

Proof: We take a δ-cover $B = \left\{\theta^1, \theta^2, \dots, \theta^N\right\}$ in $\|\cdot\|'$. Then, $B \subseteq \bigcup_{i=1}^N \left\{\theta^i + \delta B'\right\}$ and $\mathrm{Vol}(B) \leq N\mathrm{Vol}(\delta B')$. This gives us the first inequality. For the second inequality, we note that the balls $\left\{\theta^i + (\delta/2)B'\right\}$ are dosjoint and belong to the set $B + (\delta/2)B'$. It follows that $M\mathrm{Vol}((\delta/2)B') \leq \mathrm{Vol}(B + (\delta/2)B')$ and thus

$$M \leq \frac{\mathrm{Vol}(B + (\delta/2)B')}{\mathrm{Vol}((\delta/2)B')} = \frac{\mathrm{Vol}((2/\delta)B + B')}{\mathrm{Vol}(B')}.$$

■

To get some intuition, we take $B = B'$ and $\|\cdot\| = \|\cdot\|'$. Then,

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

Choosing the sup norm $\|\cdot\|_\infty$ (and thus $B_\infty^d = [-1, 1]^d$), we receive

$$\log(\mathcal{N}(\delta, B_\infty^d), \|\cdot\|_\infty) \asymp d \log\left(\frac{1}{\delta}\right).$$

Example 4.6 (Lipschitz functions on the unit interval) *Consider a class of Lipschitz functions $\mathcal{F}_L = \{g : [0,1] \to \mathbb{R} : g(0) = 0, |g(x) - g(y)| \leq L\,|x - y|\, \forall x, y \in [0,1]\}$ for some $L > 0$. Then, we can prove that*

$$\log(\mathcal{N}(\delta, \mathcal{F}_L, \|\cdot\|_\infty)) \asymp \frac{L}{\delta}.$$

A proof of this Example is given in [W].

4.1.2 Gaussian and Rademacher complexities

Given a set $T \subseteq \mathbb{R}^d$, the family $\{G_\theta : \theta \in T\}$ with

$$G_\theta := \langle w, \theta \rangle = \sum_{i=1}^{d} w_i \theta_i \text{ with } w_i \overset{iid}{\sim} \mathcal{N}(0,1)$$

defines a stochastic process known as the *canonical Gaussian process associated with the set* T. The quantity $\mathcal{G}(T) = \mathbb{E}[\sup_{\theta \in T} \langle w, \theta, \rangle]$ is referred to as *Gaussian complexity* or *Gaussian width of* T.

When we replace the w_i with Rademacher RVs $\varepsilon_i \sim U(\{\pm 1\})$, we get $R_\theta = \langle \varepsilon, \theta \rangle = \sum_{i=1}^{d} \varepsilon_i \theta_i$, and the quantity $\mathcal{R}(T) = \mathbb{E}[\sup_{\theta \in T} \langle \theta, \varepsilon \rangle]$ is referred to as *Rademacher complexity.*

One can show that

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

is always true. However, there are cases in which $\mathcal{G}(T)$ can be much larger.

Example 4.7 (Complexity of $B_2^d = \{\theta : \|\theta\|_2 \leq 1\}$) *We see that*

$$\mathcal{R}(B_2^d) = \mathbb{E}[\sup_{\theta \in B_2^d} \langle \varepsilon, \theta \rangle] = \mathbb{E}[\|\cdot\|_2] = \sqrt{d},$$

where the second equality follows from Cauchy-Schwarz. For the Gaussian complexity, we use Jensen's inequality and receive

$$\mathcal{G}(B_2^d) = \mathbb{E}[\|\cdot\|_2] \leq \sqrt{\mathbb{E}[\|w\|_2^2]} = \sqrt{d},$$

which shows that $\mathcal{R}(B_2^d) \geq \mathcal{G}(B_2^d)$. $\mathcal{G}(B_2^d) = \sqrt{d}(1 - o(1))$.

References

[W] M. Wainwright, "High-dimensional statistics: A non-asymptotic viewpoint", 2019.