36-755: Advanced Statistical Theory I **Fall 2016**

Lecture 6: September 19

Lecturer: Alessandro Rinaldo *Scribe: YJ Choe*

6.1 Metric entropy and its uses

Let $(\mathcal{X}, d)$ be a metric space. We gave some examples of metric spaces, including $(\mathbb{R}^d, \|\cdot\|_p)$, the d-dimensional real space with the ℓ^p-norm, and $L^p([0,1], \mu)$ (the L^p function space on $[0,1]$ with measure μ) for $p \geq 1$.

We are interested in measuring how "big" these spaces are.

6.1.1 Covering numbers and metric entropy

Definition 6.1 (Covering numbers) *Let $\delta \geq 0$. A δ-**covering** or δ-**net** of $(\mathcal{X}, d)$ is any set*

$$\{\theta_1, \ldots, \theta_N\} \subseteq \mathcal{X}$$

where $N = N(\delta)$, such that for any $\theta \in \mathcal{X}$, there exists $i \in [N]$ such that

$$d(\theta, \theta_i) \leq \delta$$

*The δ-**covering number** of $(\mathcal{X}, d)$, denoted as $N(\delta, \mathcal{X}, d)$, is the size of a smallest δ-covering.*

There are several remarks:

1. For any $(\mathcal{X}, d)$, its δ-covering number is unique, but there can be several δ-coverings of that size.
2. Let $B(\theta_i, d) = \{\theta \in \mathcal{X} : d(\theta, \theta_i) \leq \delta\}$. Then

$$\mathcal{X} \subseteq \bigcup_{i=1}^{N(\delta,\mathcal{X},d)} B(\theta_i, d)$$

3. We will only consider metric spaces $(\mathcal{X}, d)$ that are *totally bounded*, i.e.,

$$N(\delta, \mathcal{X}, d) < \infty$$

   for any $\delta > 0$. Note that $\text{diam}(\mathcal{X}) = \sup_{\theta,\theta'} d(\theta, \theta') < \infty$ in such case.

4. In general, $N(\delta, \mathcal{X}, d)$ decreases as δ increases and diverges to ∞ as $\delta \to 0$.

Example. Let $\mathcal{X} = [-1, 1]$ and $d(x, y) = |x - y|$ for $x, y \in \mathcal{X}$. Then,

$$N(\delta, \mathcal{X}, d) \leq \frac{1}{\delta} + 1 \leq \frac{C}{\delta}$$

for some $C > 0$. If $\mathcal{X} = [-1, 1]^p$, then

$$N(\delta, \mathcal{X}, d) \leq \frac{C}{\delta^p}$$

Definition 6.2 (Metric entropy) *The* **metric entropy** *of* $(\mathcal{X}, d)$ *is defined as*

$$\log N(\delta, \mathcal{X}, d)$$

Typically, for bounded subsets of $\mathbb{R}^p$ with $\|\cdot\|$, or any of its equivalent norms, the metric entropy scales by

$$C \cdot p \log\left(\frac{1}{\delta}\right)$$

In general, bounded subsets of $\mathbb{R}^p$ are considered as "small" spaces.

For non-Euclidean spaces, e.g. function spaces, the metric entropy scales differently. We consider these as "large" spaces.

Example. Let $\mathcal{F} = \{f : [0, 1] \to \mathbb{R} \mid f \text{ is } L\text{-Lipschitz}\}$. Then,

$$\log N(\delta, \mathcal{F}, d) \preceq \frac{L}{\delta}$$

where $\preceq$ denotes less than equal up to positive constants. The bound generalizes to L-Lipschitz functions on $[0, 1]^p$ by

$$\log N(\delta, \mathcal{F}, d) \preceq \left(\frac{L}{\delta}\right)^p$$

Further notions in the book can be useful depending on the area of interest.

6.1.2 Packing numbers

Definition 6.3 (Packing numbers) *A* δ**-packing** *of* $(\mathcal{X}, d)$ *is any set*

$$\{\theta_1, \ldots, \theta_M\} \subseteq \mathcal{X}$$

where $M = M(\delta)$*, such that*

$$d(\theta_i, \theta_j) > \delta$$

for all $i \neq j$*.*

The δ**-packing number** *of* $(\mathcal{X}, d)$*, denoted as* $M(\delta, \mathcal{X}, d)$*, is the size of a largest* δ*-packing set.*

Again, the δ-packing number may be unique while the δ-packing set that achieves the number is not.

Sometimes we would prefer using covering numbers, while sometimes we would prefer using packing numbers. Figure 6.1 shows an example of an ε-covering and an ε-packing.

The following is a classic lemma on the relationship between covering and packing numbers.

Lemma 6.4 *For any* $\delta > 0$*,*

$$M(2\delta, \mathcal{X}, d) \leq N(\delta, \mathcal{X}, d) \leq M(\delta, \mathcal{X}, d)$$

Proof: Homework. ■

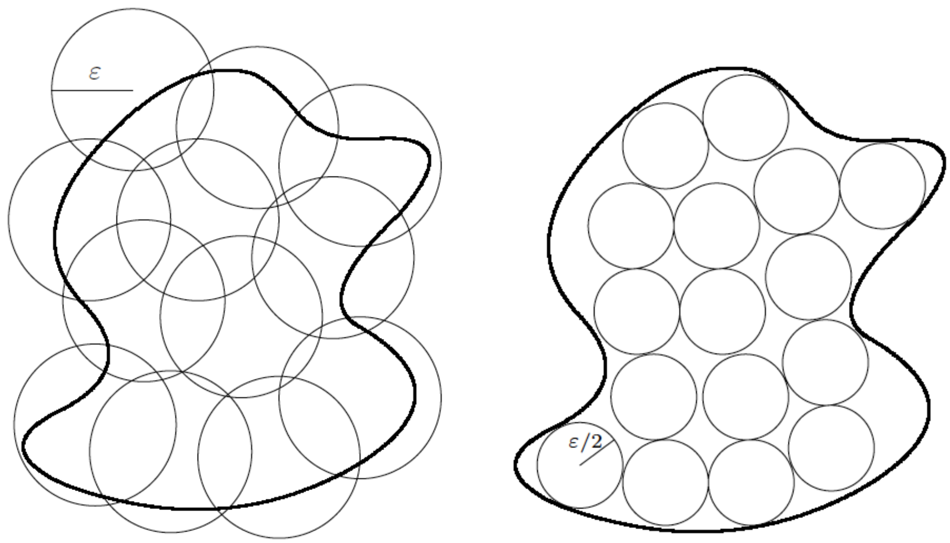

Figure 6.1: A comparison of an ε-covering (left) and an ε-packing (right). Figures from [GKKW06].

6.1.3 Volumetric ratios and covering numbers

Proposition 6.5 *Let $\|\cdot\|$ and $\|\cdot\|'$ be two norms on $\mathbb{R}^p$ (e.g. $\|\cdot\|_1$ and $\|\cdot\|_2$). Let B_p and B'_p be the corresponding unit balls.*[1]

Then,

$$\left(\frac{1}{\delta}\right)^p \frac{\mathrm{Vol}\,(B_p)}{\mathrm{Vol}\,(B'_p)} \leq N(\delta, B_p, \|\cdot\|') \leq \frac{\mathrm{Vol}\left(\frac{2}{\delta}B_p + B'_p\right)}{\mathrm{Vol}\left(B'_p\right)}$$

where, for $\alpha, \beta > 0$, $\alpha B_p = \{\alpha x : x \in B_p\}$, and $\beta B_p + B'_p = \{\beta x + y : x \in B_p,\ y \in B'_p\}$.

Proof: First note that, by homework,

$$\mathrm{Vol}\,(\delta B_p) = \delta^p \mathrm{Vol}\,(B_p)$$

for any $\delta > 0$. Also, if $\{x_1, \ldots, x_N\}$ is a δ-covering of B_p in $\|\cdot\|'$, then

$$B_p \subseteq \bigcup_{i=1}^{N} \{x_i + \delta B'_p\}$$

where $\left\{x_i + \delta B'_p\right\} = \{x : \|x - x_i\| \leq \delta\}$. Together, we get

$$\mathrm{Vol}\,(B_p) \leq N \mathrm{Vol}\left(\delta B'_p\right) \leq N\delta^p \mathrm{Vol}\left(B'_p\right)$$

Note that we assume the norm is equivalent to the L^p norm, so that we have invariance of volumes. This gives us the lower bound

$$N(\delta, B_p, \|\cdot\|') \geq \frac{\mathrm{Vol}\,(B_p)}{\mathrm{Vol}\left(B'_p\right)} \cdot \frac{1}{\delta^p}$$

[1] See previous lecture note for examples of norm balls.

To get the upper bound, let $\{y_i, \ldots, y_M\}$ be a maximal δ-packing of B_p in $\|\cdot\|'$. Then, this set is also a δ-covering of B_p in $\|\cdot\|'$, because otherwise we can find another point that will contradict the maximality of the δ-packing set.

The $\|\cdot\|'$-balls $\left\{y_i + \frac{\delta}{2}B'_p\right\}_{i=1}^{M}$ are disjoint by the maximality of the δ-packing set. Thus,

$$\bigcup_{i=1}^{M}\left\{y_i + \frac{\delta}{2}B'_p\right\} \subseteq B_p + \frac{\delta}{2}B'_p$$

Taking volumes we get

$$M\left(\frac{\delta}{2}\right)^2 \mathrm{Vol}\left(B'_p\right) \leq \left(\frac{\delta}{2}\right)^2 \mathrm{Vol}\left(\left(\frac{2}{\delta}B_p + B'_p\right)\right)$$

Note that the union simply becomes a product on the left-hand side, because the balls are disjoint.

Thus,

$$M(\delta, B_p, \|\cdot\|') \leq \frac{\mathrm{Vol}\left(\frac{2}{\delta}B_p + B'_p\right)}{\mathrm{Vol}\left(B'_p\right)}$$

Since the δ-covering number is bounded below by the δ-packing number, we have the upper bound as well. ■

In our applications, we can simply take $\|\cdot\| = \|\cdot\|'$ to conclude that

$$p\log\left(\frac{1}{\delta}\right) \leq \log N(\delta, B_p, \|\cdot\|) \leq p\log\left(1 + \frac{2}{\delta}\right) \leq p\log\left(\frac{3}{\delta}\right)$$

Note once again that this result holds for *any* norm in $\mathbb{R}^d$, including the Euclidean norm.

6.1.4 Discretization

Covering and packing numbers can be used to "discretize" a supremum over an infinite space into a maximum over a finite number of covering or packing sets. We can then give a bound on this maximum, as done in e.g. Theorem 6.7 with sub-Gaussian random vectors.

Definition 6.6 (Sub-Gaussian random vectors.) *A random vector $X \in \mathbb{R}^d$ with $\mathbb{E}[X] = 0$ is* **sub-Gaussian with parameter** σ^2*, denoted as $X \in SG_d(\sigma^2)$, if*

$$v^T X \in SG(\sigma^2)$$

for all $v \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : \|v\| = 1\}$ is the d-dimensional unit sphere.

Theorem 6.7 *Let $X \in SG_d(\sigma^2)$, and let B_d be the unit ball in $(\mathbb{R}^d, \|\cdot\|_2)$. Then,*

$$\mathbb{E}\left[\max_{\theta\in B_d} \theta^T X\right] = \mathbb{E}\left[\max_{\theta\in B_d} \left|\theta^T X\right|\right] \leq 4\sigma\sqrt{d}$$

In other words, for $\delta \in (0, 1)$,

$$\max_{\theta\in B_d} \theta^T X \leq 4\sigma\sqrt{d} + \sqrt{2\sigma\log\left(\frac{1}{d}\right)}$$

with probability $1 - \delta$.

Proof: Let $\mathcal{N}_{1/2}$ be a $\frac{1}{2}$-covering of B_d in $\|\cdot\|_2$. Then,

$$\left|\mathcal{N}_{1/2}\right| \le 5^d$$

Next, for any $\theta \in B_d$, there exists $z = z(\theta) \in \mathcal{N}_{1/2}$ such that

$$\theta = z + x$$

for some $x \in \mathbb{R}^d$ such that $\|x\| \le \frac{1}{2}$. Thus,

$$\max_{\theta \in B_d} \theta^T X \le \max_{z \in \mathcal{N}_{1/2}} z^T X + \max_{x \in \frac{1}{2} B_d} x^T X$$

Now, notice that $\max_{x \in \frac{1}{2} B_d} x^T X = \frac{1}{2} \max_{\theta \in B_d} \theta^T X$. This implies that

$$\max_{\theta \in B_d} \theta^T X \le 2 \max_{z \in \mathcal{N}_{1/2}} z^T X$$

This holds almost everywhere. Taking expectations, we get

$$\begin{aligned}
\mathbb{E}\left[\max_{\theta \in B_d} \theta^T X\right] &\le 2\,\mathbb{E}\left[\max_{z \in \mathcal{N}_{1/2}} z^T X\right] \\
&\le 2\sigma\sqrt{2\log\left|\mathcal{N}_{1/2}\right|} \\
&\le 2\sigma\sqrt{2d\log 5} \\
&\le 4\sigma\sqrt{d}
\end{aligned}$$

where we used Lemma 6.4 for the second inequality.

For the second claim, we use the union bound (second inequality below). For any $t > 0$,

$$\begin{aligned}
\mathbb{P}\left(\max_{\theta \in B_d} \theta^T X \ge t\right) &\le \mathbb{P}\left(2 \max_{z \in \mathcal{N}_{1/2}} z^T X \ge t\right) \\
&\le \sum_{z \in \mathcal{N}_{1/2}} \mathbb{P}\left(z^T X \ge \frac{t}{2}\right) \\
&\le \left|\mathcal{N}_{1/2}\right| \exp\left\{-\frac{t^2}{8\sigma^2}\right\} \\
&\le 5^d \exp\left\{-\frac{t^2}{8\sigma^2}\right\}
\end{aligned}$$

Find t such that the expression is bounded by δ:

$$t = \sigma\sqrt{8d\log 5} + 2\sigma\sqrt{2\log(1/\delta)}$$

■

6.2 Covariance estimation

Using these techniques, we will show various bounds on estimating the covariance matrix of a random vector. First, recall the following result we covered in homework 1.

Theorem 6.8 (Lemma 12, [Yuan10]; Lemma 1, [RWRY11]) *Let $(X_1, \ldots, X_d) \in \mathbb{R}^d$ be a zero-mean random vector with covariance Σ such that*

$$\frac{X_i}{\sqrt{\Sigma_{ii}}} \in SG(\sigma^2)$$

for $i = 1, \ldots, d$. Let $\hat{\Sigma}$ be the empirical covariance matrix. Then, for any $t > 0$,

$$\max_{i,j} \left|\hat{\Sigma}_{ij} - \Sigma_{ij}\right| \preceq \sqrt{\frac{t + \log d}{n}}$$

with probability at least $1 - e^{-t}$.

Note that d can be larger than n, and that the empirical covariance matrix need not be positive definite, as long as d is a polynomial in n.

We first review some basic notions in matrix algebra. For $A \in \mathbb{R}^{m \times n}$ with $\text{rank}(A) = r \leq \min\{m, n\}$, the **singular value decomposition (SVD)** of A is given by

$$A = UDV^T$$

where $D = \text{diag}(\sigma_1, \ldots, \sigma_r)$, $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are the singular values, and $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$ has r orthonormal columns.

Note that, for $j = 1, \ldots, r$,

$$AA^T u_j = \sigma_j^2 u_j$$

where $u_j \in \mathbb{S}^{m-1}$ is the jth column of U, and

$$A^T A v_j = \sigma_j^2 v_j$$

where $v_j \in \mathbb{S}^{n-1}$ is the jth column of V.

The largest singular value can also be characterized as the operator norm:

$$\sigma_{\max}(A) = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{x \in \mathbb{S}^{m-1}, y \in \mathbb{S}^{n-1}} \left|x^T A y\right|$$

If $A \in \mathbb{R}^{n \times n}$ is symmetric and positive definite, then the singular values are the square root of the eigenvalues.

In the next lecture, we will give a bound on the distance between $\hat{\Sigma}$ and Σ in the operator norm.

References

[GKKW06] L. Györfi, M. Kohler, A. Krzyzak and H. Walk, "A distribution-free theory of non-parametric regression," *Springer Science & Business Media*, 2006.

[Yuan10] M. Yuan, "High dimensional inverse covariance matrix estimation via linear programming," *Journal of Machine Learning Research 11*, 2010, pp. 2261–2286.

[RWRY11] P. Ravikumar, M. Wainwright, G. Raskutti and B. Yu, "High-dimensional covariance estimation by minimizing ℓ^1-penalized log-determinant divergence," *Electronic Journal of Statistics 5*, 2011, pp. 935–980.