ECE598: Information-theoretic methods in high-dimensional statistics	Spring 2016
Lecture 14: Packing, covering, and consequences on minimax risk	
Lecturer: Yihong Wu	*Scribe: Yingxiang Yang, Mar 10, 2016 [Ed. Mar 12]*

Last lecture, we lower bounded $\min_{\|\theta-\hat{\theta}\|} I(\theta;\hat{\theta})$ using Shannon lower bound, and we saw that for the p dimensional n sample GLM,

$$R^*(\mathbb{R}^p) \gtrsim \frac{1}{n\mathrm{vol}^{\frac{2}{p}}(B_{\|\cdot\|})}$$

with respect to a loss $\ell(\theta,\hat{\theta}) = \|\theta-\hat{\theta}\|^2$ and an arbitrary norm $\|\cdot\|$.

To understand why some sort of volume shows up, we further extend the lower bound obtained using Fano's method. We first introduce the concept of **packing**, **covering**, relate them to the notion of volume, and then plug them into the lower bound obtained using the Fano's inequality. When applied to GLM, this alternative method gives the same dependence on the dimension and the sample size for ℓ_q norms with $q < \infty$.

14.1 Covering and Packing

Definition 14.1 (ϵ-covering)**.** Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \subset V$. $\{V_1, ..., V_N\}$ is an ϵ-covering of Θ if $\Theta \subset \cup_{i=1}^N B(V_i, \epsilon)$, or equivalently, $\forall \theta \in \Theta$, $\exists i$ such that $\|\theta - V_i\| \le \epsilon$.

Definition 14.2 (ϵ-packing)**.** Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \subset V$. $\{\theta_1, ..., \theta_M\}$ is an ϵ-packing of Θ if $\min_{i\neq j} \|\theta_i - \theta_j\| > \epsilon$ (notice the inequality is strict), or equivalently $\cap_{i=1}^M B(\theta_i, \epsilon/2) = \emptyset$.

Upon defining ϵ-covering and ϵ-packing, one naturally asks what is the minimal number of ϵ-balls one needs in order to cover Θ, and what is the maximal number of $\epsilon/2$-balls one can pack in Θ. Those numbers are defined as covering and packing numbers.

Definition 14.3 (Covering number)**.** $N(\Theta, \|\cdot\|, \epsilon) := \min\{n : \exists\epsilon\text{-covering over } \Theta \text{ of size } n\}$.

Definition 14.4 (Packing number)**.** $M(\Theta, \|\cdot\|, \epsilon) := \max\{m : \exists\epsilon\text{-packing of } \Theta \text{ of size } m\}$.

Remark 14.1. Some basic remarks.

- $M(\Theta, \|\cdot\|, \epsilon)$ and $N(\Theta, \|\cdot\|, \epsilon)$ are often abbreviated as $M(\epsilon)$, $N(\epsilon)$.
- For ϵ-covering, the balls need not be disjoint.
- $N(\Theta, \|\cdot\|, \epsilon)$ is a decreasing function of ϵ when the norm and Θ are fixed. That is, if $\epsilon_0 < \epsilon_1$, and $\{V_1, ..., V_N\}$ is an ϵ-covering of Θ, then $\Theta \subset \cup_{i=1}^N B(V_i, \epsilon_0) \subset \cup_{i=1}^N B(V_i, \epsilon_1)$.
- Metric entropy: $\log M(\epsilon)$ and $\log N(\epsilon)$.
- $N(\epsilon) < \infty\ \forall\epsilon > 0 \Leftrightarrow \Theta$ is totally bounded (In topology, a metric space is said to be totally bounded if for every $\epsilon > 0$ there is a finite covering of the space by ϵ-balls). For example, a metric space is compact iff it is complete and totally bounded. Hence a compact metric space is totally bounded.

The relation between the packing number and the covering number is described in the following theorem.

Theorem 14.1. *Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \subset V$. Then*

$$M(\Theta, \|\cdot\|, 2\epsilon) \overset{(a)}{\leq} N(\Theta, \|\cdot\|, \epsilon) \overset{(b)}{\leq} M(\Theta, \|\cdot\|, \epsilon).$$

Proof. First prove part (b). Suppose $E = \{\theta_1, ..., \theta_M\}$ is a maximal packing. Then $\forall \theta \in \Theta \backslash E$, $\exists i$ such that $\|\theta - \theta_i\| \leq \epsilon$ (if this does not hold for θ then we can construct a bigger packing with $\theta_{M+1} = \theta$). Hence E is automatically an ϵ-covering. Since $N(\Theta, \|\cdot\|, \epsilon)$ is the minimal size of all possible coverings, we have $M(\Theta, \|\cdot\|, \epsilon) \geq N(\Theta, \|\cdot\|, \epsilon)$.

We next prove part (a) by contradiction. Suppose there exists a 2ϵ-packing $\{\theta_1, ..., \theta_M\}$ and an ϵ-covering $\{x_1, ..., x_N\}$ such that $M \geq N+1$. Then by pigeonhole, we must have θ_i and θ_j belonging to the same ϵ-ball $B(x_k, \epsilon)$ for some $i \neq j$ and k. This means that the distance between θ_i and θ_j cannot be more than the diameter of the ball, i.e., $\|\theta_i - \theta_j\| \leq 2\epsilon$, which leads to a contradiction since $\|\theta_i - \theta_j\| > 2\epsilon$ for a 2ϵ-packing. Hence the size of any 2ϵ-packing is less or equal to the size of any ϵ-covering. Hence $M(\Theta, \|\cdot\|, 2\epsilon)$, the maximal size of a 2ϵ-packing is at most $N(\Theta, \|\cdot\|, \epsilon)$, the minimal size of an ϵ-covering. □

When $(V, \|\cdot\|)$ is the d-dimensional Euclidean space, we can extend the previous theorem by further lower/upper bounding the covering/packing numbers. The result is given as follows.

Theorem 14.2. *Let $\Theta \subset V = \mathbb{R}^d$. Then*

$$\left(\frac{1}{\epsilon}\right)^d \frac{\text{vol}(\Theta)}{\text{vol}(B)} \overset{(a)}{\leq} N(\Theta, \|\cdot\|, \epsilon) \leq M(\Theta, \|\cdot\|, \epsilon) \overset{(b)}{\leq} \frac{\text{vol}(\Theta + \frac{\epsilon}{2}B)}{\text{vol}(\frac{\epsilon}{2}B)} \underset{\substack{\Theta \text{ convex} \\ \epsilon B \subset \Theta}}{\overset{(c)}{\leq}} \frac{\text{vol}(\frac{3}{2}\Theta)}{\text{vol}(\frac{\epsilon}{2}B)} = \left(\frac{3}{\epsilon}\right)^d \frac{\text{vol}(\Theta)}{\text{vol}(B)}.$$

where $+$ is the Minkovski sum, and $B(\epsilon)$ is the norm ball with radius ϵ and B is the unit norm ball.

Proof. First prove (a). For a covering of minimal size, $\Theta \subset \cup_{i=1}^{n} B(X_i, \epsilon)$. Hence

$$\text{vol}(\Theta) \leq \text{vol}(\cup_{i=1}^{N(\epsilon)} B(X_i, \epsilon)) \leq \sum_{i=1}^{N(\epsilon)} \text{vol}(B(X_i, \epsilon)).$$

Since $\text{vol}(B(X_i, \epsilon)) = \epsilon^d \text{vol}(B)$, we have $\text{vol}(\Theta) \leq N(\epsilon)\epsilon^d \text{vol}(B)$. Hence (a) is proved.

Next we prove (b). For an ϵ-packing, the balls $B(\theta_i, \epsilon/2)$ are disjoint, and $\cup_{i=1}^{M(\epsilon)} B(\theta_i, \epsilon/2) \subset \Theta + \frac{\epsilon}{2}B$. Taking the volume on both sides, we have

$$\text{vol}(\Theta + \frac{\epsilon}{2}B) \geq \text{vol}(\cup_{i=1}^{M(\epsilon)} B(\theta_i, \epsilon/2)) = M(\epsilon)\text{vol}(\frac{\epsilon}{2}B).$$

This proves (b).

To prove (c), we prove two statements. (1) When $\epsilon B \subset \Theta$, $\Theta + \frac{\epsilon}{2}B \subset \Theta + \frac{1}{2}\Theta$, and (2) when Θ is convex, $\Theta + \frac{1}{2}\Theta = \frac{3}{2}\Theta$.

To prove (1), notice for any $z \in \Theta + \frac{\epsilon}{2}B$, we have $z = x + y$ where $x \in \frac{\epsilon}{2}B$ and $y \in \Theta$. Since $x \in \frac{\epsilon}{2}B \Rightarrow x \in \Theta$, we immediately have $z \in \Theta + \frac{1}{2}\Theta$.

To prove (2), first notice that $\forall \theta \in \frac{3}{2}\Theta$, $\theta = \frac{1}{3}\theta + \frac{2}{3}\theta$. Since $\frac{1}{3}\theta \in \frac{1}{2}\Theta$, and $\frac{2}{3}\theta \in \Theta$, $\frac{3}{2}\Theta \subseteq \Theta + \frac{1}{2}\Theta$. On the other hand, for any $x \in \Theta + \frac{1}{2}\Theta$, we have $x = y + \frac{1}{2}z$ with $y, z \in \Theta$. When Θ is convex, $\frac{2}{3}x = \frac{2}{3}y + \frac{1}{3}z \in \Theta$. Hence $x \in \frac{3}{2}\Theta$, implying $\Theta + \frac{1}{2}\Theta \subseteq \frac{3}{2}\Theta$.

With (1) and (2), (c) follows immediately. □

Remark 14.2. Why is Theorem 14.1?

- (a) is a converse, saying that the minimal covering size cannot be too small. When combined with $N(\epsilon) \leq M(\epsilon)$, this turns into a existential statement: It is possible to construct a packing of size at least $\text{vol}(\Theta)/\text{vol}(B(\epsilon))$. From the proof we see that this corresponds to a *greedy* construction. Furthermore, for Hamming space and Hamming distance, this is exactly the *Gilbert-Varshanov* bound.

- (b) is a converse, saying that the maximal packing size cannot be too large. When combined with $N(\epsilon) \leq M(\epsilon)$, this turns into a existence statement: there exists a small covering.

Example 14.1 (Euclidean norm ball)**.** Consider $N(B_2(1), \|\cdot\|_2, \epsilon)$.[1] When $\epsilon \geq 1$, $N(B_2(1), \|\cdot\|_2, \epsilon) = 1$. When $\epsilon < 1$, activating the previous theorem we have (noticing $\Theta = B_2$)

$$\left(\frac{1}{\epsilon}\right)^d = \frac{\text{vol}(\Theta)}{\text{vol}(B_2)} \leq N(\epsilon) \leq \frac{\text{vol}((1+\frac{\epsilon}{2})B)}{\text{vol}(\frac{\epsilon}{2}B)} = \left(1 + \frac{2}{\epsilon}\right)^d \leq \left(\frac{3}{\epsilon}\right)^d.$$

Hence $d \log \frac{1}{\epsilon} \leq \log N(\epsilon) \leq d \log \frac{3}{\epsilon}$. This relationship holds for all metric norms (those such that $\text{vol}(B)\epsilon^d = \text{vol}(\epsilon B)$). If we fix the dimension and drive $\epsilon \to 0$, then because all norms in Euclidean space are equivalent, whenever Θ has interior, $\log N(\epsilon) = (d + o(1)) \log \frac{1}{\epsilon}$.

14.2 Applying metric entropy + Fano's inequality to minimax risk

We now apply metric entropy and Fano's inequality to lower bound the minimax risk. The key idea is to reduce estimation over Θ to testing between a packing $E = \{\theta_1, ..., \theta_M\}$ within $T \subset \Theta$. Then $R^*(\Theta) \geq R^*(T) \geq R^*_\pi$ where π is equi-probable over E.

Let $E = \{\theta_1, ..., \theta_M\}$ be an ϵ-packing on $T \subset \Theta$. Let $\tilde{\theta}$ be the quantized version of $\hat{\theta}$ restricted to E $(\theta - X - \hat{\theta} - \tilde{\theta})$, and consider a quadratic loss function $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$. Recall that

$$rad_{KL}(T) = \inf_Q \sup_{\theta \in T} D(P_\theta \| Q),$$

and

$$diam_{KL}(T) = \sup_{\theta, \theta' \in T} D(P_{\theta'} \| P_\theta).$$

[1]For unit ball $B_2(1) := \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ we abbreviate it as B_2.

We immediately have

$$\begin{aligned}
\mathbb{E}\left[\|\theta-\hat{\theta}\|^2\right] &\overset{Markov}{\geq} \left(\frac{\epsilon}{2}\right)^2 \mathbb{P}\left[\|\theta-\hat{\theta}\| \geq \frac{\epsilon}{2}\right] \geq \left(\frac{\epsilon}{2}\right)^2 \mathbb{P}\left[\theta \neq \tilde{\theta}\right] \\
&\overset{Fano}{\geq} \left(\frac{\epsilon}{2}\right)\left(1-\frac{I(\theta;X)+\log 2}{\log M(\epsilon)}\right) \\
&\geq \frac{\epsilon^2}{4}\left(1-\frac{rad_{KL}(T)+\log 2}{\log M(\epsilon)}\right) \\
&\geq \sup_{T\subset\Theta,\epsilon>0} \frac{\epsilon^2}{4}\left(1-\frac{diam_{KL}(T)+\log 2}{\log\frac{\mathrm{vol}(T)}{\mathrm{vol}(\epsilon B)}}\right), \qquad (14.1)
\end{aligned}$$

where in the last step, the inequality holds true for all choices of T and ϵ, and the supremum is placed to obtain a better bound.

For GLM, we can use the above method (Fano+packing) to obtain the same result (up to constant factor) by Shannon Lower Bound.

Example 14.2 (p-dimensional n-sample GLM)**.** Let $\Theta=\mathbb{R}^p$, and $T=B_2(s)$. Then $diam_{KL}(T) = \sup_{\theta,\theta'\in T} D(P_\theta\|P_{\theta'}) = \sup_{\theta,\theta'\in T} D(N(\theta,I_p)^{\otimes n}\|N(\theta',I_p)^{\otimes n}) = \sup_{\theta,\theta'} \frac{n}{2}\|\theta-\theta'\|^2 = \frac{n}{2}diam^2(T) = \frac{n}{2}s^2$. By (14.1), we have

$$R^* \geq \frac{\epsilon^2}{4}\left(1-\frac{diam_{KL}(T)+\log 2}{\log\frac{\mathrm{vol}(T)}{\mathrm{vol}(\epsilon B)}}\right) = \frac{\epsilon^2}{4}\left(1-\frac{\frac{n}{2}s^2+\log 2}{\log\frac{s^p\mathrm{vol}(B_2)}{\epsilon^p\mathrm{vol}(B_{\|\cdot\|})}}\right).$$

We now choose ϵ and s. Denote $\mathrm{vol}(B_{\|\cdot\|})=V$, and recall that $\mathrm{vol}^{1/p}(B_2)\asymp\frac{1}{\sqrt{p}}$, i.e., $c_1\frac{1}{\sqrt{p}} < \mathrm{vol}^{1/p}(B_2) < c_2\frac{1}{\sqrt{p}}$. If we choose

$$s = c_3\sqrt{\frac{p}{n}},\ \epsilon = c_4\frac{1}{\sqrt{n}V^{1/p}},$$

then

$$R^* \geq \frac{c_4^2}{4nV^{2/p}}\left(1-\frac{\frac{c_3^2 p}{2}+\log 2}{p\log\frac{c_1c_3}{c_4}}\right) \geq \frac{c_4^2}{4nV^{2/p}}\left(1-\frac{\frac{c_3^2}{2}+\log 2}{\log\frac{c_1c_3}{c_4}}\right).$$

As long as we choose c_1,c_2,c_3,c_4 such that $(\frac{c_3^2}{2}+\log 2)/\log\frac{c_1c_3}{c_4} < c < 1$, we have

$$R^* \geq \frac{c_4^2(1-c)}{4nV^{2/p}} \gtrsim \frac{1}{nV^{2/p}}. \qquad (14.2)$$

Remark 14.3. When the specified norm is $\|\cdot\|_\infty$, the norm ball becomes a cube, and the volume is (for fixed values of p)

$$V = 2^p.$$

Hence $R^*\gtrsim\frac{1}{n}$; however, we know $R^*\asymp\frac{\log p}{n}$ and we lose the dependence on the dimension p.

So what is to be blamed? It turns out our mutual information method, and in fact, its further relaxation via packing and Fano's inequality is tight in this case. What is loose is the volume ratio bound on packing number in Theorem 14.1. In the next lecture, we will prove

$$\log N(B,\|\cdot\|,\epsilon) \asymp \begin{cases} p\log\frac{1}{\epsilon\sqrt{p}}, & \epsilon\lesssim\frac{1}{\sqrt{p}} \\ \frac{1}{\epsilon^2}\log(p\epsilon^2), & \epsilon\gtrsim\frac{1}{\sqrt{p}} \end{cases}.$$

This will lead to the tight result $R^* \asymp \frac{\log p}{n}$.