Covering Number Estimates

Mathematical Analysis of Machine Learning Algorithms
(Chapter 5)
Packing Number

Definition 1 (Packing Number)

Let \((\mathcal{Y}, d)\) be a pseudometric space with metric \(d(\cdot, \cdot)\). A finite subset \(\mathcal{G}(\epsilon) \subset \mathcal{G}\) is an \(\epsilon\)-packing of \(\mathcal{G}\) if \(d(\phi, \phi') > \epsilon\) for all \(\phi, \phi' \in \mathcal{G}(\epsilon)\). The \(\epsilon\)-packing number of \(\mathcal{G}\), denoted by \(M(\epsilon, \mathcal{G}, d)\), is the largest cardinality of \(\epsilon\)-packing of \(\mathcal{G}\).
Theorem 2 (Thm 5.2)

For all $\epsilon > 0$, we have

$$N(\epsilon, \mathcal{G}, d) \leq M(\epsilon, \mathcal{G}, d) \leq N(\epsilon/2, \mathcal{G}, d).$$

It is often convenient to use packing number because an $\epsilon$-packing always belong to $\mathcal{G}$. This means that any assumption of $\mathcal{G}$ holds for an $\epsilon$-packing of $\mathcal{G}$. 
Proof of Theorem 2

Let $G(\epsilon) = \{\phi_1, \ldots, \phi_M\} \subset G$ be a maximal $\epsilon$-packing of $G$. Given any $\phi \in G$, by the definition of maximality, we know that there exists $\phi_j \in G(\epsilon)$ so that $d(\phi_j, \phi) \leq \epsilon$. This means that $G(\epsilon)$ is also an $\epsilon$ cover of $G$. Therefore $N(\epsilon, G, d) \leq M$. This proves the first inequality.

On the other hand, let $G'(\epsilon/2)$ be an $\epsilon/2$ cover of $G$. By definition, for any $\phi_j \in G(\epsilon)$, there exists $\tilde{g}(\phi_j) \in G'(\epsilon/2)$ such that $d(\tilde{g}(\phi_j), \phi_j) \leq \epsilon/2$. For $j \neq i$, we know that $d(\phi_i, \phi_j) > \epsilon$, and thus triangle inequality implies that

$$d(\tilde{g}(\phi_j), \phi_i) \geq d(\phi_i, \phi_j) - d(\tilde{g}(\phi_j), \phi_j) > \epsilon/2 \geq d(\tilde{g}(\phi_i), \phi_i).$$

Therefore $\tilde{g}(\phi_i) \neq \tilde{g}(\phi_j)$. This implies the map $\phi_j \in G(\epsilon) \rightarrow \tilde{g}(\phi_j) \in G'(\epsilon/2)$ is one to one. Therefore $|G(\epsilon)| \leq |G'(\epsilon/2)|$. This proves the second inequality.
Finite Dimensional Space

Theorem 3 (Thm 5.3)

Let \( \| \cdot \| \) be a seminorm on \( \mathbb{R}^k \). Let \( B(r) = \{ z \in \mathbb{R}^k : \| z \| \leq r \} \) be the \( \| \cdot \| \)-ball with radius \( r \). Then

\[
M(\epsilon, B(r), \| \cdot \|) \leq (1 + 2r/\epsilon)^k.
\]

Moreover,

\[
N(\epsilon, B(r), \| \cdot \|) \geq (r/\epsilon)^k.
\]
Proof of Theorem 3 (I/II)

Let \( \{z_1, \ldots, z_M\} \subset B(r) \) be a maximal \( \epsilon \) packing of \( B(r) \). Let 
\[ B_j = \{z \in \mathbb{R}^k : \|z - z_j\| \leq \epsilon/2\}, \]
then \( B_j \cap B_k = \emptyset \) for \( j \neq k \) and \( B_j \subset B(r + \epsilon/2) \) for all \( j \). It follows that

\[
\sum_{j=1}^{M} \text{volume}(B_j) = \text{volume}(\bigcup_{j=1}^{M} B_j) \leq \text{volume}(B(r + \epsilon/2)).
\]

Let \( v = \text{volume}(B(1)) \). Since \( \text{volume}(B_j) = (\epsilon/2)^k v \) and 
\( \text{volume}(B(r + \epsilon/2)) = (r + \epsilon/2)^k v \), we have

\[
M(\epsilon/2)^k v \leq (r + \epsilon/2)^k v.
\]

This implies the first bound.
Proof of Theorem 3 (II/II)

Let \( \{z_1, \ldots, z_N\} \subset \mathbb{R}^k \) be a cover of \( B(r) \). If we define \( B_j = \{z \in \mathbb{R}^k : \|z - z_j\| \leq \epsilon\} \), then \( B(r) \subset \bigcup_j B_j \). Therefore

\[
\text{volume}(B(r)) \leq \text{volume}(\bigcup_{j=1}^N B_j) \leq \sum_{j=1}^N \text{volume}(B_j).
\]

Let \( v = \text{volume}(B(1)) \). Since \( \text{volume}(B_j) = (\epsilon)^k v \) and \( \text{volume}(B(r)) = r^k v \), we have

\[
r^k v \leq N\epsilon^k v.
\]

This implies the second bound.
Theorem 4 (Thm 5.4)

Consider

\[ \{ \phi(w, Z) : w \in \Omega \} , \]  

(1)

where \( \Omega \subset \mathbb{R}^k \) is a compact set. Assume that \( \Omega \subset \mathbb{R}^k \) is a compact set so that \( \Omega \in B(r) \) with respect to a norm \( \| \cdot \| \). Assume for all \( z \), \( \phi(w, z) \) is \( \gamma(z) \) Lipschitz with respect to \( w \):

\[ |\phi(w, z) - \phi(w', z)| \leq \gamma(z)\|w - w'|. \]

Given \( p \geq 1 \), let \( \gamma_p = \left( \mathbb{E}_{Z \sim D} |\gamma(Z)|^p \right)^{1/p} \). Then

\[ N(2\epsilon, G, L_p(D)) \leq (1 + 2\gamma_p r / \epsilon)^k. \]
Proof of Theorem 4

Let \( \{w_1, \ldots, w_M\} \) be an \( \epsilon/\gamma_p \) packing of \( \Omega \). Then it is also an \( \epsilon/\gamma_p \) cover of \( \Omega \).

Let

\[
\phi_j^L(z) = \phi(w_j, z) - \gamma(z)\epsilon/\gamma_p
\]

and

\[
\phi_j^U(z) = \phi(w_j, z) + \gamma(z)\epsilon/\gamma_p.
\]

Then \( \{[\phi_j^L, \phi_j^U] : j = 1, \ldots, M\} \) is an \( 2\epsilon \) \( L_p(D) \)-bracketing cover.

We can now apply Theorem 3 to obtain the desired result.
Empirical $L_1$ Covering of VC-class

**Theorem 5 (Thm 5.5)**

If a binary valued function class $\mathcal{G} = \{\phi(w,Z) : w \in \Omega\}$ is a VC class, then for $\epsilon \leq 1$:

$$\ln M(\epsilon, \mathcal{G}, L_1(S_n)) \leq 3d + d \ln(\ln(4/\epsilon)/\epsilon).$$
Proof of Theorem 5 (I/II)

Given \( S_n = \{Z_1, \ldots, Z_n\} \). Let \( Q = \{\phi_1, \ldots, \phi_m\} \) be a maximal \( \epsilon \) \( L_1(S_n) \) packing of \( \mathcal{G} \). \( Q \) is also an \( L_1 \) \( \epsilon \)-cover of \( \mathcal{G} \). Consider the empirical distribution, denoted by \( S_n \), which puts a probability of \( 1/n \) on each \( Z_i \). We have for \( j \neq k \):

\[
\Pr_{Z \sim S_n} [\phi_j(Z) = \phi_k(Z)] = 1 - \mathbb{E}_{Z \sim S_n} |\phi_j(Z) - \phi_k(Z)| < 1 - \epsilon.
\]

Now consider random sample with replacement from \( S_n \) for \( T \) times to obtain samples \( \{Z_{i_1}, \ldots, Z_{i_T}\} \). We have

\[
\Pr(\forall \ell : \phi_j(Z_{i_\ell}) = \phi_k(Z_{i_\ell})) < (1 - \epsilon)^T \leq e^{-T\epsilon}.
\]

That is, with probability larger than \( 1 - e^{-T\epsilon} \),

\[
\exists \ell : \phi_j(Z_{i_\ell}) \neq \phi_k(Z_{i_\ell}).
\]

Taking the union bound for all \( j \neq k \), we have with probability larger than \( 1 - (m/2) \cdot e^{-T\epsilon} \), for all \( j \neq k \):

\[
\exists \ell : \phi_j(Z_{i_\ell}) \neq \phi_k(Z_{i_\ell}).
\]
Proof of Theorem 5 (II/II)

If we take \( T = \lceil \ln(m^2)/\epsilon \rceil \), then \( e^{-T \epsilon (m^2/2)} \leq 1 \). Then there exists \( T \) samples \( \{ Z_{i_\ell} : \ell = 1, \ldots, T \} \) such that \( \phi_j \neq \phi_k \) for all \( j \neq k \) when restricted to these samples. Since \( \text{VC}(G) = d \), we obtain from Sauer’s lemma:

\[
m \leq \max[2, eT/d]^d \leq \max[2, e(1 + \ln(m^2)/\epsilon)/d]^d.
\]

The theorem holds automatically when \( m \leq 2^d \). Otherwise, \( \ln m \leq d \ln(1/\epsilon) + d \ln((e\epsilon/d) + (2e/d) \ln(m)) \).

Let \( u = d^{-1} \ln m - \ln(1/\epsilon) - \ln \ln(4/\epsilon) \) and let \( \epsilon \leq 1 \), we can obtain the following bound by using the upper bound of \( \ln m \):

\[
u \leq - \ln \ln(4/\epsilon) + \ln((e\epsilon/d) + 2e(u + \ln(1/\epsilon) + \ln \ln(4/\epsilon)))
\leq \ln \frac{2e(u + 0.5 + \ln(1/\epsilon) + \ln \ln(4/\epsilon))}{\ln(4/\epsilon)} \leq \ln(4u + 7),
\]

where the last inequality is obtained by taking \( \sup \) over \( \epsilon \in (0, 1] \). By solving this inequality we obtain a bound \( u \leq 3 \). This implies the desired result.
Theorem 6 ([Haussler, 1995], Thm 5.6)

Let $\mathcal{G}$ be a binary valued function class with $\text{VC}(\mathcal{G}) = d$. Then

$$\ln M(\epsilon, \mathcal{G}, L_1(S_n)) \leq 1 + \ln(d + 1) + d \ln(2e/\epsilon).$$

Corollary 7 (Cor 5.7)

If $\text{VC}(\mathcal{G}) = d$, then for all distributions $\mathcal{D}$ over $\mathbb{Z}$, we have

$$\ln N(\epsilon, \mathcal{G}, L_p(\mathcal{D})) \leq 1 + \ln(d + 1) + d \ln(2e/\epsilon^p)$$

for $\epsilon \in (0, 1]$ and $p \in [1, \infty)$. 

VC-Subgraph Class

One may extend the concept of VC dimension to real valued functions by introducing the definition of VC subgraph class.

**Definition 8**

A real valued function class of \( z \in Z \)

\[
G = \{ \phi(w, Z) : w \in \Omega \}
\]

is a VC-subgraph class, if the binary function class

\[
G_{\text{subgraph}} = \{ 1(t < \phi(w, z)) : w \in \Omega \}
\]

defined on \((z, t) \in Z \times \mathbb{R}\) is a VC class. The VC dimension (some times also called pseudo-dimension) of \(G\) is

\[
VC(G) = VC(G_{\text{subgraph}}).
\]
Example: Linear Functions

Example 9

The $d$ dimensional linear functions of the form

$$f_w(x) = w^\top x$$

is VC subgraph class of VC dimension $d + 1$. This is because $w^\top x - t$ is linear function in $d + 1$ dimension, and we have shown that it has VC dimension $d + 1$. 
Example 10

If $\mathcal{F} = \{ f(w, x) : w \in \Omega \}$ is a VC subgraph class and $h$ is monotone function, then

$$h \circ \mathcal{F} = \{ h(f(w, x)) : w \in \Omega \}$$

is a VC subgraph class with

$$\text{vc}(h \circ \mathcal{F}) \leq \text{vc}(\mathcal{F}).$$

In particular, if $f(w, x) = w^\top x$ is a $d$-dimensional linear function, then $h(f(w, x))$ has VC dimension $d + 1$. 
Theorem 11 (Thm 5.11)

Assume that $\mathcal{G}$ is a VC subgraph class, with VC dimension $d$, and all $\phi \in \mathcal{G}$ are bounded: $\phi(Z) \in [0, 1]$. Then for any distribution $\mathcal{D}$ over $Z$, $\epsilon \in (0, 1]$ and $p \in [1, \infty)$, we have

$$\ln N(\epsilon, \mathcal{G}, L_p(\mathcal{D})) \leq 1 + \ln(d + 1) + d \ln \left(\frac{2e}{\epsilon^p}\right).$$

Moreover,

$$\ln N_{\infty}(\epsilon, \mathcal{G}, n) \leq d \ln \max[2, \frac{en}{(d\epsilon)}].$$
Proof of Theorem 11 (I/II)

Let $U$ be a random variable distributed uniformly over $[0, 1]$. Then for all $a \in (0, 1)$: $\mathbb{E}_U \mathbb{1}(U \leq a) = a$. Thus for all $\phi, \phi' \in \mathcal{G}$:

$$\mathbb{E}_D |\phi(Z) - \phi'(Z)|^p$$
$$= \mathbb{E}_D |\mathbb{E}_U [\mathbb{1}(U \leq \phi(Z)) - \mathbb{1}(U \leq \phi'(Z))]|^p$$
$$\leq \mathbb{E}_D \mathbb{E}_U |\mathbb{1}(U \leq \phi(Z)) - \mathbb{1}(U \leq \phi'(Z))|^p.$$

The last inequality used the Jensen’s inequality. Therefore

$$\ln N(\epsilon, \mathcal{G}, L_p(D)) \leq \ln N(\epsilon, \mathcal{G}_{\text{subgraph}}, L_p(D \times U(0, 1))).$$

This leads to the first desired bound.
The second bound can be proved by discretizing $U$ into intervals with thresholds $\min(1, \epsilon(2k + 1))$ for $k = 0, 1, \ldots$ with no more than $\lceil (2\epsilon)^{-1} \rceil \leq 1/\epsilon$ thresholds. This gives an $\epsilon$-cover of $U$ in Euclidean distance.

We can then approximate $\mathbb{E}_U$ by average over the thresholds to get $\epsilon$ $L_\infty$ cover with the discretization. Let the set of thresholds be $U'$.

If $D$ contain $n$ data points, then $D \times U'$ contains at most $n|U'| \leq n/\epsilon$ points, and one may apply Sauer’s lemma to obtain a cover on these points. This implies the second bound.
Regularized Linear Function Class

\[ \mathcal{F} = \{ f(w, x) = w^\top \psi(x) : w \in \Omega, x \in \mathcal{X} \} \quad (2) \]

where \( \psi(x) \) is a known feature vector.

**Theorem 12 (Thm 5.18)**

Let \( w = [w_1, w_2, \ldots] \in \mathbb{R}^\infty \) and \( \psi(x) = [\psi_1(x), \psi_2(x), \ldots] \in \mathbb{R}^\infty \). Let \( \Omega = \{ w : \|w\|_2 \leq A \} \). Given a distribution \( \mathcal{D} \) on \( \mathcal{X} \). Assume there exists \( B_1 \geq B_2 \geq \cdots \) such that

\[ \mathbb{E}_{x \sim \mathcal{D}} \sum_{i \geq j} \psi_i(x)^2 \leq B_j^2. \]

Define \( \tilde{d}(\epsilon) = \min\{ j \geq 0 : AB_{j+1} \leq \epsilon \} \). Then the function class \( \mathcal{F} \) of (2) satisfies:

\[ \ln N(\epsilon, \mathcal{F}, L_2(\mathcal{D})) \leq \tilde{d}(\epsilon/2) \ln \left( 1 + \frac{4AB_1}{\epsilon} \right). \]
Proof of Theorem 12

Given $\epsilon > 0$. Consider $j = \tilde{d}(\epsilon/2)$ such that $AB_{j+1} \leq \epsilon/2$. Let

$$F_1 = \left\{ \sum_{i=1}^{j} w_i \psi_i(x) : w \in \Omega \right\}$$

$$F_2 = \left\{ \sum_{i>j} w_i \psi_i(x) : w \in \Omega \right\}.$$

Since $\|f\|_{L_2(D)} \leq \epsilon/2$ for all $f \in F_2$, we have $N(\epsilon/2, F, L_2(D)) = 1$. Moreover, Theorem 3 implies that

$$\ln N(\epsilon/2, F_1, L_2(D)) \leq \tilde{d}(\epsilon/2) \ln \left( 1 + \frac{4AB_0}{\epsilon} \right).$$

Note that $F \subset F_1 + F_2$, we have

$$\ln N(\epsilon, F, L_2(D)) \leq \ln N(\epsilon/2, F_1, L_2(D)) + \ln N(\epsilon/2, F_2, L_2(D)).$$

This implies the result.
Example

Example 13

Assume that $B_j = j^{-q}$, then

$$\ln N(\epsilon, \mathcal{F}, L_2(D)) = O(\epsilon^{-q} \ln(1/\epsilon)) .$$

If $B_j = O(c^j)$ for some $c \in (0, 1)$, then

$$\ln N(\epsilon, \mathcal{F}, L_2(D)) = O\left((\ln(1/\epsilon))^2\right) .$$
**L₂ Regularized Empirical L∞ Covering Number**

**Theorem 14 (Thm 5.20)**

Assume that \( \Omega = \{ w : \| w \|_2 \leq A \} \) and \( \| \psi(x) \|_2 \leq B \), then the function class (2) has the following covering number bound:

\[
\ln N(\mathcal{F}, \epsilon, L_{\infty}(S_n)) \leq \frac{36 A^2 B^2}{\epsilon^2} \ln[2 \lceil (4AB/\epsilon) + 2 \rceil n + 1].
\]

Empirical \( L_{\infty} \) covering number bounds can be used in margin analysis and can be used to derive better bounds in multiclass classification.
Summary (Chapter 5)

- Packing number and relationship with covering number
- Finite dimensional function classes
- $L_p$ covering for VC class.
- VC-subgraph class.
- Regularized linear function class.