# STA 3000F Lecture 6

Recap: covering / packing #.

Thm (duality)

$$M(K;\rho,2\varepsilon) \overset{(i)}{\leq} \underbrace{N}_{\text{covering}}(\underbrace{K}_{\text{set}};\underbrace{\rho}_{\text{metric}},\underbrace{\varepsilon}_{\text{radius}}) \overset{(ii)}{\leq} \underbrace{M}_{\text{packing.}}(K;\rho,\varepsilon)$$

Proof: (i) Let $\{\theta_j\}_{j=1}^{M}$ be $2\varepsilon$-packing of K

$\{\psi_j\}_{j=1}^{N}$ be $\varepsilon$-covering of K.

(want to show $M \leq N$).

$\forall \theta_i$, $(i=1,2\ldots M)$, $\exists j$ s.t. $\theta_i \in \mathbb{B}_\rho(\psi_j,\varepsilon)$.

(more than one — choose arbitrarily).

Suppose if $\exists\ \theta_i, \theta_l$ $(i \neq l)$

s.t. $\theta_i, \theta_l \in \mathbb{B}_\rho(\psi_j,\varepsilon)$

then $\rho(\theta_i,\theta_l) \leq \rho(\theta_i,\psi_j)+\rho(\psi_j,\theta_l) \leq 2\varepsilon$

Contradiction ($\{\theta_i\}_{i\in[M]}$ is $2\varepsilon$-packing).

So different $\theta$'s maps to different $\psi$.

$M \leq N$.

(ii). Let $\{\theta_j\}_{j=1}^{M}$ be a maximal $\varepsilon$-packing.

By maximality, $\forall \theta \in K$, $\exists j \in [M]$ s.t. $\rho(\theta, \theta_j) \le \varepsilon$

(otherwise, $\theta$ can be added to the packing).

So $\{\theta_j\}_{j=1}^{M}$ is also an $\varepsilon$-covering.

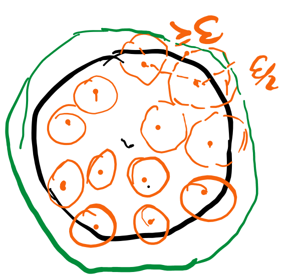
$$\underbrace{N(K; \rho, \varepsilon)}_{\text{minimal } \varepsilon\text{-covering}} \le \underbrace{M}_{\text{an } \varepsilon\text{-covering.}}$$

---

e.g. (Euclidean space).

$$K := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le 1\}. \qquad \rho = \|\cdot\|_2.$$

· Upper bound:




$$\underbrace{\bigcup_{j=1}^{M} \mathbb{B}(\theta_j, \varepsilon/2)}_{\text{disjoint.}} \subseteq \mathbb{B}\left(0, 1+\frac{\varepsilon}{2}\right)$$

$$\sum_{j=1}^{M} \mathrm{Vol}\left(\mathbb{B}(\theta_j, \varepsilon/2)\right) \le \mathrm{Vol}\left(\mathbb{B}\left(0, 1+\frac{\varepsilon}{2}\right)\right).$$

$$M \le \left(1+\frac{2}{\varepsilon}\right)^d.$$

· Lower bound: suppose

$$K \subseteq \bigcup_{j=1}^{N} \mathbb{B}(\psi_j, \varepsilon).$$

$$\mathrm{Vol}(K) \le \sum_{j=1}^{N} \mathrm{Vol}\left(\mathbb{B}(\psi_j, \varepsilon)\right) \qquad \text{So} \quad N \ge \left(\frac{1}{\varepsilon}\right)^d.$$

By duality thm:

$$\left(\frac{1}{\varepsilon}\right)^d \le N(K; \|\cdot\|_2, \varepsilon) \le \left(1+\frac{2}{\varepsilon}\right)^d.$$

---

Back to statistics.

. Consider M-estimation

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta \in \Theta} F_n(\theta)$$

where $F_n(\theta) := \frac{1}{n}\sum_{i=1}^n f(\theta; X_i)$.

from last lecture, we know $\hat{\theta}_n \xrightarrow{P} \theta^*$

is implied by

- $\theta^*$ is unique minimizer of $F(\theta) := \mathbb{E}[f(\theta; X)]$
  ($F$ is cts) — (implies $\varepsilon$-$\delta$ condition in last lecture)
- $\sup_{\theta \in \Theta} |F_n(\theta) - F(\theta)| \xrightarrow{P} 0.$ (ULLN)

Thm (Wald) Suppose $\Theta$ is a compact set, and

(i) $\mathbb{E}\left[\sup_{\theta \in \Theta} |f(\theta; X)|\right] < +\infty$. (DCT-type condition)

(ii) $f(\cdot; X)$ is cts (w.r.t. first argument).

then ULLN $\sup_{\theta \in \Theta} |F_n(\theta) - F(\theta)| \xrightarrow{P} 0$ holds

(Combining w/ uniqueness of $\theta^*$, it leads to consistency)

Proof. Since $f(\cdot;X)$ is cts, $F$ is cts on $\Theta$

so $F$ is uniformly cts

Fix $\varepsilon>0$, $\exists\,\delta$ s.t. $\forall\,\theta,\theta'\in\Theta$, $\rho(\theta,\theta')\le\delta$,

we have $|F(\theta)-F(\theta')|\le\varepsilon$.

(Recall: we're able to prove ULLN for finite $\Theta$,

Compact $\Theta$: use discretization)

Let $\{\theta_j\}_{j=1}^N$ be a $\delta$-covering of $\Theta$.

$$\sup_{\theta\in\Theta}|F_n(\theta)-F(\theta)|$$

need technicality $\quad\le\varepsilon$

$$\le \max_{i\in[N]}\ \sup_{\theta:\rho(\theta,\theta_i)\le\delta}\left\{\underbrace{|F_n(\theta)-F_n(\theta_i)|}_{①}+\underbrace{|F(\theta)-F(\theta_i)|}_{\le\varepsilon}+\underbrace{|F_n(\theta_i)-F(\theta_i)|}_{②}\right\}$$

depends on $i$ but not $\theta$.

$$\max_{i\in[N]}|F_n(\theta_i)-F(\theta_i)|\xrightarrow{P}0$$

(for fixed covering)

First term

$$=\max_{i\in[N]}\ \sup_{\theta\in B_\rho(\theta_j;\delta)}\left|\frac{1}{n}\sum_{i=1}^n\big(f(\theta;X_j)-f(\theta_i;X_j)\big)\right|$$

$$\le\max_{i\in[N]}\frac{1}{n}\sum_{i=1}^n\ \sup_{\theta\in B_\rho(\theta_j,\delta)}|f(\theta;X_j)-f(\theta_j;X_j)|$$

$$\xrightarrow[\text{(by LLN)}]{P} \max_{i\in[N]} \mathbb{E}\left[\sup_{\theta\in B_\rho(\theta_i;\delta)} |f(\theta;X) - f(\theta_i;X)|\right].$$

$$\le \sup_{\theta\in K} \mathbb{E}\left[\sup_{\rho(\theta',\theta)\le\delta} |f(\theta;X) - f(\theta';X)|\right]$$

Only need to show

$$\sup_{\theta\in K} \mathbb{E}\Big[\underbrace{\sup_{\rho(\theta',\theta)\le\delta} |f(\theta;X) - f(\theta';X)|}_{\to 0 \text{ ptwise.}}\Big] \xrightarrow{P} 0 \quad (\text{as } \delta\to 0).$$

$$M_\delta(\theta) := \sup_{\rho(\theta',\theta)\le\delta} |f(\theta;X) - f(\theta';X)| \longrightarrow 0.$$

$$|M_\delta(\theta)| \le 2\cdot \sup_{\theta\in\Theta} |f(\theta;X)| \in L^1.$$

So by DCT, $\mathbb{E}[M_\delta(\theta)] \to 0$ $(\forall\theta\in\Theta)$.

By Dini's thm, convergence is uniform in $\theta\in\Theta$.

---

Remark:

- Only need weak conditions, works for general metric spaces
- metric is used in $\begin{cases} \text{compactness of } \Theta \\ \text{cts func } f(\cdot;X) \end{cases}$
- Quantitative bound?

For ②:

$$\mathbb{P}\left( \max_{i\in[N]} |F_n(\theta_i) - F(\theta_i)| \geq \varepsilon \right)$$

$$\leq \sum_{i\in[N]} \mathbb{P}\left(|F_n(\theta_i) - F(\theta_i)| \geq \varepsilon\right)$$

$$\leq N \cdot \sup_{\theta\in\Theta} \mathbb{P}\left(|F_n(\theta) - F(\theta)| \geq \varepsilon\right)$$

- If $f$ is bdd, $\mathbb{P}(|\cdots| \geq \varepsilon) \leq \exp(-c\, n\varepsilon^2)$.

  in order to make $\mathbb{P}(\max |\cdots| \geq \varepsilon) \leq \delta$.

  we choose $\varepsilon \asymp \sqrt{\frac{\log(N/\delta)}{n}}$.

- If $f$ is heavy-tailed. (e.g. $\mathbb{E}[f(\theta;x)^2] < +\infty$)

  naive application of Chebyshev ineq gives

$$N \cdot \mathbb{P}(|\cdots| \geq \varepsilon) \leq O\left(\frac{N}{n\varepsilon^2}\right).$$

$$\varepsilon \asymp \sqrt{\frac{N}{n\delta}}.$$

This can be overcome by

"symmetrization method".

For ①: one-step discretization

is suboptimal.

use multi-level discretization.

"chaining"

---

"Symmetrization":

$$P_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i) \qquad Pf := \mathbb{E}[f(X)].$$

want to bound $\sup_{f\in\mathcal{F}} |P_n f - Pf|.$

(we replace $f(\theta;X)$ with just $f(X)$).

$$\mathcal{R}_n(\mathcal{F}) := \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i)\right|\right]$$

where $(\varepsilon_i)_{i=1}^n$ are iid Rademacher r.v.'s indp of $(X_i)_{i=1}^n$

$$\left(\varepsilon_i = \begin{cases} 1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases}\right).$$

"Rademacher complexity" of $\mathcal{F}$.

Thm. $$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |P_n f - Pf|\right] \le 2\cdot \mathcal{R}_n(\mathcal{F}).$$

Why useful: we can condition on data $(X_i)_{i=1}^n$

bound $\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \,\middle|\, (X_i)_{i=1}^n \right]$.

easy to use concentration ineq.

Proof. $\mathbb{E}\left[\sup_{f \in \mathcal{F}} |P_n f - Pf|\right]$

$$= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[f(X_i')] \right| \right]$$

$(X_i, X_i' \overset{iid}{\sim} \mathbb{P})$.

(Jensen)
$$\le \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - f(X_i')) \right| \right].$$

$$f(X_i) - f(X_i') \overset{d}{=} \varepsilon_i (f(X_i) - f(X_i'))$$

$$= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (f(X_i) - f(X_i')) \right| \right].$$

(Jensen)
$$\le 2 \cdot \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right].$$

---

For $A \subseteq \mathbb{R}^n$, want to bound

$$R_n(A) := \mathbb{E}\left[\sup_{a \in A} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i a_i \right| \right].$$

Complexity measure of set $A$.

"max correlation w/ a random direction".

related to Gaussian complexity
(replace Rademacher r.v. with $N(0,1)$).

- Finite $A \subseteq \mathbb{R}^n$.

$\forall \lambda > 0$, by Jensen's ineq.

$$R_n(A) \le \frac{1}{\lambda} \log \mathbb{E}\left[ \exp\left( \max_{a \in A} \frac{\lambda}{n} \sum_{i=1}^n \varepsilon_i a_i \right) \right] \quad \text{(MGF)}$$

$$\le \frac{1}{\lambda} \log \left( \sum_{a \in A} \mathbb{E}\left[ \exp\left( \frac{\lambda}{n} \sum_{i=1}^n \varepsilon_i a_i \right) \right] \right) \quad \text{(union bound)}$$

$$\le \frac{1}{\lambda} \log \left( \sum_{a \in A} \exp\left( \frac{\lambda^2}{2n^2} \|a\|_2^2 \right) \right)$$

$$\le \frac{1}{\lambda} \log \left( |A| \cdot \exp\left( \frac{\lambda^2}{2n^2} \max_{a \in A} \|a\|_2^2 \right) \right)$$

$$= \frac{\log |A|}{\lambda} + \frac{\lambda}{2n^2} \cdot \max_{a \in A} \|a\|_2^2 .$$

(choose optimal $\lambda$)

$$= \max_{a \in A} \frac{\|a\|_2}{\sqrt{n}} \cdot \sqrt{\frac{2 \log |A|}{n}} .$$

(Alternatively, you may also use Hoeffding + union bound to get the same result).

Discretization:

— One-step

$a^{(1)}, a^{(2)}, \dots, a^{(M)}$ $\delta$-covering of $A$.

under $\|a\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^n a_i^2}$.

$\pi(a) := \underset{a^{(i)}}{\arg\min} \|a - a^{(i)}\|_n$.

$$R_n(A) \le \underbrace{\mathbb{E}\left[\sup_{a\in A} \left|\frac{1}{n}\varepsilon^T(a-\pi(a))\right|\right]} + \underbrace{\mathbb{E}\left[\max_{j\in[M]} \left|\frac{1}{n}\varepsilon^T a^{(j)}\right|\right]}$$

$\le \mathrm{diam}(A)\cdot\sqrt{\frac{2\log M}{n}}$.

(Cauchy - Schwarz)

$\le \mathbb{E}\left[\sup_{a\in A} \|\varepsilon\|_n \cdot \|a-\pi(a)\|_n\right] \le \delta$.

So $R_n(A) \le \delta + \mathrm{diam}(A)\cdot\sqrt{\frac{2\log M(A;\delta)}{n}}$

e.g. $\mathcal{F}$ = d-dim parametric model

then $A$ is a d-dim manifold.

choose $\delta = O\left(\frac{1}{n}\right)$. $M(A;\delta) \asymp \left(\frac{1}{\delta}\right)^d$

this leads to $R_n(A) \le O\left(\sqrt{\frac{d\log n}{n}}\right)$

Gap is even larger for nonparametric models.

Chaining: $\{0\} = A_0, A_1, A_2, \dots A_m, \dots$

discrete approximations to $A$.

(approx err $\to 0$ as $m \to +\infty$)

$$\mathbb{E}\left[\sup \left|\frac{1}{n}\varepsilon^T a\right|\right] \le \sum_{m=0}^{+\infty} \mathbb{E}\left[\sup_{a\in A}\left|\frac{1}{n}\varepsilon^T(\pi_{m+1}(a) - \pi_m(a))\right|\right]$$

where $\pi_m(a)$ is the best approx to $a$ in $A_m$

$a \in A$ : $\quad a = \sum_{m=0}^{+\infty}(\pi_{m+1}(a) - \pi_m(a))$

$(\pi_0(a) = 0)$

Assuming that $A_m$ is $\delta_m$-covering of $A$.

We can bound

$$\mathbb{E}\left[\sup_{a\in A}\left|\frac{1}{n}\varepsilon^T(\pi_m(a) - \pi_{m+1}(a))\right|\right]$$

$$\le (\delta_m + \delta_{m+1})\sqrt{\frac{2\log|A_m|\cdot|A_{m+1}|}{n}}.$$

$\|\pi_m(a) - \pi_{m+1}(a)\|_n \le \delta_m + \delta_{m+1}$

- $(\pi_m(a), \pi_{m+1}(a))$ takes value in a finite set, cardinality $\le |A_m|\cdot|A_{m+1}|$

Thm (Dudley's chaining).

$\exists$ a universal const $c > 0$.

$$R_n(A) \leq \frac{c}{\sqrt{n}} \int_0^{+\infty} \sqrt{\log N(A; \|\cdot\|_n, \delta)}\, d\delta.$$

Proof: $D = \max_{a \in A} \|a\|_n$.

$A_0 = \{0\}$. $A_m :=$ minimal $\delta_m$-covering

$$\delta_m = \frac{D}{2^m}.$$

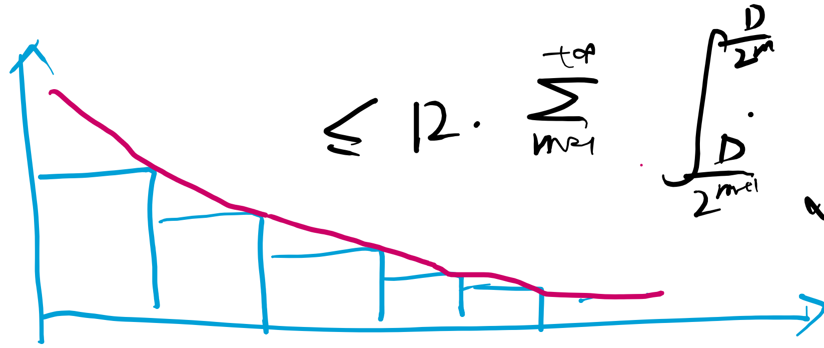
- $(\delta_m + \delta_{m+1}) = \frac{3D}{2^m}$

- $|A_m| \cdot |A_{m+1}| \leq |A_{m+1}|^2 = N(A; \|\cdot\|_n, \frac{D}{2^{m+1}})^2$.

Substitute to the bound we derived.

$$\mathbb{E}\left[\sup_{a \in A} \left|\frac{1}{n} \varepsilon^T a\right|\right] \leq \sum_{m=1}^{+\infty} \frac{3D}{2^m} \cdot \sqrt{\frac{4 \log N(\frac{D}{2^{m+1}})}{n}}$$

$$\leq 12 \cdot \sum_{m=1}^{+\infty} \int_{\frac{D}{2^{m+1}}}^{\frac{D}{2^m}} \sqrt{\frac{\log N(\delta)}{n}}\, d\delta.$$



Remarks:

- Dudley integral may diverge (V)

  of covering # upper bound.

  In that case, we use different upper bounds for diff $\delta$.

- "generic chaining" is tighter bounds (known to be optimal), but it's less convenient.

---

Back to function class.

Consider $\mathcal{F}$. assume $\exists$ envelope function $G$.

$$|f(x)| \le G(x) \quad (\forall f \in \mathcal{F})$$

(similar to Wald's consistency proof).

Thm. under above setup.

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |P_n f - Pf|\right]$$

$$\le C \cdot \sqrt{\frac{\mathbb{E}[G(X)^2]}{n}} \cdot \int_0^1 \sqrt{\log \sup_Q N(\delta \cdot \|G\|_{L^2(Q)}; \mathcal{F}, L^2(Q))}\, d\delta.$$

" worst-case $L^2$ covering # ":

Can be bounded in most cases

(see next lecture).