2 - Covering with ε-Nets

- Recap: Population risk

$$f_* = \underset{\mathcal{F}}{\operatorname{argmin}} \; R(f) = \mathbb{E}[\ell((y,x),f)]$$

$$\hat{f} = \underset{\mathcal{F}}{\operatorname{argmin}} \; \hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell((y_i,x_i),f)$$

* Excess risk: $R(\hat{f}) - R(f_*)$

! Uniform convergence ⇒ generalization

$$\mathbb{P}\left(R(\hat{f}) - R(f_*) \geq \varepsilon\right) \leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} \left|\hat{R}(f) - R(f)\right| \geq \frac{\varepsilon}{2}\right)$$

- Hoeffding's Lemma: $z_i \in [a_i, b_i]$ independent

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} z_i - \mathbb{E} z_i\right| \geq \varepsilon\right) \leq 2\exp\left\{-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right\}$$

- Last time: $\underbrace{|\mathcal{F}| < \infty}_{\text{too restrictive!}}$, $R(\hat{f}) - R(f_*) \leq O\left(\sqrt{\frac{\log|\mathcal{F}|}{n}}\right)$

sample size

complexity of $\mathcal{F}$

* Generalization of Parametric Fnc class

Let $\mathcal{F} = \{ f_\theta : \theta \in \Theta \subseteq \mathbb{R}^d \}$

Ex: Linear regression $f_\theta(x) = \langle \theta, x \rangle$ $\Theta = \{\theta : \|\theta\| \leq R\}$

$|\mathcal{F}| = \infty$, previous argument fails!

Def (Lipschitz cont): $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz cont if

$$\forall \theta, \theta', \quad |f(\theta) - f(\theta')| \leq L \|\theta - \theta'\|.$$

Remark: If f is differentiable, $\|\nabla f\| \leq L$.

- Goal: - We relied on union bound.
  $\Rightarrow$ require countable set.
  - Find a good finite representation of Θ.

- ε - Nets

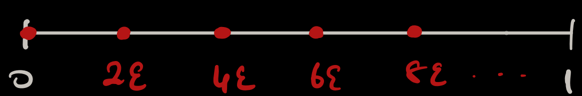
Definition (ε-cover): $\Theta \subseteq \mathbb{R}^d$ is a set and $\varepsilon > 0$.

$\mathcal{N}_\varepsilon \subseteq \Theta$ is an ε-cover of Θ if

$$\forall \theta \in \Theta, \ \exists \theta' \in \mathcal{N}_\varepsilon \ \text{s.t.} \ \|\theta - \theta'\| \leq \varepsilon.$$

- Smallest size of ε-covers of Θ is called the covering number of Θ: N_ε.

Ex: linear segment $[0,1] = \Theta$

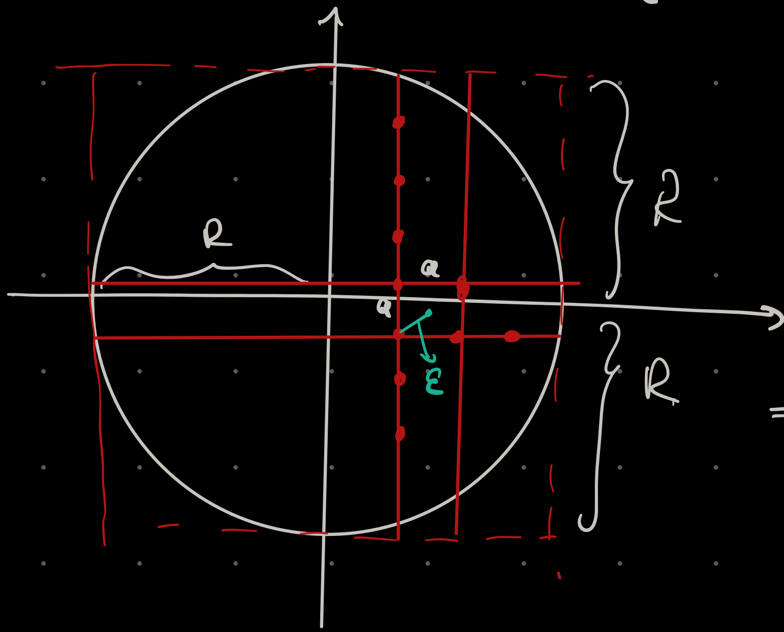


$\mathcal{N}_\varepsilon = \{0, 2\varepsilon, 4\varepsilon, \ldots\}$

is an ε-cover of Θ.

$$N_\varepsilon \leq |\mathcal{N}_\varepsilon| = \left\lfloor \frac{1}{2\varepsilon} + 1 \right\rfloor.$$

Ex: Disc $\quad \Theta = \{ \theta \in \mathbb{R}^2 : \|\theta\| \le R \}$



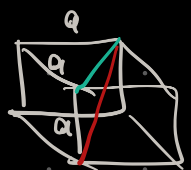
$\left(\frac{2R}{a} + 1\right)^2$ points on grid.

$\varepsilon = \frac{a}{\sqrt{2}}$

$$\Rightarrow N_\varepsilon \le |\mathcal{N}_\varepsilon| \le \left(\frac{R\sqrt{2}}{\varepsilon} + 1\right)^2$$

Ex: Sphere $\quad R \cdot S^{d-1} = \Theta = \{ \theta \in \mathbb{R}^d : \|\theta\| \le R \}$

— $\left(\frac{2R}{a} + 1\right)^d$ points in our ε-cover.

— 

$\varepsilon = \frac{a\sqrt{d}}{2}$

— $N_\varepsilon \le |\mathcal{N}_\varepsilon| \le \left(\frac{2R\sqrt{d}}{2\varepsilon} + 1\right)^d \le \left(\frac{2R\sqrt{d}}{\varepsilon}\right)^d$

not looking good

Theorem: Let $\mathcal{F} = \{ f_\theta : \underbrace{\|\theta\| \le R}_{\Theta} \}$ and define

$$\theta_* = \arg\min_{\Theta} R(\theta) = \mathbb{E}[\ell((y,x),\theta)]$$

$$\hat{\theta} = \arg\min_{\Theta} \hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell((y_i,x_i),\theta)$$

If the loss is L-Lip. cont. in θ and bounded by B, then, w.p. at least $1 - 2e^{-d/2}$,

$$R(\hat{\theta}) - R(\theta_*) \le c \cdot \sqrt{\frac{d \log n}{n}} \quad \text{for some } c > 0.$$

Remarks: – Rate $O\left(\sqrt{\frac{d\log n}{n}}\right)$ is worse than $O\left(\sqrt{\frac{d}{n}}\right)$

– $\log n$ factor is due to the covering argument

– This can be generalized to any compact Θ.

$\Rightarrow$ just replace $R = \frac{\operatorname{diam}\Theta}{2}$.

– This thm cannot cover squared error loss.

Proof: Strategy: 1 – Concentration

2 – Discretization

3 – Union bound

4 – Generalization via union bound

1 – Concentration: By Hoeffding's lemma:

Any $\theta \in \Theta$, $\mathbb{P}\left(|\hat{R}(\theta) - R(\theta)| \geq \frac{\varepsilon}{4}\right) \leq 2\exp\left\{-\frac{2n\varepsilon^2}{16B^2}\right\}$

2 – Discretization:

By Lipschitz cont. of the loss, $\hat{R}(\theta)$ and $R(\theta)$ are L-Lip. cont.

– $\forall \theta, \theta' \in \Theta$, $|\hat{R}(\theta) - R(\theta)| = |\hat{R}(\theta') - R(\theta') + \hat{R}(\theta) - \hat{R}(\theta') - R(\theta) + R(\theta')|$

by triangle ineq. and Lipschitz property $\leq |\hat{R}(\theta') - R(\theta')| + 2L\|\theta - \theta'\|$

Let $\mathcal{N}_\Delta$ be a Δ-net over Θ.

$$\Rightarrow \forall \theta \in \Theta, \ \exists \theta' \in \mathcal{N}_\Delta \ \text{s.t.} \ \|\theta - \theta'\| \leq \Delta.$$

$$\Rightarrow \forall \theta \in \Theta, \ \exists \theta' \in \mathcal{N}_\Delta \ \text{s.t.}$$

$$|\hat{R}(\theta) - R(\theta)| \leq |\hat{R}(\theta') - R(\theta')| + 2L\Delta$$

$$\sup_{\theta \in \Theta} |\hat{R}(\theta) - R(\theta)| \leq \max_{\theta' \in \mathcal{N}_\Delta} |\hat{R}(\theta') - R(\theta')| + 2L\Delta$$

3 — Union bound over $\mathcal{N}_\Delta$:

$$\mathbb{P}\left(R(\hat{\theta}) - R(\theta_*) \geq \varepsilon\right) \leq \mathbb{P}\left(\sup_{\Theta} |\hat{R}(\theta) - R(\theta)| \geq \frac{\varepsilon}{2}\right)$$

$$\leq \mathbb{P}\left(\max_{\mathcal{N}_\Delta} |\hat{R}(\theta') - R(\theta')| + 2L\Delta \geq \frac{\varepsilon}{2}\right)$$

$$= \frac{\varepsilon}{4} \Rightarrow \boxed{\Delta = \frac{\varepsilon}{8L}}$$

$$\leq \mathbb{P}\left(\max_{\mathcal{N}_\Delta} |\hat{R}(\theta') - R(\theta')| \geq \frac{\varepsilon}{4}\right)$$

$$= \mathbb{P}\left(\bigcup_{\mathcal{N}_\Delta} \left\{ |\hat{R}(\theta') - R(\theta')| \geq \frac{\varepsilon}{4} \right\}\right)$$

by union bound $$\leq \sum_{\mathcal{N}_\Delta} \mathbb{P}\left(|\hat{R}(\theta') - R(\theta')| \geq \frac{\varepsilon}{4}\right)$$

by concentration $$\leq |\mathcal{N}_\Delta| \cdot 2 \exp\left\{-\frac{n\varepsilon^2}{8B^2}\right\} \text{———} *$$

Since w is a ball of radius R, by Ex 3

$$|\mathcal{N}_\delta| \le \left(\frac{2R\sqrt{d}}{\Delta}\right)^d = \left(\frac{16LR\sqrt{d}}{\varepsilon}\right)^d$$

— $*$ $\le 2\left(\frac{16LR\sqrt{d}}{\varepsilon}\right)^d \exp\left\{-\frac{n\epsilon^2}{8B^2}\right\}$

$$= 2\exp\left\{-\frac{n\epsilon^2}{8B^2} + d\log\left(16LR\sqrt{d}\right) + d\log\frac{1}{\varepsilon}\right\}$$

$$-d \quad + \quad d\log d \quad + d\log n - d\log d$$

— Try $\varepsilon = c\sqrt{\frac{d}{n}}$:

won't work.

— I would try $c\sqrt{\frac{d}{n^{1-\delta}}}$ for $\delta > 0$.

$\Rightarrow$ which suggests $\varepsilon = c\sqrt{\frac{d\log n}{n}}$

$$= 2\exp\Big\{\underbrace{-\frac{n\epsilon^2}{8B^2} + d\log\left(16LR\sqrt{d}\right) + d\log\frac{1}{\varepsilon}}_{\delta}\Big\}$$

$$\delta = \underbrace{-\frac{\cancel{n}}{8B^2}c^2\frac{d\log n}{\cancel{n}}}_{\mathrm{I}\ \checkmark} + \underbrace{d\log\left(16LR\sqrt{d}\right)}_{\mathrm{II}\ \checkmark} + d\log\frac{1}{c}\sqrt{\frac{n}{d\log n}}$$

$$\underbrace{+\frac{d}{2}\log n}_{\mathrm{III}\ \checkmark}\ \underbrace{-\ d\log c\sqrt{d}}_{\mathrm{IV}\ \checkmark}\ \underbrace{-\ \frac{d}{2}\log\log n}_{\mathrm{V}}$$

— If $c \ge 2B \Rightarrow \mathrm{I} \ge \mathrm{III}$

— If $c \ge 16LR \Rightarrow \mathrm{IV} > \mathrm{II}$

— If $n \ge 16 \Rightarrow \log\log n \ge 1$

$\Rightarrow$ —— $*$ $\le 2\exp\left\{-\frac{d}{2}\right\}$ when $c = 2B \vee 16RL$

$\varepsilon = c\cdot\sqrt{\frac{d\log n}{n}}$ $\blacksquare$