LECTURE 40

Last Class, we proved:

Dudley's inequality: let $(X_t)_{t\in T}$ be a mean zero stochastic process on a metric space (T,d), which satisfies

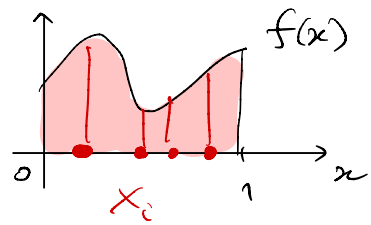
$$\|X_t - X_s\|_{\psi_2} \le K\, d(t,s) \qquad \forall\, t,s \in T.$$

Then, $\forall\, t_0 \in T$:

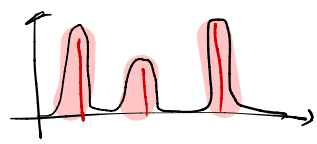
$$\mathbb{E}\sup_{t\in T} |X_t - X_{t_0}| \le CK \int_0^\infty \underbrace{\sqrt{\log N(T,d,\varepsilon)}}_{\text{metric entropy}}\, d\varepsilon.$$

Application: uniform Monte-Carlo integration:

$$\int_0^1 f(x)\,dx \overset{X\sim \mathrm{Unif}[0,1]}{=} \mathbb{E}f(X) \overset{\text{LLN}}{\approx} \frac{1}{n}\sum_{i=1}^n f(\underset{\text{iid copies of}}{X_i})$$



- Can we use one sample $X_1, \dots, X_n$ ("hardware") to integrate $\forall$ function? (ULLN) NO:



LHS small, RHS large.

- Yes for Lipschitz functions:

f is L-Lipschitz if $\boxed{|f(a)-f(b)| \le L|a-b| \quad \forall\, a,b}$

Ex: $|f'(x)| \le L \;\; \forall x \;\Rightarrow\; L$-Lipschitz.

THM Consider the class of functions

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

Let $X \in [0,1]$ be $\forall$ r.variable; X_i be independent copies of X.

Then
$$\mathbb{E}\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X) \right| \le \frac{CL}{\sqrt{n}}.$$

Remark. Thm can't be deduced from VC theory since f are not Boolean functions.

- Instead, we will use Dudley's inequality.
- WLOG $L=1$; $f(0)=0$ (translation invariance) $\Rightarrow$

$$\boxed{\mathcal{F} = \{ f : [0,1] \to \mathbb{R},\ f(0)=0,\ f \text{ is } 1\text{-Lipschitz} \}.}$$

Lem (Covering numbers of $\mathcal{F}$)

$\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|$

$$N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \begin{cases} \le e^{c/\varepsilon} & \forall \varepsilon \in (0,1) \\ = 1 & \forall \varepsilon > 1 \end{cases}$$

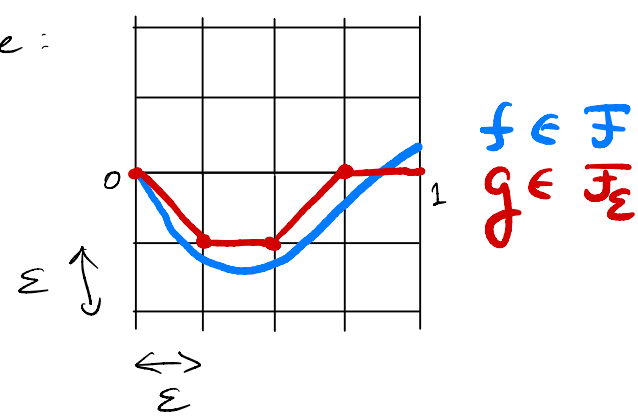
① $\varepsilon > 1$: $|f(x)| = |f(x) - f(0)| \le |x-0| \le 1 \quad \forall x \in [0,1]$

$\Rightarrow$ the zero function forms a 1-net of $\mathcal{F}$ $\Rightarrow$ QED.

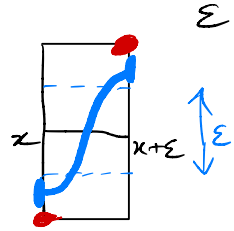
② $\varepsilon \le 1$: put the ε-grid on the plane:

$\forall f \in \mathcal{F}$ can be approximated by g that (a) takes values on the grid

(b) $g(x+\varepsilon) - g(x) \in \{-1, 0, 1\}$ $\forall x \in \varepsilon\mathbb{Z}$



(indeed, if g jumps by 2:



then f jumps by $> \varepsilon$

this contradicts being 1-Lipschitz)

# of such functions g is $\underbrace{3 \times \cdots \times 3}_{1/\varepsilon \text{ steps}} = 3^{1/\varepsilon}$

since at each step, g can go up/down/stay put (3 possibilities)

Proof of Thm

- $$\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\underbrace{\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}f(X)}_{=:\,X_f}\Big| \le \mathbb{E}\sup_{f\in\mathcal{F}}|X_f-\overset{0\text{ (zero function)}}{X_0}| \quad (\le)$$

  X_f stochastic process indexed by $f\in\mathcal{F}$

- Subgaussian increments? $\forall f,g\in\mathcal{F}$:

  $$X_f-X_g=\frac{1}{n}\sum_{i=1}^n Z_i \quad \text{where} \quad Z_i=(f-g)(X_i)-\mathbb{E}(f-g)(X)$$

  are iid mean 0 random variables with

  $$|Z_i|\le|(f-g)(X_i)|\le\|f-g\|_\infty$$

  $$\Rightarrow \|Z_i\|_{\psi_2}\lesssim\|f-g\|_\infty$$

  $$\Big\|\sum_{i=1}^n Z_i\Big\|_{\psi_2}\lesssim\Big(\sum_{i=1}^n\|Z_i\|_{\psi_2}^2\Big)^{1/2}$$ (subgaussian Hoeffding ineq. lec. 14, Oct. 3)

  $$\lesssim\sqrt{n}\,\|f-g\|_\infty \quad \Rightarrow$$

  $$\|X_f-X_g\|_{\psi_2}\lesssim\frac{1}{\sqrt{n}}\|f-g\|_\infty \qquad \forall f,g\in\mathcal{F}.$$

- Apply Dudley $\Rightarrow$

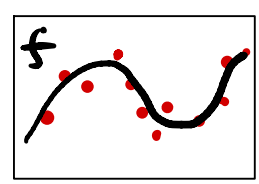
  $$(\lesssim)\ \frac{1}{\sqrt{n}}\int_0^\infty\sqrt{\log N(\mathcal{F},\|\cdot\|_\infty,\varepsilon)}\,d\varepsilon \overset{\text{lem p. 2}}{\lesssim} \frac{1}{\sqrt{n}}\int_0^1\sqrt{\frac{1}{\varepsilon}}\,d\varepsilon\lesssim\frac{1}{n}. \qquad \text{QED.}$$

Consequence in ML "Lipschitz regression":

Fit the best L-Lipschitz function to the data.

The generalization error is $O(L/\sqrt{n})$

like in linear regression ☺



$(X_1,Y_1),\dots,(X_n,Y_n)$.

Non-parametric.

Higher Dimensions?

$\mathcal{F} = \{ f : [0,1]^d \to \mathbb{R},\ f \text{ is } L\text{-Lipschitz} \}$

$\Updownarrow$

$|f(a) - f(b)| \le L \|a-b\|_\infty \quad \forall a, b \in [0,1]^d$

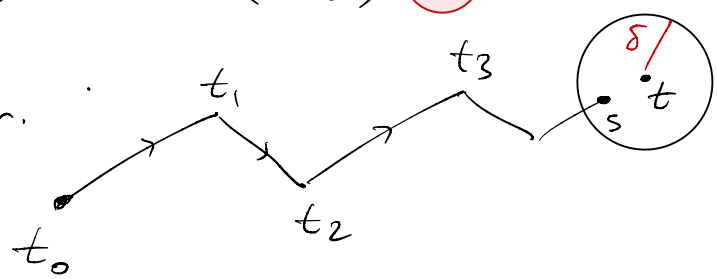
- **Lem** $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \begin{cases} \le e^{(c/\varepsilon)^d} & \forall\, \varepsilon \in [0,1] \\ = 1 & \forall\, \varepsilon > 1 \end{cases}$ (DIY)

  # grid pts $= (1/\varepsilon)^d$

- However, when we plug this into Dudley, the integral

$$\int_0^1 \sqrt{(c/\varepsilon)^d}\, d\varepsilon = \infty \quad (d \ge 2)$$

- Fix: stop chaining earlier.



Proof of Dudley yields

$$\mathbb{E} \sup_{t \in T} |X_t - X_{t_0}| \lesssim \int_\delta^\infty \sqrt{\log N(T, d, \varepsilon)}\, d\varepsilon + \mathbb{E} \sup_{\substack{t,s \in T \\ d(t,s) \le \delta}} |X_t - X_s|$$

(Check!) KW

- Apply for $X_f = \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X)$

$$\mathbb{E} \sup_{f \in \mathcal{F}} |X_f| \lesssim \frac{1}{\sqrt{n}} \int_\delta^1 \sqrt{(c/\varepsilon)^d}\, d\varepsilon + \mathbb{E} \sup_{\substack{f,g \in \mathcal{F} \\ \|f-g\|_\infty \le \delta}} \underbrace{|X_f - X_g|}_{\frac{1}{n}\sum_1^n \underbrace{(f-g)(X_i)}_{\le \delta} - \underbrace{\mathbb{E}(f-g)(X)}_{\le \delta} \ \le\ 2\delta}$$

$$\overset{d>2}{\lesssim} \frac{1}{\sqrt{n}} c^d \delta^{-d/2+1} + \delta \lesssim \frac{1}{n^{1/d}}$$

optimal choice: $\delta = \frac{c^2}{n^{1/d}}$

- $\Rightarrow$ we proved:

$$\boxed{\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X_i) \right| \le \frac{CL}{n^{1/d}}} \quad \forall \text{ dim } d > 2.$$

OPTIMAL.

- Unfortunately, to make $n^{-1/d} = o(1)$, one needs $\boxed{n > \exp(d)}$

$\Rightarrow$ Lipschitz regression suffers from the curse of high dimensionality.

Mention Wasserstein