Q1: Behaviour of VC under operations

(a) Take $d \in \mathbb{N}$.

If $f^* \in \bigcap_{i=1}^{d} F_i$, $\exists (f_1, \dots f_d) \in \bigotimes_{i=1}^{d} F_i$ s.t.

$$f^*(x) = \min_{i \in [d]} f_i(x) \quad \forall x \in \mathcal{X}.$$

Consider the map $M : \bigotimes_{i=1}^{d} F_i(\underline{x}) \to \bigcap_{i=1}^{d} F_i(\underline{x})$ defined by

$$M(\underline{x}) : \begin{bmatrix} f_1(x_1) & \dots & f_1(x_n) \\ f_2(x_1) & \dots & f_2(x_n) \\ f_d(x_1) & \dots & f_d(x_n) \end{bmatrix} \longmapsto \begin{bmatrix} \min_{i \in [d]} f_i(x_1) \\ \vdots \\ \min_{i \in [d]} f_i(x_n) \end{bmatrix}.$$

M is surjective, since $\forall x_j, j \in [n]$, $\min_{i \in [d]} f_i(x_j)$ can be achieved by picking some element of a d-tuple of functions from $\bigotimes_{i=1}^{d} F_i$

$$\left|\left(\bigcap_{i=1}^{d} F_i\right)(\underline{x})\right| = \left|\left\{\left[\min_{i \in [d]} f_i(x_1), \dots \min_{i \in [d]} f_i(x_n)\right] \mid f_i \in F_i \ \forall i \in [d]\right\}\right|$$

$$= |\text{image}(M(x))|$$

$$\leq |\text{domain}(M(x))|$$

$$= \left|\left\{\begin{bmatrix} f_1(x_1) & \dots & f_1(x_n) \\ \vdots & & \\ f_d(x_1) & \dots & f_d(x_n) \end{bmatrix} \mid f_i \in F_i \ \forall i \in [d]\right\}\right|$$

$$\leq \prod_{i=1}^{d} |F_i(x)|$$

Now with $k = \sum_{i=1}^{d} VCD(\mathcal{F}_i) = \sum_{i=1}^{d} \sup_{n \geq 1} \left\{ n \mid \sup_{X \in \mathcal{X}^n} |\mathcal{F}_i(X)| = 2^n \right\}$

The Sauer-Shelah inequality tells us that for $X_1, \dots X_n \in \mathcal{X}^n$, $n \geq k$,

$$|F_i(X)| \leq O(n^k), \quad k = VCD(\mathcal{F}_i)$$

Hence
$$\prod_{i=1}^{d} |F_i(X)| \leq \prod_{i=1}^{d} O(n^{VCD(F_i)})$$
$$= O\left(n^{\sum_{i=1}^{d} VCD(F_i)}\right)$$
$$= O(n^k)$$

Hence
$$\left|\left(\bigcap_{i=1}^{d} \mathcal{F}_i\right)(X)\right| \leq O(n^k).$$

This holds $\forall (X_1, \dots X_n) \in \mathcal{X}^n$, hence taking sup over $\mathcal{X}^n$ on LHS

$$\sup_{X \in \mathcal{X}^n} \left|\left(\bigcap_{i=1}^{d} \mathcal{F}_i\right)(X)\right| \leq O(n^k)$$

$$\iff \log\left(\sup_{X \in \mathcal{X}^n} \left|\left(\bigcap_{i=1}^{d} \mathcal{F}_i\right)(X)\right|\right) \leq O(k \log n)$$

$$\iff \sup_{n \geq 1} \left\{ n \mid \sup_{X \in \mathcal{X}^n} \left|\left(\bigcap_{i=1}^{d} \mathcal{F}_i\right)(X)\right| = 2^n \right\} \leq O(k \log n)$$

$$\iff VCD\left(\bigcap_{i=1}^{d} \mathcal{F}_i\right) \leq O(k \log n)$$

$$\iff VCD\left(\bigcap_{i=1}^{d} \mathcal{F}_i\right) = O(k \log n)$$

Althoug we assume $n \geq k$, note that $n \leq k$ since

$$k = \sum_{i=1}^{d} VCD(\mathcal{F}_i) = \sum_{i=1}^{d} \sup_{n \geq 1} \{ n \mid \sup_{x \in \mathcal{X}^n} |\mathcal{F}_i(x)| = 2^n \}$$

$$\geq \sum_{i=1}^{d} n = dn \geq n. \text{ for any } n \in \mathbb{N}.$$

Hence
$$O(k \log n) = O\left(k\left(\log k + \log\left(\frac{n}{k}\right)\right)\right)$$

$$= O(k(\log k + \log d))$$

$$= O(k \log k + O(1))$$

$$= O(k \log k).$$

By the above this tells us $VCD\left(\bigcap_{i=1}^{d} \mathcal{F}_i\right) \leq O(k \log k)$.

The same proof holds by replacing "min" with "max" in the definition of M which will still be surjective, hence we also have

$$VCD\left(\bigcup_{i=1}^{d} \mathcal{F}_i\right) \leq O(k \log k).$$

(b) $\mathcal{F}^d_{left} = \{ x \mapsto \mathbb{1}[x_i \le t_i \ \forall i \in [d], \ t \in \mathbb{R}^d] \}$.

Let $\mathcal{F}_i := \{ x \mapsto \mathbb{1}_{\{x_i \le t_i\}} \mid t_i \in \mathbb{R} \} \ \forall i \in [d]$

Then $\bigcap_{i=1}^{d} \mathcal{F}_i = \{ x \mapsto \min_{i \in [d]} \mathbb{1}_{\{x_i \le t_i\}} \mid t_i \in \mathbb{R} \}$

$= \{ x \mapsto \min_{i \in [d]} \mathbb{1}_{\{x_i \le t_i \ \forall i \in [d]\}} \mid t \in \mathbb{R}^d \}$

$$\min_{i \in [d]} \mathbb{1}_{\{x_i \le t_i \ \forall i \in [d]\}} = \begin{cases} 0 & \text{if } x_i > t_i \text{ for some } i \in [d] \\ 1 & \text{if } x_i \le t_i \ \forall i \in [d] \end{cases}$$

$= \mathbb{1}_{\{x_i \le t_i \ \forall i \in [d]\}}$.

Hence $\bigcap_{i=1}^{d} \mathcal{F}_i = \mathcal{F}^d_{left}$.

By part (1), this implies $VCD(\mathcal{F}_{left}) \le O(k \log k)$,

with $k = \sum_{i=1}^{d} VCD(\mathcal{F}_i) = \sum_{i=1}^{d} 1 = d$, since VCD of

halfspaces in $\mathbb{R}$ is 1 by example 2.3 in lecture.

Hence $VCD(\mathcal{F}_{left}) \le O(d \ln d)$. (1).

We now conclude:

Since $\mathcal{F}^d_{left}$ is a class of uniformly bounded functions (boolean class), we have by theorem 2.2

$$\| P_n - P \|_{\mathcal{F}^d_{left}} := \sup_{f \in \mathcal{F}^d_{left}} | P_n f - P f | \leq 2 R_n (\mathcal{F}^d_{left}) + \delta$$

w.p. $\geq 1 - \exp\{ - \delta^2 n / 2b^2 \}$.

Since $VCD(\mathcal{F}^d_{left}) = O(d \log d)$, by Sauer-Shelah inequality if $n \geq d \log d$,

For $(X_1, \dots X_n) \in \mathcal{X}^n$, $\quad | \mathcal{F}^d_{left}(x) | \leq \sum_{i=0}^{d \log d} \binom{n}{d} \leq (n+1)^{O(d \ln d)}$

Hence $\mathcal{F}^d_{left}$ has polynomial discrimination $O(d \ln d)$

Hence with $K = O(d \ln d)$,

$$R_n(\mathcal{F}) \leq 4 C_{\mathcal{F}^d_{left}, \mathbb{P}} \sqrt{\frac{K \ln(n+1)}{n}}, \quad C_{\mathcal{F}^d_{left}, \mathbb{P}} = \mathbb{E}_X \sup_f \sqrt{P_n f^2} < \infty$$

indeed $C_{\mathcal{F}^d_{left}, \mathbb{P}}$ is bounded since $\mathcal{F}^d_{left}$ is uniformly bounded.

i.e. $\| (P_n - P) f \|_{\mathcal{F}^d_{left}} = 8 C_{\mathcal{F}^d_{left}, P} \sqrt{\frac{K \ln(n+1)}{n}} + \delta \quad$ w.p. $\geq 1 - \exp\left\{ \frac{-\delta^2 n}{2b^2} \right\}$

Letting $\delta \downarrow 0$, we have $P_n f - P f \to 0$ uniformly in n and hence the set of distribution functions over $\mathbb{R}^d$ is Glivenko-Cantelli.

Q2: VC talent show

(a) Let $\mathcal{F} := \{ x \mapsto \mathbb{1}_{\{p(x) \geq t\}}, t \in \mathbb{R}, p \text{ degree } k \text{ polynomial of } \mathbb{R}^d \}$

$\Rightarrow \mathcal{G} = \{ x \mapsto t - p(x), t \in \mathbb{R}, p \text{ degree } k \text{ polynomial of } \mathbb{R}^d \}$

$\mathcal{F}$ is subgraph class of $\mathcal{G}$.

$\mathbb{R}_k[\mathbb{R}^d]$ has dimension $\binom{d+k}{d}$, and t is in constant term.

rewrite: $x \mapsto t - p(x)$ as $x \mapsto \tilde{p}(x)$, $\tilde{p} \in \mathbb{R}_k[\mathbb{R}^d]$

$\Rightarrow \mathcal{G} := \{ x \mapsto \tilde{p}(x), \tilde{p} \in \mathbb{R}_k[\mathbb{R}^d] \}$

$\Rightarrow \dim(\mathcal{G}) = \binom{d+k}{d}$

$\Rightarrow VCD(\mathcal{F}) \leq \binom{d+k}{d}$ by lecture.

1b) The class of convex polygons in $\mathbb{R}^2$ has infinite VC dimension.

The class of convex polygones in $\mathbb{R}^2$ can be written as

$\mathbb{1}\{Ax \leq b\}$, with $A \in \mathbb{R}^{k \times 2}$, $b \in \mathbb{R}^k$, k free, $\rho(A) \geq 1$

For $n \in \mathbb{N}$, spread n points uniformly along unit circle in $\mathbb{R}^2$

For an arbitrary labelling $\varepsilon \in \{0,1\}^n$, consider the

polygone defined by the vertices where $\varepsilon_i = 1$.

Since we can construct any labelling for any dimension,

the result follows.

(c) The VC dimension is bounded by 2^d

Note $\forall\ y \in \{0,1\}^n$, we can find a point $x \in \mathbb{R}^d$ s.t.
$\text{sgn}(x)_i = \mathbb{1}[x_i \geq 0] - \mathbb{1}[x_i < 0] = y_i \quad \forall i \in d]$.

Hence $\forall f$ as given in problem, $|\text{dom} f| = 2^d$

Suppose $(X_1, \dots X_n) \in (\{\pm 1\}^d)^n$, then with $n > 2^d$,

By pigeon hole principle $\exists\ i \neq j$ s.t. $\text{sgn}(X_i) = \text{sgn}(X_j)$

Hence $f(X_i) = f(X_j)$, even though $X_i \neq X_j$.

Since this holds $\forall f \in \mathcal{F}$, $\mathcal{F}$ can not shatter X.

Question 3. Covering number bounds from VC bounds

(a) If $\mathcal{P}(\mathcal{F}, L^1(\mathbb{P}), 1)$ is 1-packing of $(\mathcal{F}, L^1(\mathbb{P}))$,
then $\forall f, g \in \mathcal{P}(\mathcal{F}, L^1(\mathbb{P}), 1)$, we have $\|f-g\|_1 > 1$ unless $f = g$ a.s.

However, $\forall f, g \in \mathcal{F}$, $\|f-g\|_1 = \int_{\mathcal{X}} |f(x)-g(x)|\, dx \leq \int_{\mathcal{X}} 1\, dx = 1.$

But by definition of packing $\mathcal{P}(\mathcal{F}, L^1(\mathbb{P}), 1) \subseteq \mathcal{F}$
Hence if $f, g \in \mathcal{P}(\mathcal{F}, L^1(\mathbb{P}), 1)$, it must be the case that $f \equiv g$.
Hence the 1-packing number of $(\mathcal{F}, L^1(\mathbb{P}))$ is 1.

(b) Denote the probability that the sets S_i are all distinct
by $\mathbb{P}[S_i \neq S_j \quad \forall i \neq j \in [N]]$

$$\mathbb{P}[\{S_i \neq S_j \ \forall i \neq j \in [N]\}^c] = 1 - \mathbb{P}[\{\exists i \neq j \in [N] \text{ s.t. } S_i = S_j\}]$$

But we have

$$\mathbb{P}[\exists i \neq j \in [N] \text{ s.t. } S_i = S_j] = \mathbb{P}\Big[\bigcup_{i \neq j \in [N]} \{S_i = S_j\}\Big]$$

$$\leq \sum_{i \neq j \in [N]} \mathbb{P}[S_i = S_j]$$

$$\leq \binom{N}{2} \mathbb{P}[S_1 \neq S_2] \quad \text{Since } X_i \text{ iid}$$

$$= \binom{N}{2} \mathbb{P}[\{f_1(x_i) = f_2(x_i) \ \forall i \in [n]\}]$$

$$= \binom{N}{2} \mathbb{P}\left[\{ f_1(X_i) = f_2(X_i) \} \right]^n \qquad X_i \text{ iid}$$

But $\mathbb{P}\left[\{ f_1(X_i) = f_2(X_i) \} \right] = 1 - \mathbb{P}\left[\{ |f_1(X_i) - f_2(X_i)| > 0 \} \right]$

$$= 1 - \int_{\mathcal{X}} \mathbb{1}_{\{ |f_1(x) - f_2(x)| > 0 \}} \, d\mathbb{P}(x)$$

$$(*) \quad = 1 - \int_{\mathcal{X}} |f_1(x) - f_2(x)| \, d\mathbb{P}(x)$$

$$= 1 - \delta$$

(*) Since $|f_1(x) - f_2(x)| \in \{0, 1\} \Rightarrow \mathbb{1}_{\{ |f_1(x) - f_2(x)| > 0 \}} = |f_1(x) - f_2(x)|$

Hence $\mathbb{P}\left[\exists i \neq j \in [N] \text{ s.t. } S_i = S_j \right] \leq \binom{N}{2} (1 - \delta)^n$

$$\Leftrightarrow \mathbb{P}\left[S_i \neq S_j \quad \forall i \neq j \in [N] \right] \geq 1 - \binom{N}{2} (1 - \delta)^n .$$

This is what we wanted to show.

(c) We know $VCD(\mathcal{F}) \leq k$, and using part (b), if we show that the probability that the S_i's are all distinct is positive, then we have shown that $\mathcal{F}$ shatters $\underline{x} \in \mathcal{X}^n$.

If $1 - \binom{N}{2}(1-\delta)^n > 0 \iff \binom{N}{2}(1-\delta)^n < 1$

$$\iff \frac{N(N-1)}{2}(1-\delta)^n < 1$$

$\frac{N(N-1)}{2} \leq N^2$ $\quad\quad \Leftarrow (1-\delta)^n N^2 < 1$

$$\iff n \log(1-\delta) < -2\log N$$

$e^{-u} > 1-u$
$-u > \log(1-\delta)$ $\quad\quad \iff n > \frac{-2\log N}{\log(1-\delta)}$ since $\log(1-\delta) < 0$

$\log(1-\delta) < -\delta$ for $0<\delta<1$ $\quad\quad \Leftarrow n > \frac{-2\log N}{-\delta}$

$$\iff n > \frac{2\log N}{\delta}$$

Hence for $n > \frac{2\log N}{\delta}$, with $\mathcal{F}_N := \{f_1, \dots f_N\}$, $\exists\, x \in \mathcal{X}^n$ s.t. $|\mathcal{F}_N(x)| = N$. But $\mathcal{F}_N \subseteq \mathcal{F}$, so $|\mathcal{F}_N(x)| \leq |\mathcal{F}(x)| \leq (n+1)^k$, since $VCD(\mathcal{F}) = k$. Hence $N \leq k$.

From above we also note $\frac{N(N-1)}{2}(1-\delta)^n < 1 \Rightarrow (N-2)(N+1) < 0 \Rightarrow N < 2$
Since $N \geq 0$. hence $\neg(N < -1)$. So we have

$$N \leq \max\left(2, (n+1)^k\right) \leq \max\left(2, \left(\frac{2\log N}{\delta}\right)^k\right)$$

$$\leq \max\left(2, \left(\frac{5\log N}{\delta}\right)^k\right).$$

This is what we wanted to show.

(d) $\mathcal{N}(\mathcal{F}, L^1(\mathbb{P}), \varepsilon)$ is the size of the smallest possible ε-net of $\mathcal{F}$ with respect to $L^1(\mathbb{P})$.

$$\mathcal{N}(\mathcal{F}, L^1(\mathbb{P}), \varepsilon) \leq \mathcal{P}(\mathcal{F}, L^1(\mathbb{P}), \varepsilon)$$
$$= N$$
$$\leq \max\left(2, \left(\frac{5\log W}{\varepsilon}\right)^k\right)$$

Case 1: $\quad 2 > \left(\frac{5\log N}{\varepsilon}\right)^k$.

If $\varepsilon \geq 1$, $\mathcal{P}(\mathcal{F}, L^1(\mathbb{P}), \varepsilon) = 1$. Hence wlm $1 \leq \varepsilon < 10k$,

$$(2k)^{2k}\left(\frac{5}{\varepsilon}\right)^{2k} > 1 \iff 2k\left(\frac{5}{\varepsilon}\right) > 1 \iff \varepsilon < 10k$$

Hence $\mathcal{N}(\mathcal{F}, L^1(\mathbb{P}), \varepsilon) < C_k\left(\frac{5}{\varepsilon}\right)^{2k}$.

It is reasonable that $\varepsilon < 10k$: for $f, g \in \mathcal{F}$, $f \neq g$, we have $\|f-g\|_1 > \varepsilon$. But suppose $\|f-g\|_1 \geq 10k$.

Then $\|f-g\|_1 > k$

$$\Leftrightarrow \int_{\mathcal{X}} |f(x)-g(x)|\, d\mathbb{P}(x) > k$$

$$\Rightarrow 1 \geq \int_{\mathcal{X}} |f(x)-g(x)|\, d\mathbb{P}(x) > k \quad ⚡.$$

<u>If $\varepsilon < 1$,</u> then $\mathcal{N}(\mathcal{F}, L^1(\mathbb{P}), \varepsilon)$

$$\leq 2$$

$$\leq (2k)^{2k}\left(\frac{5}{\varepsilon}\right)^{2k} \qquad \text{for } k \geq 1.$$

<u>case 2:</u> $\left(\frac{5\log W}{\varepsilon}\right)^k > 2$. Then

<u>If $\varepsilon \geq 1$,</u> as before we have $\mathcal{N}(\mathcal{F}, L^1(\mathbb{P}), \varepsilon) \leq C_k\left(\frac{5}{3}\right)^{2k}$.

<u>If $\varepsilon < 1$.</u> $\left(\frac{1}{\varepsilon}5\log W\right)^k$

$$\leq \left(\frac{5}{\varepsilon}\right)^k\left(\log\max\left\{2, \left(\frac{1}{\varepsilon}5\log W\right)^k\right\}\right)^k$$

$$\leq \left(\frac{5}{\varepsilon}\right)^k\left(\log\left(k\frac{5}{\varepsilon}\log W\right)\right)^k$$

$$\leq \left(\frac{5}{\varepsilon}\right)^k\left(k\frac{5}{\varepsilon}\log W\right)^k$$

$$= \left(\frac{5}{\varepsilon}\right)^{2k} k^k (\log W)^k$$

Note since $|\mathcal{F}_N(X)| = N \leq |\mathcal{F}(X)| \leq 2^k$, we have $\log N \leq k$.

Hence $\left(\frac{1}{\varepsilon} 5 \log N\right)^k \leq \left(\frac{5}{\varepsilon}\right)^{2k} k^k \cdot k^k = C_k \left(\frac{5}{\varepsilon}\right)^{2k}$.

<u>Summary</u>: $\forall \varepsilon > 0$, $\mathcal{N}(\mathcal{F}, L^1(\mathbb{P}), \varepsilon) \leq C_k \left(\frac{5}{\varepsilon}\right)^{2k}$.

4. Fitting a Lipschitz function.

Let us define $X_f^i = (f(X_i) - Y_i)^2 - \mathbb{E}[(f(X) - Y)^2]$

$$|X_f^i - X_g^i| = \Big| f^2(X_i) - g^2(X_i) + 2h(X_i)(g(X_i) - f(X_i))$$

$$+ \mathbb{E}\left[g^2(X) - f^2(X) + 2h(X)(f(X) - g(X)) \right] \Big|$$

$$\leq \big| (f(X_i) - g(X_i))(f(X_i) + g(X_i) - 2h(X_i)) \big|$$

$$+ \big| \mathbb{E}\left[(f(X) - g(X))(f(X) + g(X) - 2h(X)) \right] \big|$$

$$\leq \|f - g\|_\infty \Big(|f(X_i) + g(X_i) - 2h(X_i)|$$

$$+ \mathbb{E}\left[|f(X) + g(X) - 2h(X)| \right]$$

$$\leq 4 \|f - g\|_\infty$$

So $X_f^i - X_g^i$ is $4\|f - g\|_\infty$- subgaussian by HW 6

Since $X_f - X_g = \frac{1}{n} \sum_{i=1}^{n} X_f^i - X_g^i$, then $X_f - X_g$ is

$\|f - g\|_\infty \frac{4}{\sqrt{n}}$ - subgaussian using HW 6. Q 2.

(1b) Take a grid on $[0,1]^2$, with ε-grid on y-axis and $\frac{\varepsilon}{L}$ on x axis.

Consider piecewise linear functions on this grid with slope no greater than L.

3 options for x_{i+1} from x_i for given f. (slopes at L, $-L$ or 0).

$\Rightarrow$ # functions from x_0 starting point $\leq 3^{L/\varepsilon}$

Consider $\{x_0 : x_0 \in [0,1], |x_0^{(1)} - x_0^{(2)}| = \varepsilon\} := A$

$\Rightarrow |A| 3^{L/\varepsilon} = \frac{1}{\varepsilon} 3^{L/\varepsilon}$ functions.

WTS: ε-net of $\mathcal{F}_L$. Take $f \in \mathcal{F}_L$.

$\forall$ $\tilde{f}$ in our net, $\hat{f}$ uniquely defined by $(\tilde{f}(x_k))$, with $x_k = \frac{k\varepsilon}{2}$ ($\check{f}$ is a peicewise linear function passing throug all the points)

chose $\tilde{f}(0)$ s.t. $|f(0) - \tilde{f}(0)| \leq \frac{\varepsilon}{2}$. ($\varepsilon$-grid on y-axis)

let's show that if $|f(x_i) - \tilde{f}(x_i)| \leq \frac{\varepsilon}{2}$, we can choose $\tilde{f}(x_{i+1})$ such that

$$|f(x_{i+1}) - \check{f}(x_{i+1})| \leq \frac{\varepsilon}{2} \quad \text{and}$$

$$\sup_{x \in [x_i, x_{i+1}]} |f(x) - \tilde{f}(x)| \leq \varepsilon$$

$f \in \tilde{\mathcal{F}}_L$ is L-Lipschitz, $|x_{i+1} - x_i| = \frac{\varepsilon}{L}$,

$\Rightarrow \quad |f(x_i) - f(x_{i+1})| \leq \varepsilon$

$\Rightarrow \quad |\tilde{f}(x_i) - f(x_{i+1})| \leq \frac{3\varepsilon}{2}$ since $|\tilde{f}(x_i) - f(x_i)| \leq \frac{\varepsilon}{2}$

i.e. $f(x_{i+1}) \in \left[\tilde{f}(x_i) - \frac{3\varepsilon}{2}, \tilde{f}(x_i) + \frac{3\varepsilon}{2}\right]$

For $\tilde{f}(x_{i+1})$, we can take any value in

$\{\tilde{f}(x_i) - \varepsilon, \tilde{f}(x_i) + \varepsilon\}$

$\Rightarrow$ chose $\tilde{f}(x_{i+1})$ s.t. $|\tilde{f}(x_{i+1}) - f(x_{i+1})| \leq \frac{\varepsilon}{2}$

For $x \in [x_i, x_{i+1}]$, $\exists t \in [0,1]$ s.t. $x = tx_i + (1-t)x_{i+1}$

$|f(x) - \tilde{f}(x)| = |f(x) - t\tilde{f}(x_i) - (1-t)\tilde{f}(x_{i+1})|$ $\tilde{f}$ linear on $[x_i, x_{i+1}]$

$= |t(f(x) - f(x_i)) + (1-t)(f(x) - f(x_{i+1}))$
$+ t(f(x_i) - \tilde{f}(x_i)) + (1-t)(f(x_{i+1}) - \tilde{f}(x_{i+1}))|$

$\leq t|x - x_i| + (1-t)|x - x_{i+1}| + \frac{t\varepsilon}{2} + \frac{(1-t)\varepsilon}{2}$

$= t(1-t)\varepsilon + t(1-t)\varepsilon + \varepsilon/2$

$\leq \varepsilon$ with $t(1-t) \leq \frac{1}{4}$ $\forall t \in [0,1]$

Hence we can build $\tilde{f}$ by induction that belongs to our set (which is in $\mathcal{F}_L$), s.t.

$$\|f - \tilde{f}\|_\infty \leq \varepsilon$$

Hence our set is an ε-net of $\mathcal{F}_L$, and

$$N(\mathcal{F}_L, \|\cdot\|_\infty, \varepsilon) \leq \frac{1}{\varepsilon} 3^{L/\varepsilon}$$

$$\leq \frac{1}{\varepsilon} \exp\left((\log 3) \frac{L}{\varepsilon}\right)$$

$$= \frac{1}{\varepsilon} \exp\left(\frac{cL}{\varepsilon}\right) \quad \text{with} \quad c = \log 3$$

(c) We use Dudley's entropy integral

$$\mathbb{E} \sup_{f \in \mathcal{F}_L} |L_n f - Lf| = \mathbb{E} \sup_{f \in \mathcal{F}_L} X_f$$

$$\leq c' r \int_0^\infty \sqrt{\frac{cL}{\varepsilon}}\, d\varepsilon$$

$$= c'' \sqrt{\frac{L}{n}}$$

Let $g \in \mathcal{F}_L$: $g : [0,1] \to [0,1]$ by $g(x) = 0$.

$$\mathbb{E} \sup_{f \in \mathcal{F}} |X_f| = \mathbb{E} \sup_{f \in \mathcal{F}} |X_f - X_g + X_g|$$

$$\leq \mathbb{E} \sup_{f \in \mathcal{F}} |X_f - X_g| + \mathbb{E}|X_g|$$

$$\leq \mathbb{E} \max \left\{ \sup_{f \in \mathcal{F}_L} X_f - X_g, \sup_{f \in \mathcal{F}_L} X_g - X_f \right\} + \mathbb{E}|X_g|$$

both supp ≥ 0 (take $f = g$)

$$\leq \mathbb{E} \sup_{f \in \mathcal{F}_L} \{X_f - X_g\} + \mathbb{E} \sup_{f \in \mathcal{F}_L} \{X_g - X_f\} + \mathbb{E}|X_g|$$

but for $f_1, f_2 \in \mathcal{F}_L$,

$$\begin{cases} (X_{f_1} - X_g) - (X_{f_2} - X_g) = X_{f_1} - X_{f_2} \\ (X_{f_2} - X_g) - (X_{f_1} - X_g) = X_{f_2} - X_{f_1} \end{cases}$$

Both $(X_{f_1} - X_{f_2})$ and $(X_{f_2} - X_{f_1})$ are subgaussian w/ param

$\frac{4\|f-g\|_\infty}{\sqrt{n}}$. Note that $\sup_{f, g \in \mathcal{F}_L} \|f-g\|_\infty = 1$, so

$\mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, \varepsilon) = 1$ if $\varepsilon > 1$. Applying the integral inequality,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}_L} |X_f|\right] \leq 2\tilde{C} \int_0^1 \frac{4}{\sqrt{n}} \sqrt{\log(\mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, \varepsilon))}\, d\varepsilon + \mathbb{E}[|X_g|]$$

for some $\hat{C} \in \mathbb{R}$.

i.e. $\mathbb{E}\left[\sup_{f \in \mathcal{F}_L} |X_f|\right] \le \frac{8\tilde{C}}{\sqrt{n}} \int_0^1 \underbrace{\sqrt{\frac{cL}{\varepsilon} + \log \varepsilon}}_{} \, d\varepsilon + \mathbb{E}[|X_g|]$

$O\left(\frac{1}{\sqrt{\varepsilon}}\right)$ as $\varepsilon \to 0$,

so integral converges.

$\le \frac{\tilde{C}'}{\sqrt{n}} + \mathbb{E}[|X_g|]$.

$\mathbb{E}[|X_g|] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n Y_i^2 - \mathbb{E}[Y^2]\right]$

$\left|\frac{1}{n}\sum_{i=1}^n Y_i^2 - \mathbb{E}[Y^2]\right| \in [0,1]$, $Y^2 \in [0,1]$, so

$\mathbb{E}[|X_g|] = \int_0^1 \mathbb{P}(|\bar{Y}_n^2 - \mathbb{E}Y^2| > t)\, dt$

$\le 2\int_0^1 \exp(-2t^2 n)\, dt$ by Azuma-Hoeffding corollary

$\le \sqrt{\frac{2}{n}} \int_0^{2n} \exp(-t^2)\, dt$

$\le \sqrt{\frac{2}{n}} \int_0^{\infty} \exp(-t^2)\, dt$

$= \frac{\tilde{C}''}{\sqrt{n}}$, $\tilde{C}'' = \sqrt{2}\int_0^\infty e^{-t^2}\, dt < +\infty$

$\Rightarrow \mathbb{E}\left[\sup_{f \in \mathcal{F}_L} |X_f|\right] \le \frac{C}{\sqrt{n}}$, $C \in \mathbb{R}$.

(d) Generalization: excess risk is

$$L(\hat{f}) - L(h) \le |L(\hat{f}) - L_n(\hat{f})| + (L_n(\hat{f}) - L_n(h)) + |L_n(h) - L(h)|$$

- $\mathbb{E}\left[\sup_{f \in \mathcal{F}_L} |L_n(f) - L(f)|\right] \le \frac{c'}{\sqrt{n}}$ by (c)
- $L_n(\hat{f}) = \min_{f \in \mathcal{F}_L} L_n(f) \le L_n(h)$
- $L_n(h) - L(h) = P_n((h(x) - h(x))^2) - \mathbb{E}[(h(x) - h(x))^2] = 0$

$$\begin{aligned}
\mathbb{E}[L(\hat{f}) - L(h)] &\le \mathbb{E}\left[\sup_{f \in \mathcal{F}_L} (L(f) - L(h))\right] \\
&\le \mathbb{E}\left[\sup_{f \in \mathcal{F}_L} |L(f) - L_n(f)|\right] + 0 + 0 \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}_L} |X_f|\right] \\
&\le \frac{c'}{\sqrt{n}}
\end{aligned}$$

5. Lower bounds on supremum of Gaussian process

(a) With $X_i = Z_i + \xi_i$,
$Z \sim N(0, 1-\varepsilon)$ and ξ_i sampled iid $N(0, \varepsilon)$

$$\begin{aligned} \mathbb{E}\left[(X_t - X_s)^2\right] &= \mathbb{E}\left[(Z + \xi_t - Z - \xi_s)^2\right] \\ &= \mathbb{E}\left[(\xi_t - \xi_s)^2\right] \\ &= \mathbb{E}\,\xi_t^2 + \mathbb{E}\,\xi_s^2 - 2\underbrace{\mathbb{E}\,\xi_t \xi_s}_{=0} \\ &= 2\varepsilon \end{aligned}$$

$$\begin{aligned} \mathbb{E}\left[(Y_t - Y_s)^2\right] &= \mathbb{E}\left[Y_t^2\right] + \mathbb{E}\left[Y_s^2\right] - 2\underbrace{\mathbb{E}\left[Y_t Y_s\right]}_{\text{independence}} \\ &= 2 \end{aligned}$$

We note $\mathbb{E}\left[\sup_{t \in T} X_t\right] = \mathbb{E}\left[\sup_{t \in T} Z + \xi_t\right] = \mathbb{E}\left[\sup_{t \in T} \xi_t\right]$

as $\xi \overset{d}{\sim} \sqrt{\varepsilon}\, Y$ vector-wise

Note $\mathbb{E}\left[(X_t - X_s)^2\right] \leq \mathbb{E}\left[(Y_t - Y_s)^2\right] \Leftrightarrow \varepsilon \leq 1$

In this set-up the Sudakov Inequality makes sense because X_1 has smaller variance than Y_1 hence we expect smaller supremum.

(d) Let X_t be a Gaussian RV distributed according to $N(0, \varepsilon)$ for every $t \in P$, where P is a maximal δ-packing of T.

First consider another process Y_t on P with $\mathbb{E}[(Y_t - Y_s)^2] = \delta$. This can be constructed by part (a), with $\delta = \sqrt{\varepsilon}$

By definition of the packing $\rho(s,t) > \delta$, so the assumption of the Sudakov-Fernique inequality holds, and we conclude with

$$\mathbb{E}(Y_t - Y_s)^2 \geq \delta^2 = \mathbb{E}[(X_t - X_s)^2] \quad \text{that}$$

$$\mathbb{E}\left[\sup_{t \in T} Y_t\right] \geq \mathbb{E}\left[\max_{t \in P} Y_t\right] \geq \mathbb{E}\left[\max_{t \in P} Y_t\right]$$

$$= \mathbb{E}\left[\max_{t \in P} \varepsilon Z_t\right] \qquad Z_i \overset{iid}{\sim} N(0,1)$$

$$= \varepsilon\, \mathbb{E}\left[\max_{t \in P} Z_t\right]$$

$$(*) \qquad \geq \varepsilon\, c \sqrt{\log |P|}$$

Since $|\mathcal{N}(T, \rho, \varepsilon)| \leq |P(T, \rho, \varepsilon)|$, this concludes the proof.

$(*)$ Because $\mathbb{E}\left[\max_{i \in [n]} Z_i\right] \geq C \log n$ for Z_i IID standard Gaussian.

Indeed: $\mathbb{E}\left[\max_{i\in[n]} Z_i\right] \geq \alpha\, \mathbb{P}\left[\max_{i\in[n]} Z_i > \alpha\right] +$

$$\mathbb{E}\left[\max_{i\in[n]} Z_i \,\middle|\, \max_{i\in[n]} Z_i < 0\right] \mathbb{P}\left[\max_{i\in[n]} Z_i < 0\right]$$

$$= \alpha\left(1 - \Phi(\alpha)^n\right) + \mathbb{E}\left[Z_1 \mid Z_1 < 0\right] (1/2)^n$$

$$\geq \alpha\left(1 - \left(1 - \frac{1}{\alpha\sqrt{2\pi}} e^{-\alpha^2/2}\right)^n\right) + c'/2^n$$

$\mathbb{E}\left[\max_{i\in[n]} Z_i\right] \geq c\sqrt{\log n}$ with $\alpha = c''\sqrt{\log n}$

where we used $\mathbb{E}\left[\max_{i\in[n]} Z_i \mid Z_i < 0 \ \forall i \in [n]\right]$

is monotonically increasing in n, and 'anti' concentration

$$\mathbb{P}[Z > \alpha] \geq \frac{\phi(\alpha)}{\alpha}$$